FACULTAD DE INGENIERÍA Y ARQUITECTURA

ESCUELA PROFESIONAL DE INGENIERÍA DE COMPUTACIÓN Y SISTEMAS

SISTEMA DE PREDICCIÓN DE HECHOS DELICTIVOS PARA LA MEJORA DEL PROCESO DE PREVENCIÓN DEL DELITO EN EL DISTRITO DE LA MOLINA UTILIZANDO MINERÍA DE DATOS

PRESENTADO POR

JORGE JULIO JAULIS RUA

JONATHAN RENATTO VILCARROMERO GIRALDO

TESIS

PARA OPTAR EL TÍTULO PROFESIONAL DE INGENIERO DE COMPUTACIÓN Y SISTEMAS

LIMA – PERÚ

2015

USMP
UNIVERSIDAD DE SAN MARTIN DE PORRES
FACULTAD DE INGENIERÍA Y ARQUITECTURA

ESCUELA PROFESIONAL DE INGENIERÍA DE COMPUTACIÓN Y SISTEMAS

SISTEMA DE PREDICCIÓN DE HECHOS DELICTIVOS PARA LA MEJORA DEL PROCESO DE PREVENCIÓN DEL DELITO EN EL DISTRITO DE LA MOLINA UTILIZANDO MINERÍA DE DATOS

TESIS

PARA OPTAR EL TÍTULO PROFESIONAL DE INGENIERO DE COMPUTACIÓN Y SISTEMAS

PRESENTADO POR

JAULIS RUA, JORGE JULIO

VILCARROMERO GIRALDO, JONATHAN RENATTO

LIMA – PERÚ

2015

Dedicatoria

A Dios, porque nos da las fuerzas de para seguir adelante. A nuestra familia ya que con su apoyo incondicional damos un paso más hacia nuestro éxito profesional.

Agradecimiento

Agradecemos a nuestros asesores por el apoyo incesante cada semana, a nuestra alma mater la USMP por hacer de nosotros unos profesionales con una formación sólida.

A nuestros familiares por siempre ser ese motor que nos impulsa a seguir adelante y nos inculca a no desistir a nuestros sueños.

ÍNDICE

Lista de tablas

Página

Lista de figuras

Página

Lista de anexos

RESUMEN

El proyecto llamado sistema de predicción de hechos delictivos para la mejora del proceso de prevención del delito en el distrito de La Molina utilizando minería de datos, tiene como objetivo mejorar el proceso relacionado a prevenir delitos que realizan las 3 comisarías de la jurisdicción, centrándose en delitos con dolo, para así cumplir el objetivo principal de las comisarias que es el de salvaguardar y mantener el orden en la zona.
Gracias a las nuevas tecnologías de Información, y en especial a la minería de datos, es posible obtener predicciones de delitos que podrían pasar a futuro en base a los registros de datos históricos; y con ello desarrollar una solución que plasme esta información y sirva como herramienta de conocimiento en el proceso de prevención del delito, inicialmente, en la comisaria de Santa Felicia.

El modelo de minería de datos recoge información histórica de todo el año 2015 de las denuncias registradas en cada una de las comisarías, y en base a un algoritmo de aprendizaje automático arroja las zonas más propensas a la ocurrencia de algún hecho delictivo, para ello se optó por mostrar esta información a través de mapas de la zona y que esto puedan ser accedidos desde cualquier dispositivo (Web o móvil).

La solución facilitó a los efectivos policiales en la identificación de zonas de riesgo, asignación de recursos policiales e incrementando el número de rondas y patrullajes. El sistema de predicción de hechos delictivos cumplió con su objetivo principal (de mejora del proceso), sin embargo; existen otras limitaciones que poseen los efectivos policiales, como por ejemplo; la toma de denuncias, registro de incidencias, trabajando separado al serenazgo y CSI (entidades que reciben el apoyo de la municipalidad de La Molina).

Palabras claves: Prevención del delito, sistema de predicción, minería de datos, zonas de riesgo.

ABSTRACT

This project, called crime predictive system for improve the prevention crime process in La Molina district using data mining, has as an objective improve the process related to prevent crime in 3 police station into this district, making emphasize in crime with intention to get the main objective of the police station: keeping and making order around the zone.

Thanks to new IT technologies, and specially thanks to data mining, it Is possible obtain predictions of crime which could be possible to pass in the future based on registers of historic data and with this, develop a solution which show that information and generate knowledge in the process to prevent crime. First to all in Santa Felicia police station.

Data mining modeling uses historic information about complaint in each police station from 2015, and based to an algorithm of machine learning, give us the most prone zone to occur crime. For this application, it decided show the information in thematic maps for this can be accessed by any device (web o mobile)

This solution helped to policeman to identify risk zones, assign police resources, and increase rounds and patrols. The system of crime prediction did its main ibjetive (improve the process), however; the police has other limitations, for instance: take the complaint, registering issues, working separate to “serenazgo” and CSI (units which get support to municipality)

Key words: Prevent the crime, prediction system, data mining, risk zones

INTRODUCCIÓN

Uno de los grandes problemas que afecta al país en los últimos años es la inseguridad ciudadana. A lo largo de los años se ha visto cómo ha ido aumentando el número de delitos y la impunidad con que son realizados, convirtiéndose en un problema a nivel nacional.

El distrito de La Molina, como parte del territorio nacional no es ajeno a esta realidad, tal es así que en base a información recogida en la comisaria de Santa Felicia, las denuncias en el distrito son mayores en comparación al año pasado.

Esto da una primera indicación que el proceso de prevención del delito en el distrito se ve debilitado debido a diferentes factores, pero sobre todo a la falta de rondas, patrullajes o redadas realizadas de manera rápida por parte de los efectivos policiales y en ocasiones a la falta de recursos existentes en las comisarías.

La minería de datos, como tecnología para analizar grandes volúmenes de datos orientado a la predicción de sucesos, ha sido muy utilizada en los últimos años (Caridad, 2014) en problemas de índole social, tales como predicción de desastres natrales, predicción del tiempo, predicción del crimen, etc.

En la presente tesis se ha realizado un análisis en base a los registros históricos de las denuncias ocurridas en el distrito de La Molina en el año 2015 para poder crear un modelo de predicción de hechos delictivos que se pueda mostrar en un sistema de información fácil de usar para los policías y así ellos puedan mejorar sus tomas de decisiones en lo referido a la prevención del delito.

Esta investigación ha sido dividida en cinco capítulos, los cuales se detallan a continuación:

El Capítulo I llamado Marco Teórico explica los antecedentes de la investigación considerando artículos o tesis relacionadas a nuestra área temática, donde se muestran algunas investigaciones utilizando algoritmos de aprendizaje automático que sirvieron como herramienta de ayuda mejorar algún problema social en la localidad. Adicionalmente se explica qué es la minería de datos y cómo esta tecnología puede ayudar trabaja para tener predicciones con cierto grado de probabilidad.

El Capítulo II llamado Metodología especifica los recursos y herramientas que utilizaremos para la realización del Sistema de predicción, que comprende los principales hardware, software y talento humano; cabe mencionar que la elección de la metodología se plasma en este capítulo y se determina mediante la comparativa entre las dos más usadas.

El Capítulo III llamado Desarrollo del Proyecto comprende la serie de pasos realizados para la creación del Sistema de Predicción, y que toma como base de desarrollo las distintas fases descritas por la metodología (elegida n el capítulo 2) y que están sujetas al cronograma de actividades, en nuestro caso se pueden diferenciar dos puntos claves, la primera concerniente a la creación del modelo de predicción y la segunda el formato de presentación del sistema a través de un Sistema de predicción accesible vía web y móvil.

El Capítulo IV llamado Pruebas y Resultados valida el correcto funcionamiento del sistema en base a distintas pruebas relacionadas al desarrollo de software, entre ellas las principales son de funcionalidad, integridad y rendimiento

El Capítulo V llamado Discusiones y Aplicacione**s** muestra la comprobación de las hipótesis en base a pruebas estadísticas considerando las variables detalladas en la metodología de investigación.
Para finalizar el documento se mencionan las Conclusiones y recomendaciones de los miembros del equipo de proyecto como resultado de la investigación.

1. Situación problemática

Actualmente en nuestro país, la percepción de seguridad ciudadana por parte de la mayoría de peruanos es calificada como baja o nula, esto debido a que la delincuencia se ha convertido en el principal tema de preocupación tanto de los ciudadanos mismos como de las autoridades (Fernández y Rivas, 2015). Para comprobar esto sólo basta con ver las noticias sobre lo que ocurre diariamente en las calles y darse cuenta que esta percepción es real.

Adicionalmente a lo dicho, en los últimos años se han ido realizando encuestas de victimización a nivel latinoamericano que demuestran que las ocurrencia de delitos con dolo aumentan considerablemente en nuestro país, y que son mucho más grande a comparación de los demás países en América Latina (El Comercio, 2015). Estos delitos abarcan las diferentes modalidades de robos, toda clase de hurto, secuestro, extorsiones, homicidios, entre otros.

Estos aumentos en las denuncias en el país reflejan que los planes de prevención del delito están siendo elaborados de manera incorrecta, desencadenando la realidad que se vive hoy en día.

Si bien los delitos han aumentado al pasar de los años también lo ha hecho la tecnología, y gracias a ello la fuerzas del orden han podido mejorar sus acciones ante esta plaga. En Turquía Fatih y Bekir (2015), oficiales de la Policía Nacional de Turquía, afirman que gracias al uso de las TI han podido combatir el crimen Organizado de su país, gracias a elementos como Sistemas de Geo localización (GIS), reconocimiento de voz, reconocimiento facial y Sistemas de CCTV de forma integrada.

De igual manera en marzo del 2012 se decretó la ley de Geo localización en México, que permitía a las autoridades mexicanas a poder localizar de manera inmediata cualquier número de celular o telefónico, gracias a esto se pudo frenar delitos como extorsiones y sicariatos desde las cárceles y por

consecuencia prevenir y disminuir estos delitos con ayuda de las nuevas TI y decretos legislativos (FayerWayer, 2012).

Por otro lado según Valera (2012) afirma que es posible establecer patrones de comportamiento delictivo en base a registros históricos para así establecer diversas acciones preventivas en base a la acción policial.

Adicionalmente, el autor Perversi (2007) hizo una investigación basada en minería de datos para explorar y detectar patrones delictivos en Argentina, dicha investigación abarcó todas las provincias de dicho país, logrando explicar tendencias y predicciones de delitos que podrían venir al año siguiente.

De la misma forma, Martínez (2009) plantea una investigación en México D.F con el objetivo de realizar predicciones espacio-temporales de patrones delictivos, esto lo hace con ayuda de algoritmos de aprendizaje automático y centra su alcance en la criminología que vivía México en el año mencionado. Abarca tanto hechos delictivos como tentativas de la realidad nacional de dicho país.

Otros autores que se pueden citar son Henderson, Wolfers y Zitzewitz (2008) quienes describen, en un artículo como resumen de su investigación en la universidad de chicago, los mercados de predicción que hay actualmente detallando que son bien entendidos, pero adicionándole ideas con implicancia al crimen. Así finalmente enfatiza su investigación en analizar el uso de los mercados de predicción actuales para pronosticar la tasa de delincuencia.

Gibbs (2014) plantea una solución en Londres basada en Big Data para predecir el crimen, esto es basado en técnicas de análisis predictivo y haciendo uso de datos de teléfonos celulares para predecir la geográficamente el crimen. El resultado de este trabajo es un mapa donde muestra puntos rojos en Londres, que significa la ocurrencia de hechos delictivos.

Entrando un poco más al análisis de nuestro país, Según El INEI (2014) y en base a una Encuesta Nacional de programas estratégicos realizado el mismo año y que abarcó principalmente la seguridad ciudadana, se pudo recoger la visión de personas mayores a 15 años para conocer si han sido víctimas de algún hecho delictivo.

El módulo de Seguridad Ciudadana tiene como cobertura geográfica el área urbana a nivel nacional, siendo el tamaño de muestra anual de 28 mil viviendas particulares.

Esta investigación, realizada a finales del año 2014 muestra como resultados la curva de delitos cometidos mostrando como punto más resaltante que la inseguridad ciudadana va en aumento cada día.

Adicionalmente en el informe también se muestra a nivel nacional urbano, los principales motivos por los cuales la población de 15 años a más que fue víctima de algún hecho delictivo no denuncia. El 27,6% evita poner una denuncia porque desconoce al delincuente y el 27,2% porque consideran que es una pérdida de tiempo.

A diferencia de las otras ciudades de más de 20 mil habitantes donde el 28,5% de la población no denuncia porque consideran que es una pérdida de tiempo, el 27,1% afirma que no lo hace porque desconoce al delincuente, mientras que a nivel de centros poblados urbanos un 29,6% tiene como principal motivo el desconocer al delincuente, según los resultados del último semestre en análisis.

La investigación cierra analizando la percepción de las personas en lo referido a la labor de las autoridades de su respectivo distrito frente a la delincuencia, donde lo más resaltante es la baja confianza que tiene la población a la policía y a las autoridades de la zona, pues en su mayoría, las personas se sienten inseguras.

Si se centra un poco más el análisis y el enfoque va estrictamente al distrito de La Molina, podremos ver que este distrito sigue la línea de la tendencia que ocurre en la ciudad y en el país (INEI, 2015), es decir, el incremento de denuncias.

1.1 Definición del problema

En el distrito de La Molina desde el año 2012, se empieza todo un nuevo programa para combatir el problema de la inseguridad ciudadana en su jurisdicción, debido al aumento del aumento del número de denuncias.

Ante este aumento de criminalidad del distrito, y con los recursos limitados que poseen, las tres comisarías actuales (Santa Felicia, Las Praderas y La Molina) que en promedio tienen 60 efectivos policiales y 5 patrulleros por comisaria; por ende nace una problemática relacionada a la prevención de delitos. Pues las población del distrito es muy grande comparado con la cantidad de recursos disponibles que se manejan.

El proceso de prevención de delitos que actualmente existe en las comisarías, se basa esencialmente en revisar las denuncias que llegan a cada comisaria y determinar las zonas poco resguardadas de cada una de sus zonas, todo este estudio no tiene un sustento estadístico-probabilístico claro y cada comisaría vela por su propia jurisdicción, aquí se ve un claro problema de comunicación a nivel organizacional, salvo que se realicen operativos grandes y soliciten apoyo de la Águilas Negras o La DIRINCRI.

Cabe mencionar que el CSI y el serenazgo de La Molina también son otros entes separados manejados por la Municipalidad de La Molina, pero estos no tienen la potestad de realizar intervenciones, lo que hace que tengan que informar a la PNP del distrito para que estos puedan realizar intervenciones, lo que genera una demora en los tiempos de acción.

1.2 Problema general

- ¿Se puede mejorar el rendimiento del proceso de prevención de delitos en las comisarías de la Molina utilizando minería de datos?

1.3 Problemas específicos

- ¿Es posible mejorar los tiempos de acción de los efectivos policiales en la identificación de zonas de riesgos en el distrito de la molina?
- ¿Es posible mejorar el servicio policial frente a la prevención de hechos delictivos en el distrito de La Molina?
- ¿Es posible estructurar y categorizar la información de las denuncias de las distintas comisarías del distrito de la Molina?

2. Objetivos

2.1 Objetivo general

Mejorar el rendimiento del proceso de prevención del delito de las comisarías de La Molina utilizando Minería de Datos.

2.2 Objetivos específicos

- Mejorar los tiempos de acción de los efectivos policiales en la identificación de zonas de riesgo en el distrito de La Molina.
- Mejorar el servicio policial frente a la prevención de hechos delictivos en el distrito de La Molina.
- Unificar y Estructurar la información que manejan las comisarías del distrito de La Molina que permita predecir la ocurrencia de hechos delictivos en base a datos históricos.

3. Justificación

3.2 Justificación teórica

- La utilización de la Minería de Datos, mediante algoritmos de aprendizaje automático, permiten crear patrones de predicción en

base a datos históricos, en este caso aplicando a la ocurrencia de hechos delictivos en el distrito de La Molina.

- Aplicar óptimamente la metodología de minería de datos (CRISP-DM) es clave para que las predicciones sean las correctas y precisas.

3.3 Justificación práctica

Este sistema beneficiaría a la municipalidad de La Molina, específicamente a cada una de las comisarías, ya que la aplicación móvil les ayudará a realizar rondas preventivas y además a elaborar planes de contingencia según las zonas más riesgosas.

3.4 Justificación social

La implementación de este sistema causaría un impacto social, ya que beneficiaría directamente a los efectivos policiales que operan en las distintas comisarías de La Molina, otorgando un mejor servicio policial y así causar un impacto favorable en la seguridad ciudadana.

4. Alcance

La presente tesis está determinada para desarrollar un modelo de predicción de delitos en el distrito de La Molina, el cual utilizará una técnica de aprendizaje automático que será determinada según una comparación entre 2 o más algoritmos, de la misma manera se evaluarán diversas herramientas disponibles en el mercado para determinar cuál es la que más se acomoda a nuestra necesidad.

Al final, este modelo será implementado como un sistema de información, el cual mostrará en un mapa, la probabilidad de ocurrencia de distintos hechos delictivos, estos serán debidamente categorizados por zonas y se podrá realizar un filtro de búsqueda de delito. Al entrar al detalle de cada zona se podrá visualizar la fecha, rango de hora, tipo de lugar y probabilidad de ocurrencia de un hecho delictivo.

Como funcionalidad adicional, el sistema mostrará una sección de reportes donde se muestran gráficos de barras con las cantidades de denuncias

según zonas, modalidades, comisarias, fechas, etc. en base a las denuncias históricas que se han registrado en las comisarías del distrito.

La solución estará disponible en aplicativos móviles que tengan como sistema Operativo Android y las descargas se realizarán a los equipos móviles que posean los efectivos policiales de las comisarias del distrito.

Adicionalmente también se plantea la solución en versión web con las mismas funcionalidades descritas.

5. Hipótesis

5.1 General

La aplicación correcta de la minería de datos se logrará mejorar el rendimiento del proceso de prevención de delitos del distrito de La Molina, la automatización y/o simplificación de las actividades que se realizan en el proceso actual.

5.2 Específicas

- La implementación de una aplicación móvil que muestre un mapa demarcado por zonas donde se pueda apreciar los posibles delitos a ocurrir, podrá mejorar los tiempos tiempo de acción de los policías al poder identificar zonas de riesgo de una manera más rápida.
- La implementación de una herramienta de predicción de hechos delictivos en el distrito de La Molina logrará mejorar el servicio policial en base a la realización de acciones preventivas por zonas.
- La implementación de una herramienta de predicción de hechos delictivos en el distrito de La Molina logrará mejorar el servicio policial en base a la realización de acciones preventivas por zonas.

CAPÍTULO I
MARCO TEÓRICO

1.1 Antecedentes

Es esencial revisar investigaciones y documentos relacionados al área temática, para así verificar el funcionamiento de sistemas ya consolidados, que hayan resuelto un determinado problema de la sociedad ya que esto va a permitir tener un panorama general sobre el área temática a desarrollar y así alcanzar los objetivos que fueron señalados previamente.

1.1.1 Aplicación de minería de datos para la exploración y detección de patrones delictivos en argentina

Esta investigación realizada en Argentina, estuvo referida a la implementación de minería de datos en base a información histórica de hechos criminales que ocurrían en el mencionado país y a la vez comprobar su efectividad del mismo.

El autor Perversi (2007) resume su investigación en la identificación y detección de patrones de homicidios dolosos cometidos en Argentina durante 2005 en base a información suministrada por la Dirección Nacional de Política Criminal del Ministerio de Justicia y Derechos Humanos de la Nación, organismo encargado de realizar las estadísticas oficiales de criminalidad en Argentina. Se puso énfasis en los registros criminales como valor fundamental para la prevención del delito.

Otro objetivo destacado está relacionado al diseño de políticas y planes de prevención a realizar, ya que en Argentina este tipo de análisis se había realizado históricamente mediante herramientas estadísticas descriptivas básicas, considerando básicamente variables y relaciones primarias.

En general, la tesis abarcó el descubrimiento de patrones delictivos y la aplicación de medidas preventivas para atacar los mismos.

1.1.2 Caracterización y predicción espacio temporal de patrones delictivos mediante modelos lógicos combinatorios.

Martínez (2009) afirma que su investigación:

> Se basa en técnicas de aprendizaje inductivo y busca hacer un análisis y predicción delictiva en México D.F. El análisis delictivo se realiza mediante la construcción de definiciones inductivas para cada una de las familias delictivas que se plantean en la investigación; y, en base a estas definiciones se logra determinar el comportamiento de la actividad delictiva dentro de un espacio-tiempo seleccionado.
>
> Luego del análisis, se llega a predecir la cantidad de hechos delictivos esperados para cada familia de hechos delictivos.
>
> Los resultados son orientados como ayuda a las autoridades de seguridad pública, para que estos mejoren en temas de prevención y asignación de recursos humanos y financieros. (p.15)

1.1.3 Aplicando minería de datos al marketing educativo

Cadena (2011) nos muestra en su investigación realizada en Colombia que la minería de datos puede abarcar distintos campos de estudio, en este caso, enfocándose al sector educativo; donde en base al análisis de los registros históricos de los estudiantes se pudo establecer la tasa de deserción de los estudiantes y de los factores que influyen en la misma, para ello tomaron como muestra a un grupo de estudiantes de la Escuela de Marketing y Publicidad de la Universidad Sergio Arboleda.

Al finalizar el estudio se concluyeron muchos aspectos resaltantes, entre ellas; que la mayoría de alumnos que desertaban eran de sexo femenino y que además pertenecían a los 3 primeros ciclos de la carrera, quienes alegaban que el motivo era el bajo rendimiento académico y problemas económicos. Otro factor importante es que la mayoría de ellas vivía cerca de la ciudad de Bogotá, y sus edades oscilaban desde los 19 a los 22 años.

Con respecto a los varones, estos generalmente desertaban entre los primeros dos ciclos con un promedio de edad de 19 años, el motivo principal era el bajo rendimiento académico y la falta de atención en el Idioma Inglés.

Otros factores que también se analizaron fueron la proveniencia de los alumnos, registros de notas de colegio, nivel cultural, deportes practicados y gracias a esta información la Universidad supo tomar mejores decisiones para posicionar sus carreras en el sector social adecuado.

1.1.4 Aplicando minería de datos en el cuidado de la salud de la diabetes en pacientes jóvenes y viejos

El informe realizado por Salman (2015) perteneciente a la Universidad de King Saud refiere que:

> Gracias a la minería de datos se pudo realizar un análisis profundo y determinar que personas son proclives a sufrir este mal, para ello se registraron los datos de los pacientes enfermos, su edad, planes de tratamiento, sexo, etc. La muestra fue de alrededor 300 pacientes, donde se les hizo seguimiento por alrededor de dos años (p.22).

Este estudio se realizó en Arabia Saudita, que según datos del mismo informe tuvo una gran alza de personas que sufrían de este mal.

Los resultados del informe de Salman (2015) concluyeron que:

> Las personas de mayor edad tenían un factor común en sus diversos estilos de vida mientras que en el caso de los más jóvenes la causa principal era la obesidad, el estudio también plantea ciertos planes de tratamiento para cada tipo de paciente y los controles que estos debían llevar, esto también fue basado en la tasa de éxito de pacientes que se recuperaron y/o lograron alargar su vida (p. 23).

1.1.5 Redes neuronales artificiales en la predicción de insolvencia. Un cambio de paradigma ante recetas tradicionales de prácticas empresariales

Según Gelmar y Lastre (2014) quienes realizaron una:

> Revisión y análisis de las principales teorías y modelos que abordan la predicción de la insolvencia y quiebra empresarial. Las redes neuronales fueron tomados como un instrumento debido a su reciente aparición y además porque en los últimos años han recibido considerable atención por parte del mundo académico y profesional, y ya empiezan a implantarse en diversas organizaciones modelos de análisis de la insolvencia basados en la computación neuronal. El objetivo de la tesis fue arrojar evidencias de la utilidad de las Redes Neuronales Artificiales, en la problemática de predicción de insolencia o quiebra por lo cual se comparó su capacidad predictiva con la de los métodos utilizados habitualmente en dicho contexto.
>
> El significado de la expresión riesgo de insolvencia para una empresa infiere que experimenta dificultades financieras evidenciando esto cuando se observan hechos que caracterizan a un “estado de cesación de pagos” como incumplimiento de sus obligaciones o interrupción en el pago a los proveedores. De igual manera dicho estado es nocivo para cualquier entidad, debido a que son desviadas de sus objetivos principales que son la creación de valor y la maximización de sus utilidades. Si esta situación perdurase en el tiempo, la empresa se puede declarar en quiebra (p.18).

Las conclusiones del estudio apuntaron a que se pueden lograr altas capacidades predictivas empleando las redes neuronales artificiales u otro modelo de aprendizaje automático que cuenten con variables cualitativas y cuantitativas.

1.1.6 Usando redes bayesianas para predecir defectos de software.

Según Fenton etl al. (2007), quienes hicieron una revisión del uso de redes bayesianas en la predicción de defectos de software y la fiabilidad del mismo, afirman que el este enfoque permite a los analistas incorporar factores del proceso causal al igual que al combinar las medidas cualitativas y cuantitativas, por lo tanto, la superación de algunas de la conocidas limitaciones de los métodos tradicionales de métricas de software. El

enfoque ha sido utilizado y reportado por organizaciones como Motorola, Siemens, y Philips. Pero por otra parte, uno de los impedimentos para un uso más generalizado de las Redes Bayesianas para este tipo de aplicación fue que algoritmos de este tipo sufrían de un punto débil muy visible, los cuales estaban relacionados a que no eran capaces de manejar nodos continuos correctamente. Lo que forzó a los modeladores a tener que definir previamente intervalos de discretización por adelantado y resultados en predicciones inexactas donde el rango de, por ejemplo, los recuentos de defecto fue grande.

Finalmente, los recientes avances en los algoritmos Redes Bayesianas han hecho posible para llevar a cabo la inferencia en Redes Bayesianas con nodos continuos en los últimos tiempos, sin la necesidad de pre especificar niveles de discretización. El uso de algoritmos de discretización dinámico resulta una mejora en exactitud de los tipos de modelos de predicción de defectos y fiabilidad.

1.1.7 Aplicación de las redes bayesianas dinámicas a la predicción de series de datos y a la detección de anomalías

Según Reguero (2010) quien muestra un estudio de:

> Las redes bayesianas dinámicas aplicadas a la detección de anomalías y a la predicción en series de datos. Primero se da una breve noción general sobre las redes bayesianas: sus inicios, sus tipos y sus usos. Entre los diferentes tipos de redes bayesianas existentes en el estado del arte, se específica el caso de las híbridas en el capítulo 2, interesantes por su capacidad de operar con cualquier tipo de dato, tanto cualitativo como cuantitativo. En este capítulo se describe la teoría de las gaussianas condicionales, que permiten tener un modelo en el que se combinen variables discretas y continuas sin necesidad de llevar a cabo la discretización de estas últimas.
>
> Luego se incorpora la noción de ‘tiempo a la red’. Cada instante se representa por una distancia de la red y entre dichas instancias se

> realizan las conexiones que representan el flujo temporal de la información. De esta forma se añade información de estados anteriores en el tiempo al modelo. Una vez definida la estructura general de la red se introducen los algoritmos de entrenamiento utilizados para llevar a cabo la inferencia sobre la estructura propuesta.
>
> El más relevante es una modificación del algoritmo 'Forward' propuesto por Rabiner, que nos permite calcular la probabilidad de los diferentes estados de las variables ocultas dada una secuencia temporal de datos (p. 35).

En la parte final se aplican las redes bayesianas híbridas y dinámicas, estudiadas durante la tesis, a la creación de un modelo para realizar el control de la medición de temperatura en estaciones del bosque Andrews Forest de Oregon (USA). Al terminar estos pasos, se valida la implementación realizada de la teoría introducida con los experimentos y finalmente, el modelo creado se usa para detectar anomalías en los datos recibidos de los sensores de temperatura.

1.1.8 Clasificación supervisada basada en Redes Bayesianas. Aplicación en biología computacional

Según Robles (2003) quien detalla una tesis basada en:

> El estudio de Redes Bayesianas donde realiza una clasificación dentro de dos grandes campos que son la clasificación supervisada con modelos gráficos probabilísticos y su aplicación a la biología computacional.
>
> La idea fundamental de las propuestas que se han realizado dentro del campo de la clasificación supervisada con modelos gráficos probabilístico, es el uso de los algoritmos heurísticos de optimización EDA en la búsqueda de estructuras de redes Bayesianas para clasificación.

Gracias a la aplicación de los algoritmos EDA, se ha desarrollado un nuevo algoritmo de clasificación supervisada denominado Interval Estimation naíve-Bayes y se han mejorado varios de los algoritmos de clasificación propuestos en la literatura. Los resultados experimentales obtenidos han sido muy satisfactorios, ya que demuestran la superioridad de nuestra idea.

Además, con el objetivo de mejorar su rendimiento, se ha desarrollado una versión paralela de nuestro algoritmo, el Parallel Interval Estimation naíve-Bayes.

Las pruebas experimentales han superado nuestras expectativas iniciales, ya que no sólo se ha logrado un speedup superlineal, si no que se han obtenido mejores resultados que en la versión secuencial.

En el campo de la biología computacional la predicción de la estructura secundaria de las proteínas es de vital importancia, ya que proporciona un punto de partida para la predicción de su estructura tridimensional, lo cual ayuda a la determinación de sus funciones.

Dentro de este campo, se ha estudiado la aplicación de los métodos de clasificación supervisada en dos niveles diferentes. Por un lado, se ha desarrollado un nuevo método basado en redes Bayesianas, para la predicción de la estructura secundaria de las proteínas.

Aunque en primera instancia los resultados obtenidos no han sido brillantes, en esta tesis se sugieren refinamientos de la idea original que, confiamos, los mejorarán.

Por otra parte, se ha creado un multiclasificador con los métodos de predicción existentes, basado en el paradigma stacked generalization. Los resultados obtenidos por este multiclasificador han sido altamente satisfactorios, ya que se han mejorado los resultados de los métodos individuales. Como resultado de las propuestas realizadas, han

surgido multitud de futuras líneas de investigación, que se recogen a lo largo de esta tesis (p.43).

1.1.9 Aportaciones de las redes bayesianas en meteorología. Predicción probabilística de precipitación

El autor Ancell (2003) recopila informes, referidos principalmente a:

> Investigadores interesados en la aplicación de técnicas de minería de datos en meteorología y otras ciencias medioambientales afines, aunque algunas de las innovaciones también son susceptibles de aplicaras a otras áreas científicas, siempre que traten de sistemas definidos por muchas variables, cuyas relaciones de dependencia no seas conocidas a priori y que tengan que ser inferidas a partir de un conjunto de datos que describan el sistema.
>
> El conocimiento hallado se basa en un intercambio entre meteorólogos y matemáticos, en el afán de aplicar los recientes desarrollos producidos en el ara de la minería de daros a problemas prácticos relacionados con el diagnóstico y la predicción probabilística local en meteorología, considerando el problema de la coherencia espacial.
>
> En concreto el eje central de esta tesis es el desarrollo de modelos gráficos probabilístico, en particular Redes Bayesianas (RBs), para su aplicación en la predicción probabilística local, aunque también se ha realizado otras aportaciones interesantes.
>
> Hay que decir también que esta tesis trata sobre un tema que es considerado como punto de partida en el campo de la predicción, por ello se busca tener una herramienta escíticamente para una zona que dé predicción probabilística solo para esa localidad (p. 20).

1.1.10 Redes bayesianas temporales: Aplicaciones médicas e industriales

Según Severino (2009) cuya tesis abarca el desarrollo de un sistema que capaz de modelar la evolución de la extensión de un cáncer de nasofaringe. Este sistema sirve de ayuda a los oncólogos de una unidad de oncología radioterápica en la determinación del desarrollo alcanzado por este tipo de cáncer en un paciente antes de aplicar la terapia adecuada.

En la primera parte se muestra el marco conceptual de modelado de conocimiento en que fue realizado el trabajo. Debido a la naturaleza de los procesos que tienen lugar en el dominio médico del que ocupa el autor, se basó en el uso de redes bayesianas con representación explıcita del tiempo para la definición del modelo de crecimiento cancerígeno.

En la segunda parte, se revisan los diferentes tipos existentes de redes bayesianas para razonamiento temporal y analizamos las ventajas e inconvenientes de cada uno. Luego se formaliza un nuevo método que lo denominan red de eventos probabilistas en tiempo discreto, el cual resulta adecuado para el modelado de los mecanismos causales de carácter incierto a los que se ven sujetos un conjunto de eventos temporales. Como principal aportación del nuevo método se destaca el uso de diferentes modelos temporales de interacción causal para cada familia de nodos de la red. Estos modelos representan una adaptación para procesos temporales de los modelos canónicos probabilísticos tradicionales. Por esta razón se decide denominarlos puertas probabilistas temporales. También se incluye la aplicación de un nuevo algoritmo que permite la factorización de las probabilidades condicionales correspondientes a familias de nodos que interactúan según un modelo de puerta probabilista temporal.

Al finalizar la investigación se describe el sistema NasoNet, el cual aplica al dominio de cáncer de nasofaringe el nuevo método para razonamiento temporal con redes bayesianas desarrollado en el capítulo previo. Por otra parte, se muestra el carácter general de las redes de eventos probabilistas en tiempo discreto a través de su aplicación en un dominio industrial: el

diagnóstico y predicción de los fallos que ocurren en un generador de vapor de una central termoeléctrica.

1.1.11 Uso de redes neuronales para la Asistencia de voz en base a aplicaciones de Minería de Datos

Según Engels y Theusinger (2015) quienes describen un proyecto realizado en las inmediaciones del país de Cataluña España donde en base a datos históricos se ha podido crear un asistente de voz inteligente, este actualmente es en vías de culminación ya que uno de los principales problemas que se enfrentó fue en la realización de testeos, y la denominada ley de Turing, el objetivo de este test es verificar y comprobar si un ente artificial es capaz de simular el comportamiento humano en este caso su raciocinio.

Es asistente de voz en base a distintos patrones de reconocimiento de Voz e Inteligencia Artificial es capaz de asistir a diversos tipos de dudas de los usuarios, este proyecto se ha realizado para que el asistente sepa contestar a las distintas temáticas con las que interactúe, actualmente se sigue en la etapa de pruebas, y en esta investigación, se enfocan en hacer referencia a las tecnología que utilizaron para desarrollar este proyecto.

1.1.12 Desarrollando aplicaciones de minería de datos acopladas a sistemas gestores de bases de datos relacionales

Según Agrawall y Shim (2015) quienes mencionan que uno de los grandes avances en la tecnología fue la creación de repositorios integrados en nuestro caso popularmente denominados bases de Datos, en su principal oficio estas almacenan información y han servido de muchas formas para la toma de decisiones y un gran ahorro en trabajos manuales muchas veces.
Realizar una aplicación de Minería de datos no es algo sumamente difícil si es que se siguen los principales principios y distintos patrones de desarrollo, estos no se encuentran altamente relacionado a un determinado lenguaje de programación ya que trabajan de forma independiente y más se enfoca en el algoritmos de análisis de datos, este trabajo menciona las buenas prácticas

de desarrollo como son CRISP-DM y de diversas herramientas de minería de datos que usaron para su desarrollo de aplicaciones.

1.1.13 Mecanismos de aprendizaje y minería de datos

El autor Mitchell (2013) detalla en su investigación que:

> Los distintos tipos de algoritmos de aprendizaje que existen para el análisis de datos y que se han surgido en las últimas décadas. Si bien muchas de las técnicas estadísticas ya empleadas en la actualidad ya existían, gracias al avance de la tecnología en especial del hardware que son capaces de realizar cálculos estadísticos sumamente complejos que al realizarlo humanos simplemente sería demasiado laborioso, y como esta gran cantidad de información se ha ido aumentado año a año teniendo esta nueva tecnología de Minería de Datos, donde se tiene una gran cantidad de información que tal vez no es tan relevante pero que dentro de todo ese universo existen pequeñas informaciones de gran valor significativo y que permitan tomar decisiones anticipadas (p.14).

1.2 Bases teóricas

1.2.1 Seguridad ciudadana

El término seguridad ciudadana como binomio gramatical fue incorporado en nuestro país en la Constitución de 1993.

Según Murazzo (2014):

> La inseguridad ciudadana es un concepto dinámico que está vinculado con la condición de habitantes de una ciudad tanto en áreas urbanas como rurales. Se subraya, al mismo tiempo, que la inseguridad no se restringe a las ciudades.
>
> La inseguridad ciudadana está esencialmente referida a los efectos generados de manera directa por los riesgos y peligros físicos existentes en la comunidad y la delincuencia en sus distintas formas; por tanto, en esta etapa también se deben incorporar aspectos de

inseguridad ciudadana producidos por fenómenos naturales o conmoción civil por causas políticas. (p.23).

Además de ello se ha visto que el aumento de criminalidad en nuestro país se ha ido agravando año tras año, prueba de ello es un informe de investigación realizado por el Instituto de Estudios peruanos en el 2014 donde se cataloga al Perú como el país que tiene la más alta tasa de delincuencia entre todos los países de Latinoamérica.

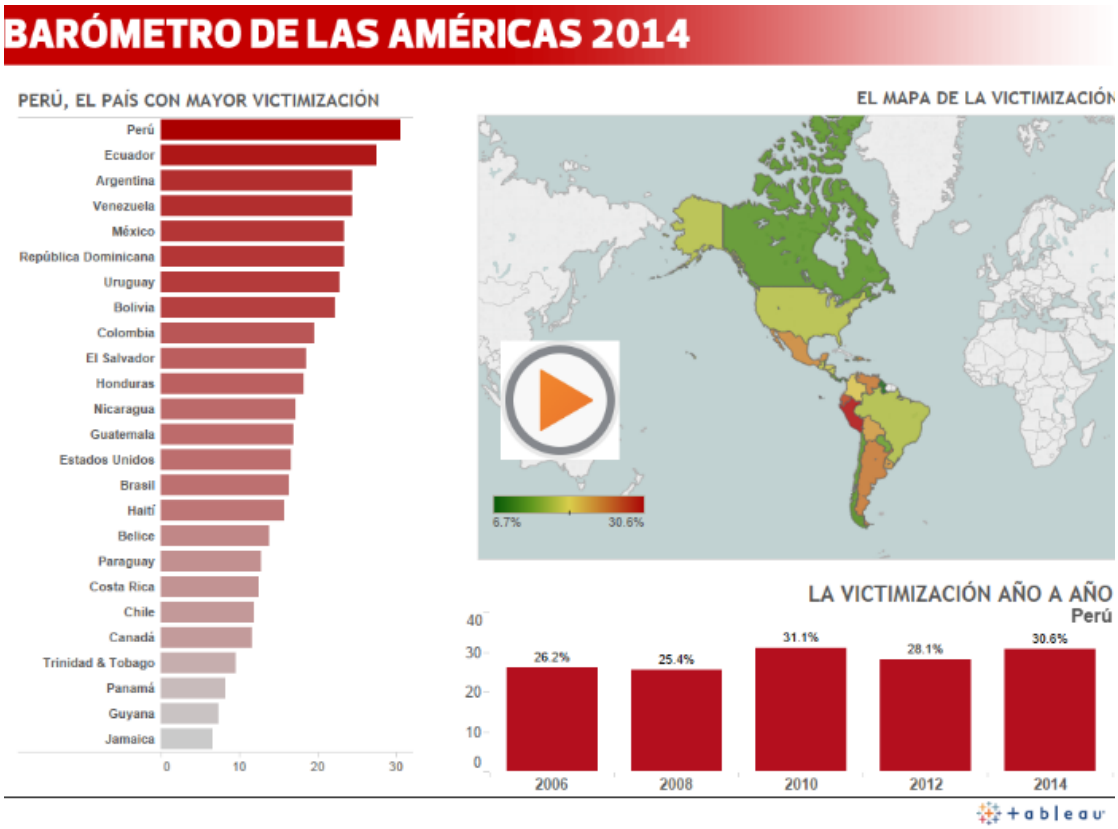


Figura 1. Tasa de Criminalidad en el Perú
Fuente: (Instituto de Estudios Peruanos, 2014)

Actualmente en nuestro país, la delincuencia se ha convertido en el principal tema de preocupación de las autoridades y ciudadanos, esto debido a que las denuncias de delitos han ido incrementado en los últimos años.

El gráfico que se expone a continuación detalla la curva de incremento de denuncias cada año y adicionalmente se muestran las cantidades totales; en este, se puede apreciar claramente que la tendencia va en aumento.

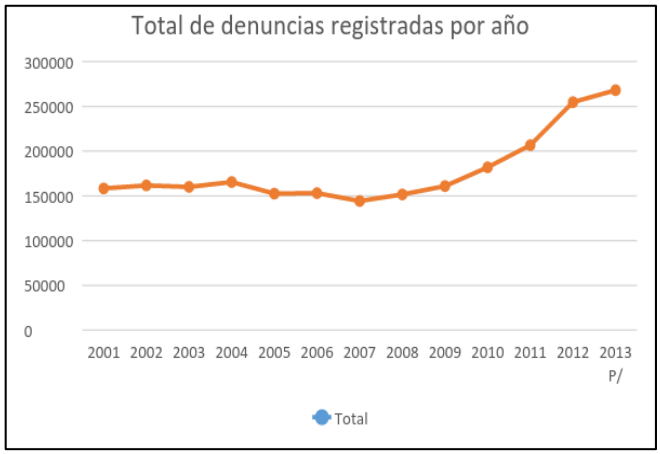


Figura 2. Denuncias registradas de 2001 al 2103
Fuente: INEI

Además del gráfico estadístico mostrado anteriormente, en el año 2014, el instituto Nacional de Estadística e Informática realizo un estudio en el cual recogió los tipos de delitos más comunes en Lima Metropolitana, los resultados se muestran en la figura siguiente.

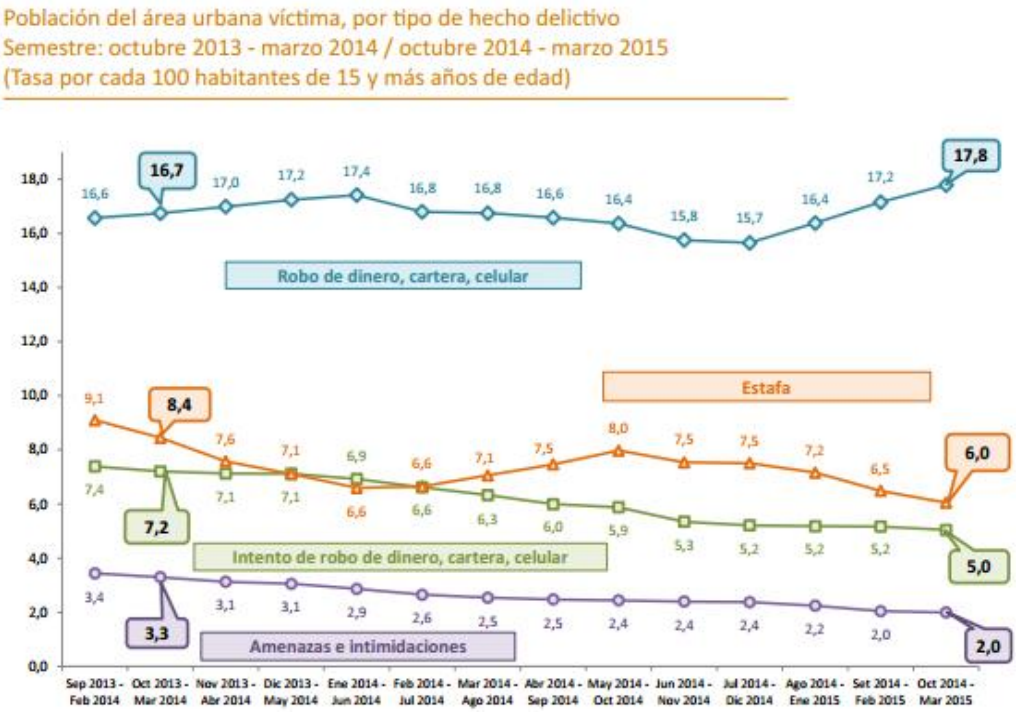


Figura 3. Cuadro de hechos delictivos 2014 – 2015
Fuente: (INEI, 2015)

Al analizar los gráficos podemos ver claramente un alza en los robos al paso entre los cuales se encuentran el robo de dinero y el de pertenencias

personales, además diariamente podemos ver en las noticias y periódicos que el nivel de salvajismo de los atracos se han agravado (El comercio, 2014), así que podemos inferir un deficiente desempeño hacia la prevención de delitos en Lima

Para continuar con el análisis, la siguiente tabla muestra la cantidad de denuncias registradas en el Distrito de La Molina en el año actual. Sólo en la comisaría de Santa Felicia se han registrado alrededor de 8000 denuncias de distinto tipo siendo el más común las denuncias especiales que hacen referencia generalmente a robos al paso a mano armada, mediante arma blanca y arrancamientos de cartera y celulares.

Tabla 1. Denuncias Registradas de Enero a Agosto 2015

DELITO	PENDIENTE	RESUELTA	TOTAL	PORCENTAJE
DENUNCIAS ESPECIALES	64	4,947	**5,011**	55.85%
INTERVENCION POLICIAL	168	1,558	**1,726**	19.24%
PATRIMONIO (DELITO)	133	880	**1,013**	11.29%
ACCIDENTES DE TRANSITO	429	307	**736**	8.20%
FALTAS	58	168	**226**	2.52%
VIOLENCIA FAMILIAR	32	102	**134**	1.49%
Otros	32	95	**127**	1.42%
TOTAL			**8,973**	**100.00%**

Fuente: (Comisaria de Santa Felicia de la Molina, 2015)

Analizando esto claramente podemos darnos cuenta que no existe un lineamiento claro hacia la prevención del delito, de lo contrario se vería una disminución ante estas considerables cifras. La investigación abarcará lo marcado en color rojo.

1.2.2 Minería de Datos

Según Mainmon, y Rokach (2010) la minería de Datos es un campo de las ciencias de la computación referido al proceso que intenta descubrir patrones en grandes volúmenes de conjuntos de datos. Utiliza los métodos de la inteligencia artificial, aprendizaje automático, estadística y sistemas de bases de datos. El objetivo general del proceso de minería de datos consiste en extraer información de un conjunto de datos y transformarla en una estructura comprensible para su uso posterior. Además de la etapa de análisis en bruto, que involucra aspectos de bases de datos y de gestión de datos, de procesamiento de datos, del modelo y de las consideraciones de inferencia, de métricas de Intereses, de consideraciones de la Teoría de la complejidad computacional, de post-procesamiento de las estructuras descubiertas, de la visualización y de la actualización en línea.

Los actuales modelos de procesos de minería de datos en sus marcos de trabajo se sustentan haciendo uso de un ciclo de vida, las mismas que van modelando el comportamiento de los datos y de los conocimientos, y así obtener un resultado que tenga significado en el mundo real (Valcárcel Ascencios, 2011). Según la figura 4 CRISP-DM (Cross – Industry Standard Process for Data Mining) es una de las metodologías más usadas.

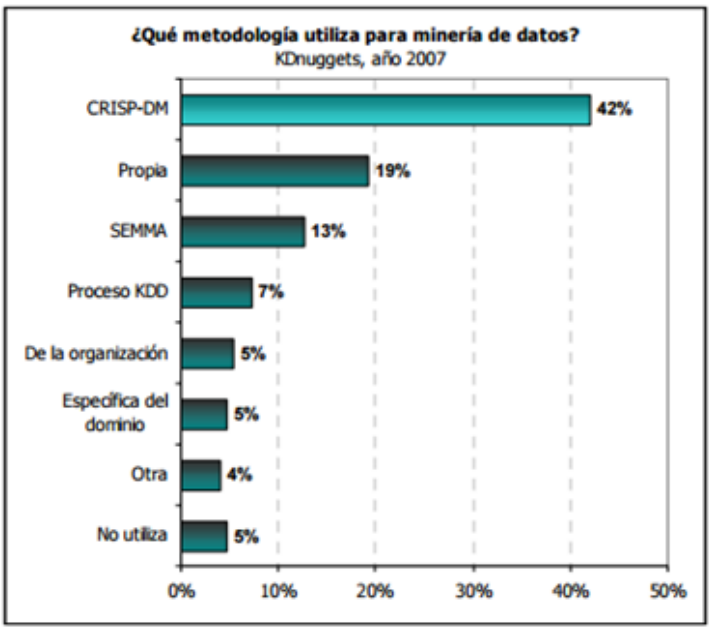


Figura 4. Metodologías utilizadas en Data Mining
Fuente: (Moine y otros, 2011)

En los apartados posteriores describiremos las dos metodologías más utilizadas en el mercado, esto con el objetivo de elegir la que más se adecue a nuestras necesidades. Resaltando criterios de selección y pesos a cada uno para poder llegar exitosamente a la elaboración del proyecto.

1.2.2.1 Evolución de la minería de Datos en las soluciones para la toma de decisiones

Según ARKSolutions (2014) en este documento informativo muestra el avance que se han tenido en cuestión a cantidad de información, contándonos la transición que ha ocurrido desde las principales bases de datos hasta los software altamente potenciales que se tienen hoy en día, y los grandes aportes que se han hecho y han beneficiado a la resolución de distintos problemas de la sociedad básicamente a la toma de decisiones. En muchas de estas se hicieron para un beneficio propio (toma de decisiones financieras) o para un impacto social (desarrollo ecológico, estadístico) tal como lo refiere en el informe.

1.2.3 Metodología de minería de datos – CRISP DM

La metodología CRISP-DM estructura el ciclo de vida de un proyecto de explotación de datos en seis fases cuya sucesión no es rígida e interactúan entre ellas de forma iterativa durante el desarrollo del proyecto. Las flechas indican la dependencia más importante y frecuente entre las fases.

El círculo exterior simboliza la naturaleza cíclica de los proyectos de este tipo. Las fases son como sigue:

- **Fase 1. Comprensión del negocio**

  Se centra en la comprensión de los objetivos y requisitos del proyecto desde una perspectiva de la empresa, para convertirlos en objetivos técnicos y en un plan de proyecto. El conocimiento adquirido del negocio se convierte en un problema de Data Mining (DM). Las tareas que se realizan son: determinar los objetivos del negocio, evaluar la situación, determinar los objetivos del DM y elaborar un plan de proyecto.

• Fase 2. Comprensión de los datos

Comienza con una recopilación inicial de datos para establecer un primer contacto con el problema, familiarizarse con los datos, identificar la calidad de los datos y establecer relaciones evidentes que permitan definir la hipótesis de información oculta. Las tareas que se realizan son: recolectar los datos iniciales, describir los datos, explorar los datos y validar la calidad.

• Fase 3. Preparación de los datos

Competen las actividades de preparación de los datos iniciales para adaptarlos a las técnicas de DM, tales como: visualización, búsqueda de relaciones entre variables para la exploración. Las tareas que se realizan son: selección de datos, limpieza de datos, estructurar los datos, integra los datos y formatear los datos.

• **Fase 4. Modelado de datos**:

Diversas técnicas de modelado son seleccionadas y aplicadas y sus parámetros son calibrados a valores óptimos, por ejemplo análisis de regresión, redes neuronales, razonamiento basado en casos (RBC), etc.

Algunas técnicas tienen requerimientos específicos según las características de los datos y la precisión que se requiera. Las tareas que se realizan son: selección de la técnica de modelado, generar el plan de pruebas, construir el modelo y evaluar el modelo.

• Fase 5. Evaluación del modelo

Una evaluación detallada del modelo y la revisión de los pasos ejecutados para construir el modelo para asegurar que se han alcanzado los objetivos de negocio. Un punto importante es determinar si hay algún objetivo que no ha sido considerado. Las tareas que se realizan son: evaluar los resultados, revisar el proceso y determinar la implementación.

- **Fase 6. Despliegue o implementación**

El conocimiento o resultado adquirido debe ser documentado y presentado de manera comprensible para que el cliente pueda usarla. Sin embargo, dependiendo de la cantidad de requerimientos, está fase puede ser sencilla o compleja. Las tareas que se realizan son: elaborar un plan de implantación, elaborar un plan de monitoreo y mantenimiento, informar sobre el proyecto y la revisión del proyecto.

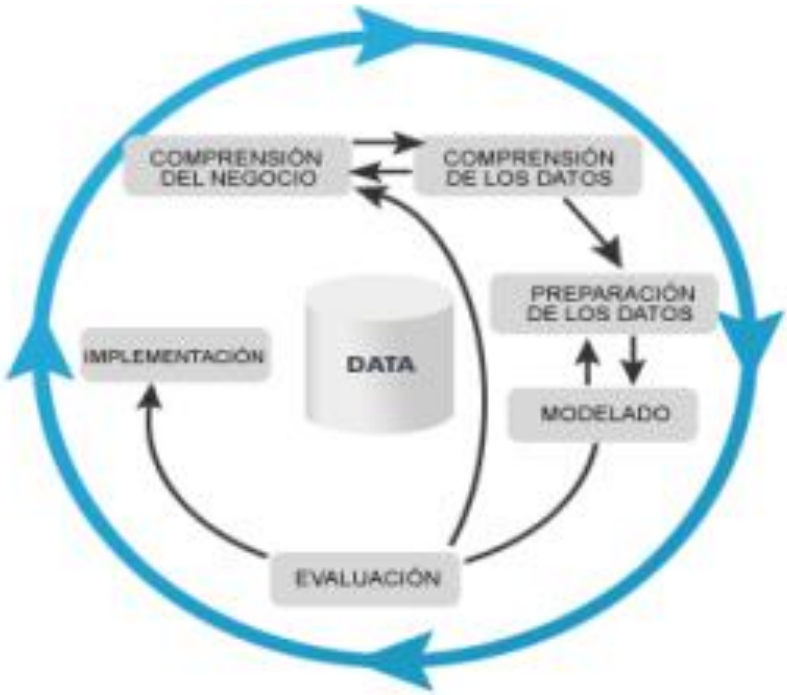


Figura 5. Ciclo de Vida del Modelo CRISP – DM
Fuente: (Chapman, y otros, 2000)

Estas fases que se identifican en el grafico anterior, a su vez se subdividen en actividades bien definidas (en alguno casos predecesoras unas de las otras, mientras que en otros, son independientes), que hacen posible un proyecto de minería de dados eficaz.

La figura 6 muestra las fases con sus respectivas actividades y posibles reportes o entregables finales según cada proyecto.

Algunos entregables podrían variar según el negocio y la solución de minería de datos que se esté desarrollando, pero las fases y actividades se deben respetar en cualquier caso si se sigue esta metodología.

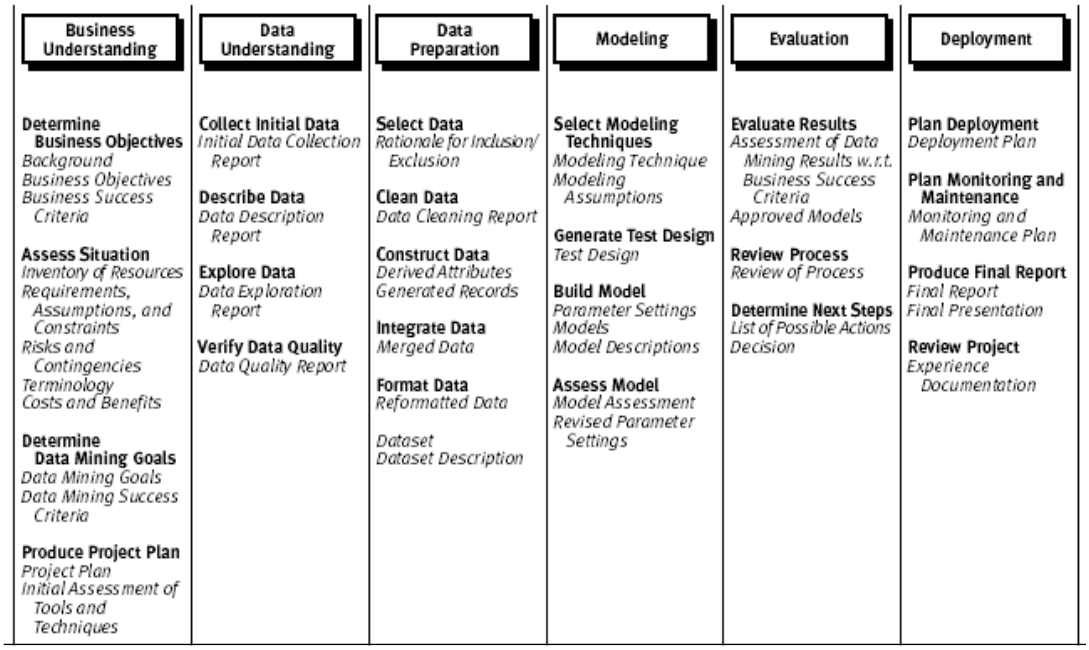


Figura 6. Actividades de la metodología CRISP-DM
Fuente: (Chapman y otros, 2000)

1.2.4 Metodología de Minería de datos – SEMMA

Para SAS – España (2006):

SEMMA es un acrónimo que significa Sample, Explore, Modify, Model and Assess que en español significa Muestra, Exploración, Modificación Modelo y Valoración. Es una lista de pasos secuenciales desarrollados por SAS Institute Inc., uno de los mayores productores de estadísticas y software de inteligencia de negocios. Orienta la implementación de aplicaciones de minería de datos. A pesar que SEMMA es a menudo considerado como una metodología general de minería de datos, es más bien una organización lógica del sistema de herramienta funcional de uno de sus productos, SAS Enterprise Miner, para la realización de las tareas principales de la minería de datos.

SEMMA es una metodología más corta y menos extensa que el CRISP-DM porque se centra más en el desarrollo del proceso de Minería de datos y no se orienta a objetivos empresariales

La metodología presentada tiene 5 fases, cada una representando a sus siglas SEMMA:

- ✓ **Sample**: Extracción de una muestra representativa

En esta primera fase de la metodología, se realiza la extracción de un conjunto de datos que sean una buena representación de la población a analizar, esto se hace con el objetivo de facilitar los procesos de minado sobre los datos, reduciendo los tiempos que se necesita para determinar la información valiosa para el negocio.

- ✓ **Explore**: Exploración de los datos en la muestra.

En esta fase, se hace un recorrido a través de los datos extraídos en la muestra para detectar, identificar y eliminar datos anómalos, ayudando a refinar los procesos de descubrimiento de información en fases siguientes del proceso. En este punto del proceso, la exploración se puede realizar a través de medios visuales, aunque muchas veces no es suficiente este método, es por eso, que además de la visualización se pueden manejar diferentes técnicas estadísticas como análisis de factores, análisis de correspondencias, entre otros.

- ✓ **Modify**: Modificación de los datos.

Esta modificación de los datos se puede realizar creando, seleccionando y transformando las variables en las cuales se va a enfocar el proceso de selección del modelo. Muchas veces se tendrá la necesidad de realizar modificaciones cuando los datos que se están analizando cambien. Esto se debe a que el entorno en el que se trabaja la minería de datos es dinámico e iterativo.

- ✓ **Model**: Modelación de los datos

En esta fase, las herramientas de software se encargan de realizar una búsqueda completa de combinaciones de datos que juntos predecirán de una manera confiable los resultados buscados.

Es en esta parte donde las técnicas y métodos de minería de datos entran a jugar un papel importante para la solución de los problemas que fueron identificados al iniciar el proyecto de minería de datos.

- ✓ **Assess**: Evaluación de los datos obtenidos

Después de que la fase de modelación presente los resultados obtenidos de la aplicación de los métodos de minería de datos al conjunto de datos. Se deberá realizar un análisis de los resultados para ver si estos fueron exitosos de acuerdo a las entradas que se tuvieron para analizar el problema. Una buena práctica para identificar si los resultados con el modelo creado son los esperados, es aplicar este modelo a una porción de datos diferente. Si el modelo funciona correctamente para esta muestra y para la muestra utilizada para el proceso de creación del modelo, se tiene una buena probabilidad de tener un modelo valido (s/p).

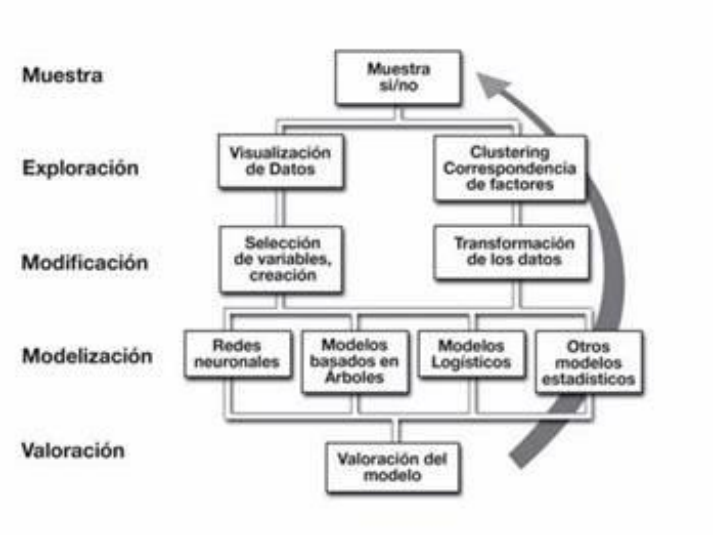


Figura 7. Fases de la metodología SEMMA
Fuente: (SAS - España, 2006)

1.2.5 Weka software de minería de datos

Hall (2015) afirma que Weka está en constante evolución, la cual está enfocada principalmente en las nuevas funcionalidades del software, actualmente se tiene más de 4 años de desarrollo continuo, donde se tiene varios contribuidores ya que es un software Open Source y posee de un respaldo de la Universidad de Waikato, en este documento aparte de las nuevas funcionalidades también se habla de las futuras funcionalidades que se desean implementar y de los casos de uso a los que se ha empleado, figurando sus casos de éxito en distintos áreas temáticas y orientándolos diferentes instituciones que han dado su visto bueno para la utilización del software.

1.2.6 Metodología de gestión del proyecto: PMBOK

Según el PMI (2015), PMBOK es una guía que contiene estándares globales para la gestión de Proyectos, estas son dictadas por el PMI (Project Management Institute).

Según PMBOK Fifth Edition (2014), un proyecto es un esfuerzo temporal que se lleva a cabo para crear un producto, servicio o resultado único. La naturaleza temporal de los proyectos implica que un proyecto tiene un principio y un final definidos. El final se alcanza cuando se logran los objetivos del proyecto, cuando se termina el proyecto porque sus objetivos no se cumplirán o no pueden ser cumplidos, o cuando ya no existe la necesidad que dio origen al proyecto. Asimismo, se puede poner fin a un proyecto si el cliente desea terminar el proyecto. Asimismo (PMBOK Fifth Edition, 2014) nos dice que existen 47 procesos de la dirección de proyectos que a su vez se agrupan en diez Áreas de Conocimiento claramente diferenciadas.

Un Área de Conocimiento representa un conjunto completo de conceptos, términos y actividades que conforman un ámbito profesional, un ámbito de la dirección de proyectos o un área de especialización. Estas diez Áreas de Conocimiento se utilizan en la mayoría de los proyectos, durante la mayor parte del tiempo. Los equipos de proyecto deben utilizar estas diez Áreas de Conocimiento, así como otras áreas de conocimiento, de la manera más adecuada en su proyecto específico.

Las Áreas de Conocimiento son: Gestión de la integración del proyecto, Gestión del alcance del proyecto, Gestión del tiempo del proyecto, Gestión de la calidad del proyecto, Gestión de recursos humanos del proyecto, Gestión de las comunicaciones del proyecto, Gestión de los riesgos del proyecto, Gestión de las adquisiciones del proyecto y Gestión de los interesados del proyecto.

Cada una de las Áreas de Conocimiento se trata en una sección específica de la Guía del PMBOK.

1.3 Definición de términos básicos

a) **Bloquer:** Conocido por ser una falla que al ocurrir, no es posible continuar con el proceso de la función que se había seleccionado.

b) **Data Set:** Se denomina Data Set al conjunto de datos a analizar con el software de minería de datos.

c) **ETL:** Es el proceso que permite a las organizaciones mover datos desde múltiples fuentes para, limpiarlos, unificarlos y cargarlos en un repositorio, este generalmente es una base de datos, data Warehouse o Data Mart.

d) **Matriz de confusión:** Es una herramienta muy usada en inteligencia artificial, que sirve para ver las predicciones correctas de cada clase y haca donde fueron las predicciones erradas. Cada columna representa el número de predicciones de cada clase y cada fila representa las instancias reales de la clase.

e) **MD:** Siglas pertenecientes a la tecnología de Minería de datos.

f) **Minería de Datos:** Es el proceso de detectar la información de grandes conjuntos de datos, utilizando un análisis matemático para deducir los patrones y tendencias que existen en los datos. Normalmente, estos patrones no se pueden detectar mediante la exploración tradicional de los datos porque las relaciones son demasiado complejas o porque hay demasiado datos (Microsoft corporation, 2015).

g) **Redes Neuronales Artificiales:** son un paradigma de aprendizaje y procesamiento automático inspirado en la forma en que funciona el sistema nervioso de los animales. Se trata de un sistema de interconexión de neuronas que colaboran entre sí para producir un estímulo de salida (Norvig y Russell, 2014).

h) **Servicio Policial:** El servicio policial se basa en garantizar, mantener y restablecer el orden interno, además de Prestar protección y ayuda a las personas y a la comunidad. Garantiza el cumplimiento de las leyes y la seguridad del patrimonio público y del privado. Previene, investiga y combate la delincuencia. Vigila y controla las fronteras.

i) **UAT** Son las siglas de User Aceptance Testing, que se refiere a todas las personas (usuarios) que van a probar la aplicación manualmente.

j) **Weka:** Software de aprendizaje automático y Minería de Datos desarrollado por la Universidad de Waikato en Nueva Zelanda.

CAPÍTULO II
METODOLOGÍA

2.1 Materiales

A continuación se describen los distintos recursos y Herramientas que se van a utilizar para el desarrollo del proyecto.

2.1.1 Recursos Humanos

Los 2 recursos humanos disponibles (ejecutores del proyecto) desempeñarán ciertos roles, por lo que cada uno de los recursos tendrá más de un rol y además estos serán multidisciplinarios para poder abarcar todas las disciplinas de la metodología seleccionada.

A continuación el detalle de las funciones principales:

k) Jaulis Rua Jorge (JJ)

l) Vilcarromero Giraldo Jonathan (JV)

Tabla 2. Roles y funciones de los ejecutores del proyecto

ROL	RESPONSABLE	FUNCIONES GENERALES
Analista de BI	JV	• Análisis y modelado del esquema de base de datos (según información recolectada) • Unificación la información de las denuncias registradas en las distintas comisarías (La Molina, Las Praderas, Santa Felicia) de la Molina. • Modelado y carga de los procesos ETL.
Analista Estadístico	JV / JA	• Selección de variables input para el Análisis Estadístico. • Establecer coeficientes y/o pesos de variables input. • Establecer ejecuciones del modelo de aprendizaje automático. • Evaluar la incertidumbre en los resultados. • Determinación de las pruebas estadísticas a utilizar

		• Realizar el informe de Estadístico de validación de las hipótesis Planteadas
Diseñador	JJ	• Realizar los Mockups de los flujos de navegación que tendrá la aplicación. (móvil, web)
Desarrollador	JJ	• Análisis de la arquitectura de la aplicación. • Desarrollo de capa BackEnd (lógica de negocio) de la Aplicación móvil. • Desarrollo de FrontEnd (capa cliente) de la aplicación.
Analista QA	JJ/JV	• Realizar las pruebas funcionalidad del sistema para ratifiquen el correcto funcionamiento del Sistema. • Pruebas de stress para simular la concurrencia a la aplicación.

Fuente: elaboración propia

2.1.2 Hardware

En la tabla que se muestra a continuación se detalla el hardware necesario para poder ejecutar el proyecto.

Tabla 3. Hardware utilizado en el proyecto

DESCRIPCIÓN	CANTIDAD
Laptop Toshiba Satellite 845C Core i3	1
Laptop Toshiba Satellite I845 Core i5	1
SmartPhone Huawei Y3C con Android 4.4	1

Fuente: elaboración propia.

2.1.3 Software

Tabla 4. Software utilizado en el proyecto

DESCRIPCIÓN	CANTIDAD	LICENCIA
Base de Datos		
Mysql 5.6	1	Open Source
ETL		
Pentaho Data Integration	1	Open Source
Análisis Estadístico y Data Mining		
Weka 6.3.12	2	Open Source
Gestión de Proyectos		
TeamGantt	2	Open Source
Programación		
Android Studio	1	Open Source
Diseño		
Balzamiq Mockups	1	Trial
Pruebas		
Apache Jmeter	1	Open Source
Sistemas Operativos		
Windows 8.1 ServicePack 1	1	Propietario
Windows 8 ServicePack 1	1	Propietario
Documentación		
Microsoft Office 2010	2	Propietario

Fuente: elaboración propia.

2.1.4 Cronograma

El proyecto está planteado para realizarlo en 4 meses, por ende la asignación de actividades para la creación del modelo predictivo de hechos delictivos se encuentran según la metodología CRISP-DM de minería de datos.

Sin embargo la metodología del desarrollo del proyecto está basada en PMBOK 5ta Edición, es por este motivo que según esta última se determinarán las fases en las que se agrupará nuestro cronograma de actividades. Este documento se encuentra de forma detallada en el Anexo 1.

En la imagen siguiente se muestran las fases del Inicio del proyecto hasta la fase de cierre comprendiendo la creación del modelo hasta el despliegue del

sistema de predicción en forma geográfica y posteriormente el levantamiento de correcciones respectivo en base a las observaciones realizadas por cada uno de los miembros de jurado. El detalle a nivel de actividades del cronograma se puede ver en el ANEXO 1.

Tabla 5. Tabla General del Cronograma de Actividades según PMBOK

CRONOGRAMA DE ACTIVIDADES			
Nombre	**Duración**	**Comienzo**	**Fin**
SISTPRE (SISTEMA DE PREDICCION DE HECHOS DELICTIVOS)	118 días	08/08/15	04/12/15
Fase I : Inicio del Proyecto	7 días	12/08/15	19/08/15
Fase II : Organización y Preparación	69 días	21/08/15	21/11/15
Fase III : Ejecución del Trabajo	45 días	21/08/15	22/10/15
Fase IV : Cierre	69 días	21/08/15	25/11/15

Fuente: elaboración propia

2.1.5 Presupuesto

Debido a que se decidió utilizar en su mayoría, software libre, los costos que se incurrirán básicamente están relacionado a recursos humanos y hardware (los pocos software propietarios incurren en costos mínimos pero que de la misma manera se detallarán).

Los costos por Recursos Humanos hacen referencia a los salarios que se incurren al pagar al personal asociado al proyecto.

Para establecer los meses de dedicación de cada uno de los intervinientes en el proyecto, se ha seleccionado el periodo total del proyecto para todos los roles (debido a que los ejecutores del proyecto abarcaran todos los roles) Para los costes por mes de cada uno de los intervinientes del proyecto, se ha tomado el siguiente sueldo promedio, el cual se obtuvo revisando las ofertas laborales que se ofrecen en el mercado de tecnología en el Perú, según cada rol son los siguientes:

- S/.36.000/año para el analista BI.
- S/. 30.000/año para el analista estadístico.

- S/. 24.000/año para el diseñador.
- S/. 30.000/año para el desarrollador.
- S/. 33.600/año para el analista QA.

Para todos los costes expuestos, se considera que trabajan 45 horas semanales.

Tras realizar los cálculos asociados al coste de personal, que pueden ser consultados en la Tabla 6, podemos concluir que los rostes de recursos humanos del proyecto ascienden a S/. 9150.

Tabla 6. Costo de Recursos Humanos.

Rol	Responsable	Dedicación (meses)	Costo hombre por mes (S/.)	Costo (S/.)
Analista BI	JV	1	3000	3000
Analista estadístico	JV	0.5	2500	1250
Diseñador	JJ	0.5	2000	1000
Desarrollador	JJ	1	2500	2500
Analista QA	JJ/JV	0.5	2800	1400
			TOTAL	**9150**

Fuente: elaboración propia

Con lo que respecta al Hardware, se tomará en cuenta un costo derivado a lo referido al uso de equipos como laptops, tablets y otros que fueron necesarios en el proyecto.

En la tabla 7 se muestra detalladamente el material hardware utilizado para el desarrollo del proyecto. El costo imputado será calculado en función de la amortización del dispositivo. Se decidió que su amortización se realice en un periodo de 60 meses (basado en el costeo que diferentes autores realizan).

A continuación se muestra la fórmula utilizada para el coste amortizado, la cual servirá para el cálculo de los costes asociados al hardware y software usado durante el proyecto.

$$\frac{A}{B} xCxD$$

Donde:

A = número de meses desde la fecha de facturación en que el equipo es utilizado

B = periodo de depreciación (en meses)

C = costo total del equipo

D = % del uso que se dedica al proyecto

Tabla 7. Costo de hardware

Descripción	Costo (S/.)	% Uso en el proyecto	Dedicación (meses)	Periodo de depreciación	Costo imputable
Laptop Toshiba Satellite 845C Core i3	2500	100%	3	60	125
Laptop Toshiba Satellite I845 Core i5	3000	100%	3	60	150
Smartphone Huawei Y3C con Android 4.4	500	100%	3	60	25
				TOTAL	**300**

Fuente: elaboración propia

Con relación a los costos relacionados con Software se tendrán en cuenta todas las herramientas de software utilizadas para el desarrollo del proyecto. El periodo de depreciación será de 3 años (36 meses).

Tabla 8. Costos de software**.**

Descripción	Costo (S/.)	% Uso dedicado al proyecto	Dedicación (meses)	Periodo de depreciación	Costo imputable
Weka	0	100%	3	36	0
MS Office 2010	189.9	100%	3	36	15.825
TeamGantt	0	100%	3	36	0
Pentaho Data Integration	0	100%	3	36	0
Apache Jmeter	0	100%	3	36	0
Maria DB 10.1.7	0	100%	3	36	0
				TOTAL	**15.83**

Fuente: Elaboración propia.

Los costos totales se detallan a continuación:

Tabla 9. Costos totales del proyecto.

Tipo de Costo	Subtotal (S/.)
Recursos Humanos	9150.00
Hardware	15.83
Software	300.00
TOTAL	**9465.83**

Fuente: Elaboración propia.

2.1.5.1 Costos del pase a producción

Para el pase a producción del sistema de predicción de hechos delictivos, se optaron por herramientas Open Source que no generen costo a lo largo de la vida útil de proyecto que se estima en un periodo de 3 años. Los gastos Post Mantenimiento del sistema son los siguientes:

Tabla 10. Herramientas de Hosting usadas en producción.

Servidor	Herramienta	Costo en soles	Costo en el ciclo de vida(3 años)
Servidor de aplicaciones	Cloud 9	--	0
Servidor de Base de datos	Google SQL Cloud	--	0
Soporte (Equipo de Proyecto)	--	0	0
TOTAL		--	**0**

Fuente: elaboración propia

El servidor de aplicaciones alojado en cloud9.io que es una organización que apoya el software libre, te permite un despliegue automático de la aplicación, otorgándote un dominio permanente siempre y cuando el consumo de la aplicación de no supere 1 GB de espacio en disco y 512MB en Memoria RAM.

Actualmente SITPRE consume 128 MB en memoria RAM, 80MB en disco con una concurrencia de 20 conexiones.

El servidor de Base de datos alojado en Google SQLCloud ofrece una versión de prueba de 2 meses, y posteriormente de pago; a lo largo del desarrollo del proyecto para saltarnos esta restricción simplemente se crea otra cuenta y se migra la base de datos actual, se planea hacer lo mismo hasta el fin ciclo de vida de proyecto.

El costo del soporte a cargo del equipo de proyecto no tiene un costo adicional, ya que se ha establecido a los Stakeholders mutuamente.

2.1.6 Evaluación social del proyecto

A diferencia de algunos proyectos con fines de lucro, este trabajo no pretende generar ganancias y ser viable económicamente a lo largo del tiempo, ya que como se ha mencionado en la introducción, nuestro objetivo es mejorar el desempeño de las comisarias hacia la prevención de delitos concurrentes, para ello nuestra métrica de evaluación será mediante:

- La validación del modelo de predicción de hechos delictivos, es decir que el sistema dé resultados precisos con una tasa de error mínima definida en la sección validación.
- La satisfacción manifestada por cada uno de los usuarios que hayan usado la aplicación, intentando recoger la apreciación sobre cuánto ha ayudado la aplicación en sus labores diarias.

2.2 Métodos

2.2.1 Metodología de desarrollo del proyecto

Como se ha mencionado en el capítulo I: Bases Teóricas PMBOK es un estándar de administración de proyectos que comprenden un conjunto de procesos y áreas de conocimiento aceptadas y reconocidas como los mejores dentro de la gestión de proyectos, por ello se optó por dicha metodología para la gestión del proyecto “Sistema de Predicción”, el cual tendrá el siguiente ciclo de Vida.

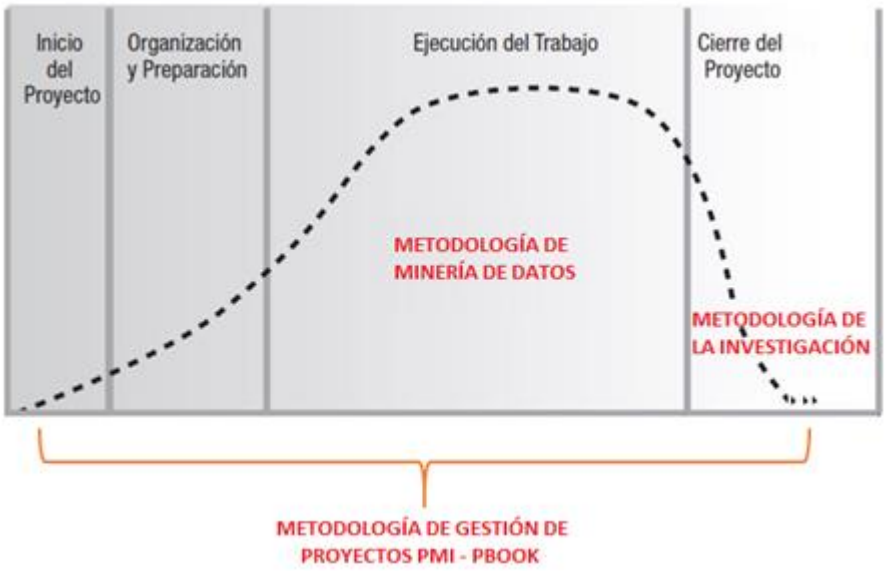


Figura 8. Ciclo de Vida del Proyecto
Fuente: elaboración Propia

La primera fase de Inicio de Proyecto y la segunda denominada Organización y Preparación, presentarán entregables basados en la guía PMBOK v5, donde el equipo de proyecto identificó las áreas del Conocimiento descritas en dicha guía y que se acomodan al proyecto a realizar.

Debido a que no todas las áreas del conocimiento son aplicables al producto a desarrollar, y basándonos en que la guía PMBOK indica que se puede determinar o conformar el ciclo de vida del proyecto sobre la base de los aspectos únicos de la organización, de la industria o de la tecnología empleada. Mientras que cada proyecto tiene un inicio y un final definido, los entregables específicos y las actividades que se llevan a cabo variarán ampliamente dependiendo del proyecto se decidió aplicar las siguientes áreas del conocimiento.

- Gestión de la Integración: debido a que se necesita tener las características, comunicación y acciones integradoras cruciales para que el proyecto se lleve de una manera controlada y llegue a cumplir los requisitos. Se tienen los siguientes entregables: Anexo 2 Acta de constitución del proyecto**.** Anexo 3 Plan de proyecto

- Gestión de Interesados: Al ser una investigación aplicada a una institución, se tienen sujetos interesados en que el proyecto dé beneficios a la organización o a un área específica. Anexo 4 Plan de gestión de interesados**.**
- Gestión de Riesgos**:** debido a que los riesgos existen en todo tipo de proyecto sin importan el tamaño o la envergadura del mismo. Anexo 5 Plan de gestión de riesgos

Para la fase de Ejecución del trabajo se optó por una metodología de minería de datos, debido a la naturaleza de nuestro proyecto a realizar. La metodología de MD seleccionada y el desarrollo de la misma serán descrita en los apartados siguientes siguiendo los lineamientos mencionados en el capítulo de Bases Teóricas.

Específicamente, el desarrollo de esta fase se encuentra en el siguiente Capítulo III: Desarrollo de Proyecto donde se especifican las actividades realizadas para la creación del Modelo de predicción de delitos y del respectivo desarrollo del sistema y posterior despliegue.

Para la fase de Cierre del Proyecto se contempla los objetivos alcanzados y las pruebas estadísticas empleadas, estas se encuentran en el Capítulo V Discusiones y Aplicaciones y está basado en la “metodología de la investigación” que se plantea en el apartado 2.2.2.

Adicionalmente el equipo de proyecto ha decidido mantener la comunicación con los Stakeholders para futuras implementaciones y mejoras para el sistema de predicción por ende aún no se da por terminado el proyecto pues podrían haber más fases. Por último, establecemos un tiempo de vida útil del Sistema por un mínimo de 3 años.

2.2.2 Metodología de la investigación

A) Tipo de investigación

Según Hernández, Fernández y Batista (1997) existen tres tipos de investigación los cuales son:

- Exploratoria: Se efectúan cuando el objetivo es examinar un tema o problema que nunca ha sido estudiado, o solo existen guías o ideas poco relacionadas con el tema de estudio.
- Descriptiva: Muestran y describen situaciones o eventos que pueden ser medidos, pues desde el punto de vista científico describir es medir. Seleccionando una serie de cuestiones para medir cada una de ellas independientemente y así describir lo que se investiga.
- Correlacional: Pretenden medir el grado de relación que existe entre dos o más conceptos variables. Responde preguntas de relación de variable.
- Explicativa: Están dirigidos a responder a las causas de los eventos físicos o sociales, centrándose en explicar porque ocurre un fenómeno y en qué condiciones sucede este.

Ya que nuestra investigación es de tipo descriptiva, debido a que se va medir situaciones en base a dos escenarios en diferentes tiempos.

B) Variables a Analizar

Según las hipótesis planteadas en la Introducción se debe considerar que cada hipótesis tiene una variable a ser medida, además para la aceptación de la hipótesis general nuestro criterio de validación sería en base a la aceptación de las hipótesis específicas.

Siendo nuestra hipótesis general Y en base a la mejora en el proceso de prevención de delitos:

$$Y = F(x,y,z)$$

x: hipótesis específica 1 en base a tiempos de acción

y: hipótesis específica 2 en base al número de acciones preventivas

z: hipótesis específica 3 en base a la estructuración de la información de las comisarias

C) Operacionalidad de Variables

Tabla 11. Operacionalidad de variables

VARIABLE	Tiempo de identificación de zona de riesgo		Nro. de Acciones Preventivas	
DEFINICION CONCEPTUAL	Se refiera al tiempo que le toma a un efectivo policial determinar una zona de riesgo.		Se refiere al número de acciones preventivas que realiza un efectivo policial diariamente , esta comprende patrullajes, rondas y operativos	
TIPO	Cualitativo	Cuantitativo	Cualitativo	Cuantitativo
INDICADOR	Conocimiento de riesgo por zona	Tiempo promedio empleado	Criterio para la realización de Operativos	Asignación de Recursos
FUENTES	Efectivos Policiales	Efectivos Policiales	Efectivos Policiales	Efectivos Policiales
TECNICA DE RECOLECCIÓN DE INFORMACION	Encuestas	Encuestas	Encuestas	Encuestas
ESCALA DE MEDICION	Nominal: Alta Media Baja	Continuo	Nominal: SI NO	Discreta

Fuente: elaboración propia

D) Población

Nuestra población de estudio viene a ser todos los efectivos policiales, que se encuentran asignados a lo largo de las tres comisarías de La Molina, siendo un promedio de 150 efectivos policiales.

E) Muestra

Para la determinación de la muestra se ha hecho un muestreo no probabilístico es decir nuestra muestra será en base a nuestro criterio y conocimiento, por ello se ha escogido a la comisaría de Santa Felicia como

el plan piloto de implementación de nuestro sistema, donde seleccionaremos 20 efectivos policiales para las mediciones.

F) Técnica de recolección de datos

La técnica de recolección de datos será en base a encuestas realizadas a los efectivos policiales, tal como se mostró en la matriz de Operacionalidad de Variables.

G) Análisis y comprobación de hipótesis

La comprobación de la hipótesis se realiza en el capítulo V Discusiones y Aplicaciones.

2.2.2 Metodologías de minería de datos

Como se ha mencionado previamente se tiene que seleccionar una la metodología de minería de datos para la elaboración del modelo de predicción de hechos delictivos, para ello se realizará una comparación entre SEMMA y CRISP-DM.

La elección de la metodología se basó según una comparativa de entre las dos metodologías investigadas sobre minería de datos. El cuadro comparativo se muestra las diferencias principales de ambas metodologías para la selección de la que más se acomode al trabajo a realizar.

Tabla 12. Comparación de metodologías investigadas.

SEMMA	CRISP – DM
Se centra en las fases técnicas necesarias en todo proyecto de minería de datos	Basa su orientación en los objetivos de negocio alineados a los objetivos de MD.
Metodología desarrollada por SAS	Metodología Libre.
Se inicia analizando los datos que se tiene disponibles	Se inicia analizando los objetivos del negocio para alinearlos a los del proyecto
Se define una muestra significativa, luego se explora dicha muestra	Muestra y exploración de datos se realizan juntos de forma paralela
Evalúa el éxito del modelo mediante validaciones a los datos	Evalúa el éxito del modelo mediante el logro de objetivos de negocio – proyecto

No elabora planes finales.	Elabora plan de monitoreo luego de implantar la solución.
Abarca iteraciones a las fases anteriores, es un proceso cíclico	Tiene naturaleza cíclico
No se orienta más allá que la MD	Tiene una orientación similar a la gestión de proyectos.

Fuente: elaboración propia.

Según el análisis hecho al realizar el cuadro comparativo, el equipo de investigación decidió asignarle pesos en una escala del 1 al 3 (poco importante, importante y muy importante respectivamente) según los criterios que consideramos más importantes para que el modelo llegue a estar listo en el tiempo establecido y alcanzando los objetivos previamente planteados.

Adicionalmente se estableció un rango del 1-5 para establecer qué tan bueno es una metodología en dicho criterio.

Dónde: 1 – Muy malo, 2 – Malo, 3 – regular, 4 – bueno, 5 muy bueno

Tabla 13. Criterios para la elección de la metodología de MD.

CRITERIO	IMPORTANCIA	SEMMA	CRISP - DM	TOTAL SEMMA	TOTAL CRISP – DM
Orientación a objetivos de negocio	2	2	4	4	8
Selección de muestra significativa como fase de partida	2	5	3	10	6
Adaptación a cualquier herramienta	3	2	5	6	15
Agilidad y al momento de documentar	3	3	4	9	12
realización de		3	4	9	12

actividades en paralelo	3				
Proceso cíclico para corregir errores en fases previas	1	5	5	5	5
Orientación ligada a gestión de proyectos	2	2	4	4	8
			TOTAL	**47**	**66**

Fuente: elaboración propia.

Según los resultados del cuadro mostrado, la mejor opción es utilizar la metodología CRISP – DM debido a las siguientes razones:

- El objetivo de nuestra investigación está muy relacionado con la ayuda a las comisarías de la molina, es decir nos orientamos a los objetivos de la entidad policial.
- Al utilizar software libre, necesitamos una metodología que se pueda acoplar a cualquier tipo de herramienta.
- Debido al corto tiempo de desarrollo que se tiene, se necesita una metodología que haga una o más actividades en paralelo (como la muestra y exploración de datos).

CAPÍTULO III
DESARROLLO DEL PROYECTO

3.1 Fase 1: Comprensión del negocio

3.1.1 Determinar los objetivos de negocio

La primera actividad de esta fase es determinar los objetivos de las comisarias del distrito de La Molina, esto debido a que nuestra solución de predicción debe apoyar directamente a estos objetivos generando valor a los mismos.

Los objetivos de la comisaría son los que cumple la Policía Nacional del Perú, pero segmentado al distrito donde se investiga, en este caso La Molina.

Los objetivos son:

- Garantizar, mantener y restablecer el orden interno de la jurisdicción.
- Prestar protección y ayuda a las personas de la comunidad.
- Garantizar el cumplimiento de las leyes y la seguridad del patrimonio público y privado.
- Prevenir, investigar y combatir la delincuencia.
- Vigilar y controlar las fronteras con el propósito de defender a la sociedad y las personas en el marco de una cultura de paz y de respecto a los derechos humanos

Entonces, según los objetivos de la comisaria de La Molina, los objetivos relacionados a nuestra investigación y donde nuestra herramienta ayudará, son los que se muestran en color rojo.

3.1.2 Evaluar la situación

Esta actividad de la fase de conocimiento del negocio de la metodología, abarca una descripción general del proceso al que vamos a atacar en el negocio donde se implantará la solución de MD.

En capítulos anteriores (especialmente el 2) se describió la situación de Las Comisarias del distrito, además de las estadísticas referidas a los hechos

delictivos en el distrito y en Lima, es por ello que en esta fase se detallará solo las actividades del proceso de “Prevención del delito” que es el que se busca mejorar con nuestro sistema.

Según el conocimiento de los procesos de las comisarías de la molina, y la información recolectada, pudimos describir el proceso de prevención del delito, el cual posee diversas actividades bien diferenciadas.

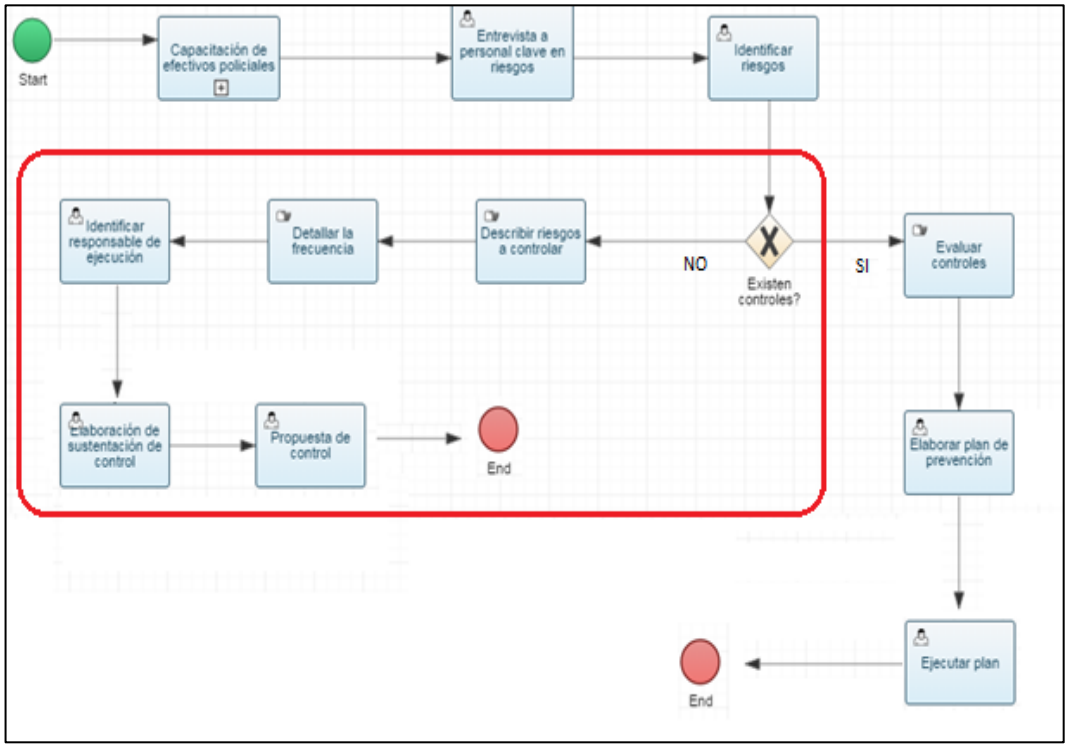


Figura 9. Proceso de Prevención de delitos.
Fuente: Propia

Como se puede observar en la figura, la solución abarcará la parte sombreada; es decir, servirá como herramienta de información en las actividades relacionadas cuando no existen controles de hechos delictivos.

Cuando en el proceso no existen controles hacia un hecho delictivo específico (detallado en el atestado policial de las denuncias).
Lo primero que hacen los efectivos encargados de la unidad de estadística es analizar y detallar la frecuencia de estos hechos delictivos esto con el objetivo de determinar qué tan recurrente son, si se da con una frecuencia

moderada o si se da en esporádicamente (abarca una tarea manual de buscar información pasada relacionada a dicha denuncia)

3.1.3 Determinar los objetivos de minería de datos

- Determinar un modelo eficaz de minería de datos para predecir los hechos delictivos en el distrito de La Molina.
- Conseguir como un porcentaje mínimo de aciertos según los objetivos del negocio, en este caso las comisarias (detallado como criterio de éxito).
- Determinar las zonas donde podría ocurrir hechos delictivos según una selección de variables de entrada.
- Mostrar los resultados de las predicciones en un mapa geográfico.

Criterios claves de éxito:

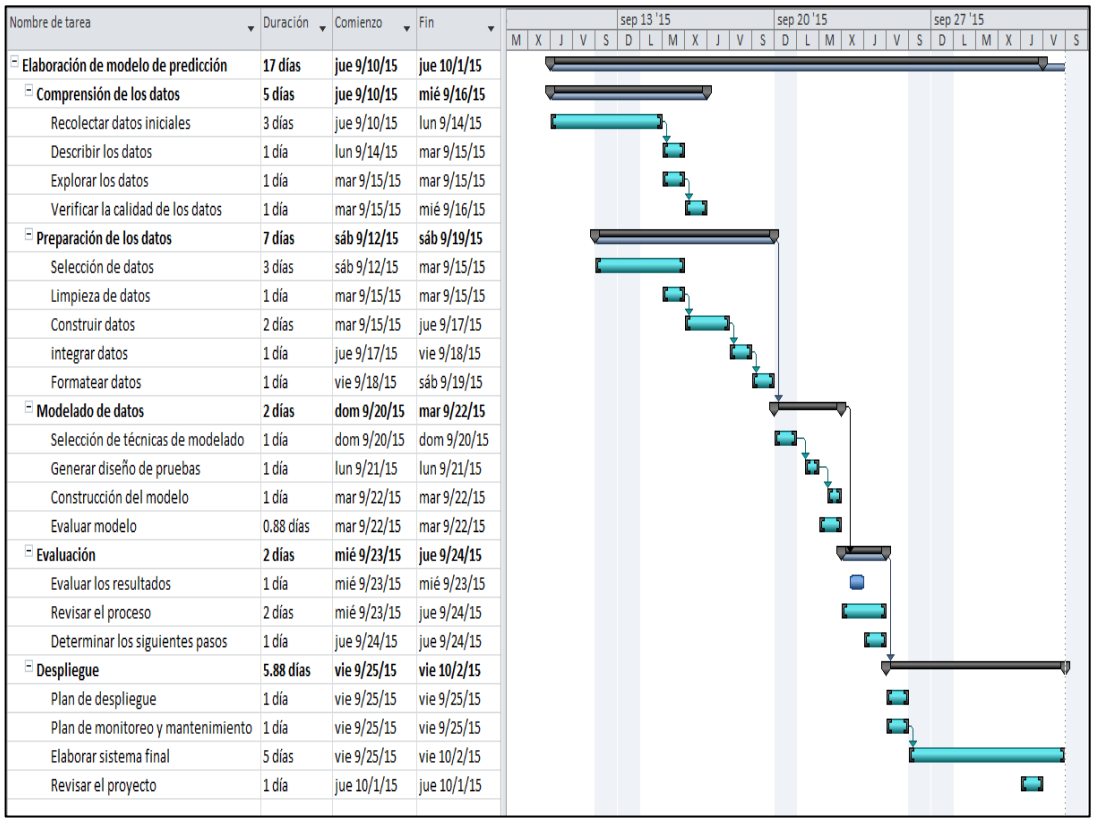
Según los objetivos del negocio, y las denuncias que se registran mensualmente podemos tener como criterio de éxito que:

- Se obtenga como mínimo un 60% de probabilidad de aciertos en las predicciones.
- Se tenga un detalle de las zonas con los delitos más comunes con su probabilidad de ocurrencia.
- Se pueda realimentar el modelo con más data histórica.
- Se puedan predecir más hechos delictivos según vaya pasando el tiempo (no solo para uno o dos meses siguientes, sino que se pueda volver continuo en el futuro).

3.1.4 Producir plan de proyecto

Esta actividad comprende dos bloques importantes e independientes. El primero se refiere a elaborar un plan general del proyecto a realizar, y el segundo está referido a elaborar una evaluación preliminar de la herramienta que se utilizará en el proyecto de minería de datos.

Plan de proyecto.



Nombre de tarea	Duración	Comienzo	Fin
Elaboración de modelo de predicción	**17 días**	**jue 9/10/15**	**jue 10/1/15**
Comprensión de los datos	**5 días**	**jue 9/10/15**	**mié 9/16/15**
Recolectar datos iniciales	3 días	jue 9/10/15	lun 9/14/15
Describir los datos	1 día	lun 9/14/15	mar 9/15/15
Explorar los datos	1 día	mar 9/15/15	mar 9/15/15
Verificar la calidad de los datos	1 día	mar 9/15/15	mié 9/16/15
Preparación de los datos	**7 días**	**sáb 9/12/15**	**sáb 9/19/15**
Selección de datos	3 días	sáb 9/12/15	mar 9/15/15
Limpieza de datos	1 día	mar 9/15/15	mar 9/15/15
Construir datos	2 días	mar 9/15/15	jue 9/17/15
integrar datos	1 día	jue 9/17/15	vie 9/18/15
Formatear datos	1 día	vie 9/18/15	sáb 9/19/15
Modelado de datos	**2 días**	**dom 9/20/15**	**mar 9/22/15**
Selección de técnicas de modelado	1 día	dom 9/20/15	dom 9/20/15
Generar diseño de pruebas	1 día	lun 9/21/15	lun 9/21/15
Construcción del modelo	1 día	mar 9/22/15	mar 9/22/15
Evaluar modelo	0.88 días	mar 9/22/15	mar 9/22/15
Evaluación	**2 días**	**mié 9/23/15**	**jue 9/24/15**
Evaluar los resultados	1 día	mié 9/23/15	mié 9/23/15
Revisar el proceso	2 días	mié 9/23/15	jue 9/24/15
Determinar los siguientes pasos	1 día	jue 9/24/15	jue 9/24/15
Despliegue	**5.88 días**	**vie 9/25/15**	**vie 10/2/15**
Plan de despliegue	1 día	vie 9/25/15	vie 9/25/15
Plan de monitoreo y mantenimiento	1 día	vie 9/25/15	vie 9/25/15
Elaborar sistema final	5 días	vie 9/25/15	vie 10/2/15
Revisar el proyecto	1 día	jue 10/1/15	jue 10/1/15

Figura 10. Plan del proyecto.
Fuente: elaboración propia.

Evaluación preliminar de la herramienta.

A continuación se muestra una serie de criterios relevantes para la construcción de nuestro modelo en particular, estos serán los datos de entrada para poder seleccionar la herramienta que más se adapte a nuestra solución de MD.

Tabla 14. Criterios para elegir herramienta de modelado.

CRITERIO	WEKA	MATLAB	SPSS	EASYNN PLUS
Facilidad de uso	Baja: No necesita instalar, posee tutoriales oficiales de descarga libre.	Alta: Posee muchas opciones, es además un entorno de desarrollo	Media: tutoriales limitados sin costo.	Media: herramienta no muy difundida, no muchas guías oficiales.

Acceso a diferentes fuente de datos	Diversas: bases de datos, archivos planos y formato propio.	Enfocado a la comunicación con otros programas o hardware	Diversas, BD, archivos planos. Al ser e IBM se dificulta la conexión con BD externas.	Diversas: CSV, bases de datos.
Bajo costo o libre	Libre.	Licencia.	Licencia.	Libre.
Opción para utilizar RNA	Sí	Sí	Sí	Sí
Campo central de orientación	Modelos de aprendizaje automático y MD.	Entorno de desarrollo, enfocado al desarrollo de modelos a medida del usuario.	Modelos de aprendizaje automático.	Aprendizaje automático.

Fuente: elaboración propia

Según estos criterios y los datos recolectados de cada herramienta, pasaremos a asignarles unos pesos según nuestras necesidades de desarrollo, Estos pesos significan:

- 1 - Muy malo
- 2 - Malo
- 3 - regular
- 4 - bueno
- 5 - muy bueno

Tabla 15. Calificación por criterio de cada herramienta.

CRITERIO	WEKA	MATLAB	SPSS	EASYNN PLUS
Facilidad de uso	4	4	3	2
Acceso a diferentes fuente de datos	4	2	3	4
Bajo costo o libre	5	2	2	5
presenta el algoritmo	5	5	5	5

RNA				
Campo de orientación	4	2	4	4
TOTAL	22	15	17	20

Fuente: elaboración propia

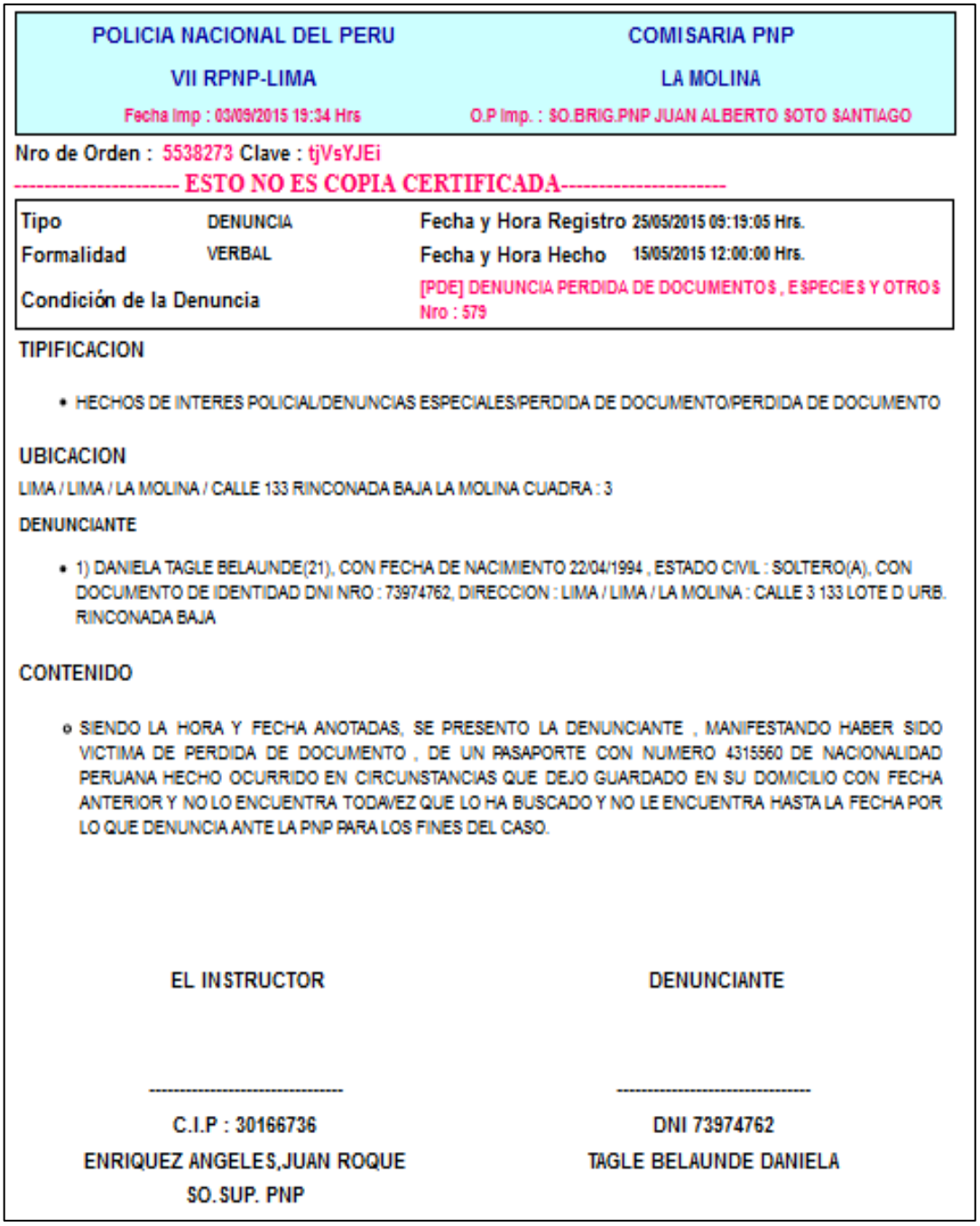
Según los criterios identificados y lo pesos asignados, la herramienta con la que se trabajará será **WEKA** de la Universidad de Waikato, Nueva Zelanda.

3.2 Fase 2 Compresión de los datos

3.2.1 Recolectar datos iniciales

Como primer paso, previo a la recolección de datos, se debe tener una base de datos donde se tenga almacenada toda la información que se va a obtener de las comisarías y a la vez soporte las funcionalidades del sistema a desarrollar (incluyendo el modelo de minería de datos), posterior a ello, se hará una discriminación de atributos, pues algunos de estos serán utilizados en el modelo de minería de datos, mientras que otros solo serán utilizados para la visualización del sistema o para los reportes que se construirán en fases posteriores. Para cualquiera de los casos mencionados, se debe tener un soporte de base de datos.

Para la construcción del modelo relacional se utilizó ingeniería inversa, donde el input fueron los atestados policiales, los cuales tienen toda la información relacionada a los hechos delictivos. La siguiente figura muestra un atestado policial de la comisaria de La Molina.



POLICIA NACIONAL DEL PERU — COMISARIA PNP

VII RPNP-LIMA — LA MOLINA

Fecha Imp : 03/09/2015 19:34 Hrs — O.P Imp. : SO.BRIG.PNP JUAN ALBERTO SOTO SANTIAGO

Nro de Orden : 5538273 Clave : tjVsYJEi

---------- ESTO NO ES COPIA CERTIFICADA----------

Tipo	DENUNCIA	Fecha y Hora Registro	25/05/2015 09:19:05 Hrs.
Formalidad	VERBAL	Fecha y Hora Hecho	15/05/2015 12:00:00 Hrs.
Condición de la Denuncia		[PDE] DENUNCIA PERDIDA DE DOCUMENTO$, ESPECIE$ Y OTRO$ Nro : 579	

TIPIFICACION

- HECHOS DE INTERES POLICIAL/DENUNCIAS ESPECIALES/PERDIDA DE DOCUMENTO/PERDIDA DE DOCUMENTO

UBICACION

LIMA / LIMA / LA MOLINA / CALLE 133 RINCONADA BAJA LA MOLINA CUADRA : 3

DENUNCIANTE

- 1) DANIELA TAGLE BELAUNDE(21), CON FECHA DE NACIMIENTO 22/04/1994 , ESTADO CIVIL : SOLTERO(A), CON DOCUMENTO DE IDENTIDAD DNI NRO : 73974762, DIRECCION : LIMA / LIMA / LA MOLINA : CALLE 3 133 LOTE D URB. RINCONADA BAJA

CONTENIDO

- SIENDO LA HORA Y FECHA ANOTADAS, SE PRESENTO LA DENUNCIANTE , MANIFESTANDO HABER SIDO VICTIMA DE PERDIDA DE DOCUMENTO , DE UN PASAPORTE CON NUMERO 4315560 DE NACIONALIDAD PERUANA HECHO OCURRIDO EN CIRCUNSTANCIAS QUE DEJO GUARDADO EN SU DOMICILIO CON FECHA ANTERIOR Y NO LO ENCUENTRA TODAVEZ QUE LO HA BUSCADO Y NO LE ENCUENTRA HASTA LA FECHA POR LO QUE DENUNCIA ANTE LA PNP PARA LOS FINES DEL CASO.

EL INSTRUCTOR — DENUNCIANTE

---------- ----------

C.I.P : 30166736 — DNI 73974762

ENRIQUEZ ANGELES,JUAN ROQUE — TAGLE BELAUNDE DANIELA

SO.SUP. PNP

Figura 11. Ejemplo de Atestado Policial
Fuente: DEINPOL

A partir de los atestados policiales se pudo elaborar un modelo de base de datos que dé soporte a todos los datos se va a recoger y al modelo que se elaborará en las fases posteriores. La imagen que se muestra a continuación es el modelo físico elaborado luego de hacer la ingeniería inversa a partir de los atestados:

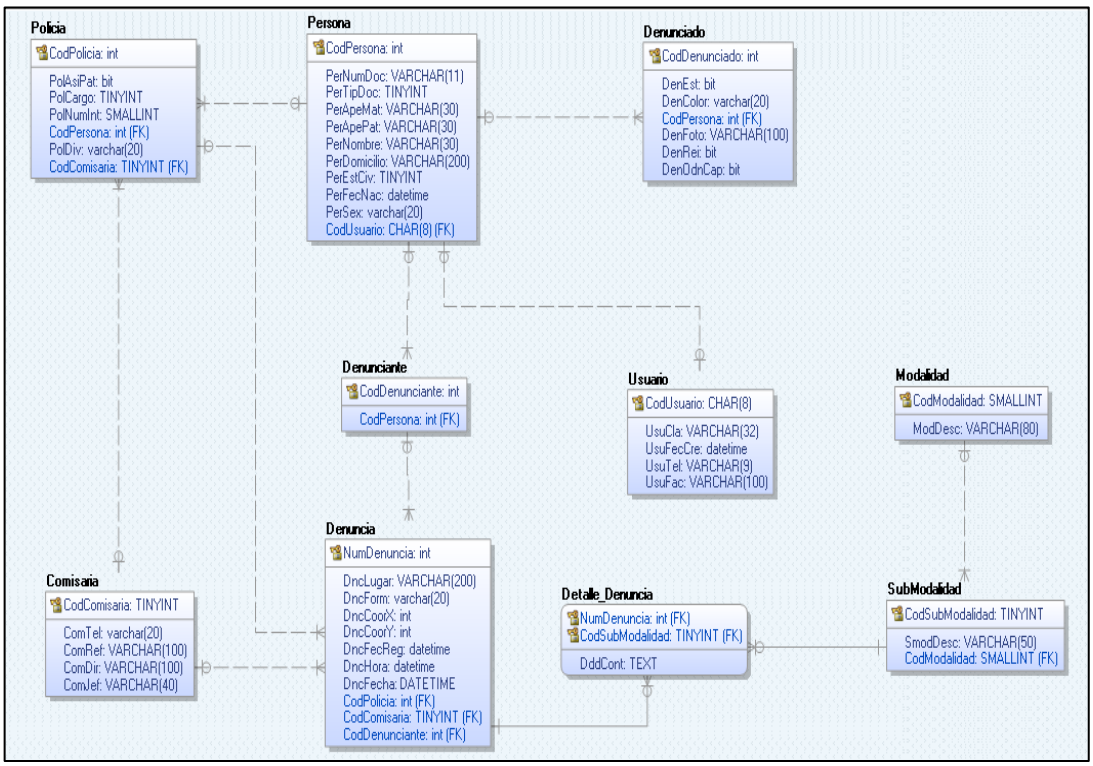


Figura 12. Modelo de base de datos.
Fuente: Elaboración propia

Luego de tener la base de datos modelada, se debe obtener la información por parte de las tres comisarías de La Molina. Esta información fue recogida en formato Excel, debido a que solo tuvimos acceso a los atestados policiales y a las fechas de denuncia. Al tener toda esta información en tablas de Excel, tuvimos que desarrollar un ETL para clasificar y cargar esta información.

Posteriormente se procedió a normalizar la información según la base de datos previamente modelada.

Finalmente se desarrolló el ETL (utilizando técnicas de inteligencia de negocio) para la carga de las tablas.

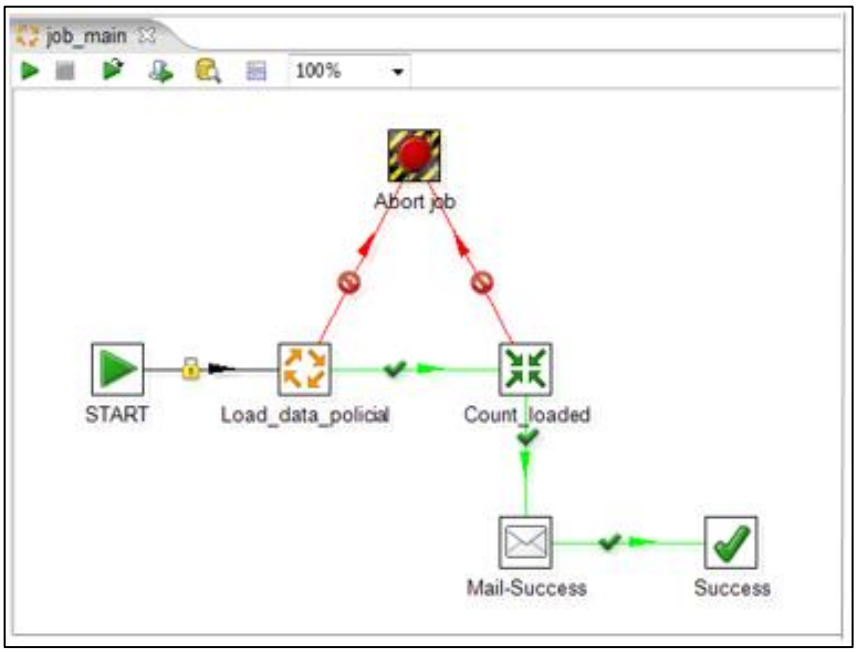


Figura 13. Job inicial del ETL
Fuente: Elaboración propia

La imagen muestra el Job principal que posee a su vez un Job de carga de datos y una transformación que hace el conteo de los datos en Excel vs los datos que se cargaron a nuestra base de datos (esto se realizó por temas de validación de data).

Posteriormente se muestra el contenido del Job “Load_data_policial” que contiene la lectura de los datos desde Excel y los lleva a la base de datos relacional. En este caso se utilizó una técnica de Inteligencia de negocios llamada “Carga por pasos” la cual define cada carga como un paso individual dependiente del anterior.

Esta técnica es utilizada para poder controlar desde qué paso se quiere empezar la carga, en qué paso detener la carga para hacer validaciones, aumentar información, etc.

Con esto se tiene el ETL que llena la base de datos que servirá para obtener las variables del modelo de predicción y a la vez para poder consultar la información necesaria para las funcionalidades adicionales del sistema.

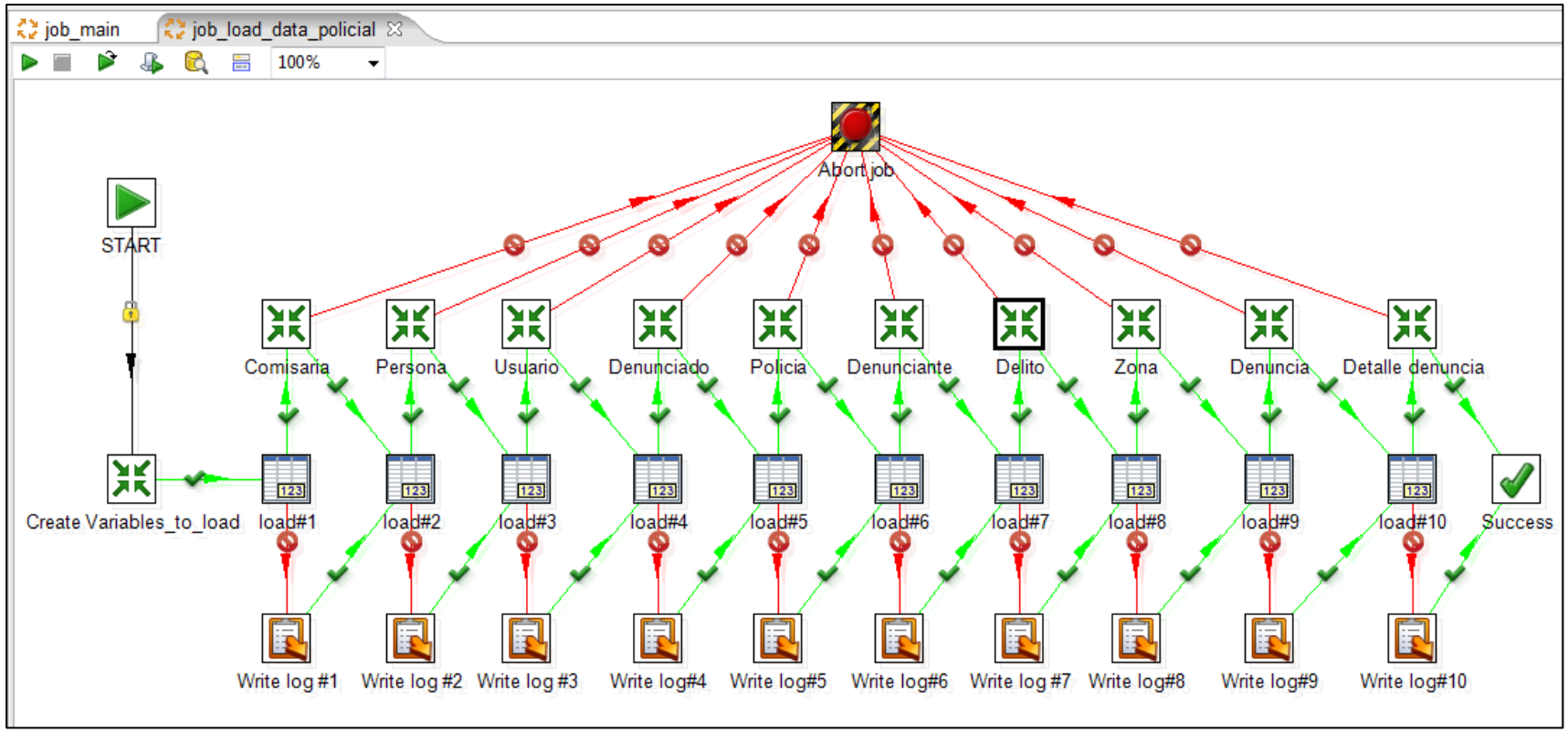


Figura 14. Job carga de tablas
Fuente: Elaboración Propia

Adicionalmente se muestra las transformaciones realizada para la carga de las tablas “denuncia” y “detalle_denuncia”.

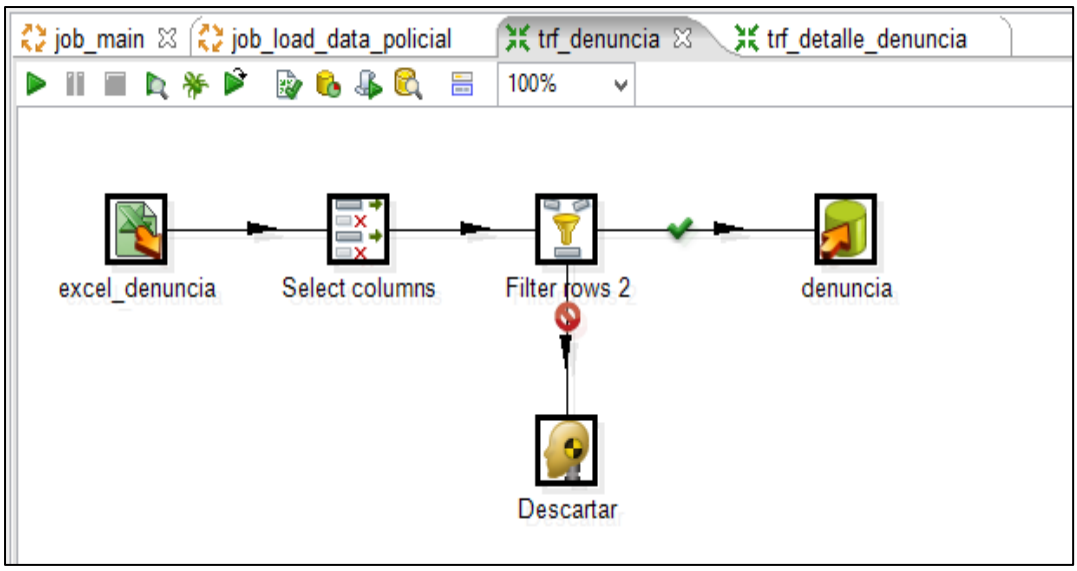


Figura 15. Carga de tabla Denuncia
Fuente: Elaboración propia

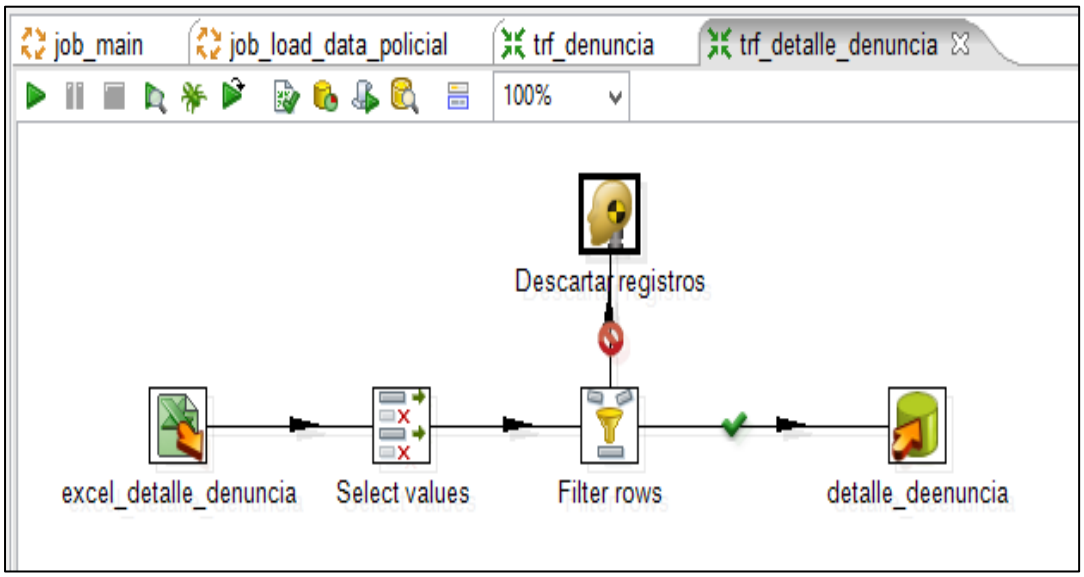


Figura 16. Carga tabla Detalle_denuncia
Fuente: Elaboración propia

3.2.2 Describir los datos

Tabla 16. Diccionario de datos.

TABLA	COLUMNA	DESCRIPCIÓN
PERSONA	PerNumDoc	Numero de documento PK
	PertipDoc	Tipo de documento (DNI, carné de extranjería, etc.)
	PerApePat	Apellido paterno
	PerApeMat	Apellido materno
	PerNombre	Nombres de la persona
	PerDomicilio	Dirección de la persona
	PerEstCiv	Estado civil (soltero, casado, etc.)
	PerFecNac	Fecha de nacimiento
	PerSex	Sexo de la persona
	CodUsuario	usuario para entrar al sistema (en caso de ser policía)
POLICIA	PolAsPat	identificador del jefe superior
	PolCargo	cargo del policía
	PolNumInt	numero interno del policía
	CodPersona	fk heredada de la tabla persona
	PolDiv	División que pertenece
DENUNCIADO	DenColor	rasgo físico
	CodPersona	FK de la tabla persona
	DenFoto	Foto del denunciado
	DenReg	Nro. de registro del denunciado
	DenOrnCap	flag si el denunciado tiene orden de captura
DENUNCIANTE	CodDenunciante	identificador del denunciante
	CodPersona	FK de la tabla persona
COMISARIA	CodComisaria	Identificador de la comisaria
	ComTel	Teléfono de la comisaria
	ComRef	Representa el nombre de referencia de la comisaria
	ComDIr	Dirección de la comisaria
	ComJef	datos del jefe de la comisaria
DENUNCIA	NumDenuncia	identificador de la denuncia
	DncForm	Forma en que fue puesta la denuncia (verbal, escrita)
	DncCoorX	coordenada x de la ocurrencia del hecho
	DncCoorY	coordenada y de ocurrencia del hecho
	DncFecReg	fecha de registro del hecho
	DncHora	hora de registro del hecho
	DncFecha	Fecha en que ocurrió el hecho
	CodPolicia	FK de la tabla policía
	CodComisaria	FK de la tabla comisaria

	CodDenunciante	Fk de la tabla denunciante
DETALLE_DENUNCIA	NumDenuncia	FK de la tabla denuncia
	CodSubmodalidad	FK de la tabla submodalidad
	DddCont	atestado policial
SUBMODALIDAD	CodSubmodalidad	identificador de la submodalidad
	SmodDesc	descripción de la submodalidad
	CodModalidad	Fk de la tabla modalidad
MODALIDAD	Cod submodalidad	identificador de la modalidad
	ModDesc	descripción de la modalidad
USUARIO	CodUsuario	Identificador del usuario
	UusuCla	clave del usuario
	UsuFecCre	fecha de creación del usuario
	UsuTel	teléfono del usuario
	UsuFac	descripción del tipo de usuario

Fuente: elaboración propia

La base de datos construida contiene los 2741 hechos delictivos ocurridos desde el 1 de enero hasta el 31 de agosto del 2015, los cuales fueron registrados en las tres comisarías del distrito.

Según la Base de datos, la tabla principal seria la tabla Denuncia, en la cual cada registro indica las circunstancias de la denuncia.

Según la opinión de los especialistas de las comisarías de La Molina las tablas secundarias que están relacionadas a los datos personales de los denunciantes o efectivos policiales son poco relevantes y se pueden excluir del análisis. A continuación se va a dar una descripción de cada uno de los campos de dichas tablas:

- **DncLugar**: Este campo representa el lugar detallado dónde ocurrió el hecho delictivo, esto abarca las inmediaciones aproximadas donde ocurrió el hecho o la dirección del domicilio o comercio si fuera el caso.
- **DncForm**: Representa la forma en que fue puesta la denuncia (puede ser verbal o escrita).
- **DncFecReg**: Este capo está referido a la fecha exacta (con hora) en que fue realizada la denuncia del hecho.
- **DncFecha**: Es la fecha que es narrada por el denunciante (en la cual ocurrieron los hechos).

- **CodPolicia**: Representa el identificador único del policía que redacto el atestado policial.
- **CodComisaria:** Representa el identificador único de la comisaria donde se está registrando el hecho.
- **CodDenunciante:** Representa el identificador único de la persona que denuncia la ocurrencia del hecho.
- **CodSubModalidad:** Corresponde a la modalidad de la denuncia, cada denuncia tiene una modalidad específica.
- **DddCont:** Es el atestado policial que redacta el policía según lo que narra el denunciante.

3.2.3 Explorar los datos

Luego de tener la descripción de los datos, se debe saber en qué cantidades están representados; sin embargo, esto tomaría un tiempo significativo si se hiciera con cada uno de los atributos de las tablas (donde no todos son significativos para nuestro problema), por eso fundamentalmente se debe hacer esta actividad en forma de un reporte o listado inicial donde se muestren los datos que resulten más relevantes para luego poder seleccionarlos en la fase 3, ya que en dicha fase s mostrará la frecuencia y gráficos de distribución de estos datos. Con esto evitaremos repetir la información de las frecuencias de los atributos de las tablas que sean seleccionados como variables de entrada del modelo.

Cabe resaltar que de un campo de una tabla podría salir más de una variable o que la unión de dos campos podría representar una variable de entrada.

A continuación se muestran los datos como un reporte preliminar, detallando por qué fueron considerados como relevantes.

- El campo ComRef de la tabla comisaria, que representa la comisaria donde fue registrado el hecho, es una posible variable de entrada al modelo debido a que comúnmente las denuncias están registradas en la comisaria de la zona donde ocurrió el hecho.

- Lugar_de_ocurrencia: que representa el lugar (detallado o aproximado) donde ocurrió el hecho delictivo, este atributo es importante porque detalla donde se ocurren los hechos (se podrían sacar las zonas con más frecuencia de ocurrencia de hechos delictivos)
- DddCont, que contiene el atestado policial, de este campo se podrían sacar subtipos o variantes del delito (si fue con arma, en banda, etc.) esto serviría como una variable para identificar qué delito es más común.
- DncFecha, que representa la hora del hecho delictivo. Este atributo se convertiría en una variable clave para el modelo debido que con esta se pueden identificar los días del mes, días de la semana u horas donde ocurren más delitos.

- PerFecNac, que representa la fecha de nacimiento de la persona, este atributo es relevante porque podría servir al modelo a identificar que rango de edades son las principales víctimas de hechos delictivos.

3.2.4 Verificar la calidad de los datos

La información obtenida es de calidad debido a que nuestro ETL hizo la carga de la información necesaria para tener una bases de datos OLTP, sobre esta se elabora el data set que utilizará en el modelo.
Las tablas adicionales de referencia serán consideradas en la medida que describan la codificación de algún campo de interés de la tabla principal.

3.3 Fase 3 Preparación de los datos

3.3.1 Selección de datos

Los campos de interés se convertirán luego en nuestras variables de entrada del modelo de predicción, estos se deben elegir según un juicio en el cual se obtenga una relevancia importante de cada una.

A continuación se realiza una descripción de cada uno de estos campos. Para cada uno de los campos seleccionados se determinan:
• Estados posibles: valores que puede tomar cada campo;

• Descripción de cada uno de los estados;

• Frecuencia: cantidad de registros para cada estado.

Para los campos omitidos, se explicará el porqué de su omisión.

Los datos que se van a recoger y que servirán para crear el modelo de predicción de hechos delictivos son los siguientes:

- **Zona de ocurrencia:** El distrito de La Molina posee 6 grandes sectores claramente identificados, las cuales fueron determinadas por la municipalidad,
- **Zona 1 – Camacho:** Zona oeste del distrito, comprende la Urb. Camacho, Santa Sofía Magdalena, La Fontana y los Cerros de Camacho, las cuales conforma el 15% del distrito.
- **Zona 2 – La Molina Vieja:** Es la zona sur del distrito. Comprende las urbanizaciones La Molina Vieja, La Alameda de la Molina Vieja, Los Sirius y El Remanso, Corregidor, Isla del Sol, Las Viñas de La Molina, Portada del sol, La Capilla, El Valle de La Molina, Las Lomas de la Molina Vieja y constituyen un 20% del distrito.
- **Zona 3 - Santa Patricia:** Cuadrante comprendido entre la Av. La Molina y la Av. Melgarejo, La Av. Separadora Industrial y la Av. La Universidad. Incluye Covima y Santa Raquel. Representa el 30% del distrito.
- **Zona 4 - UNALM:** Comprende la Universidad Agraria de La Molina. Ocupa el 10% del distrito. Esta zona no entra en nuestro análisis (a menos que el hecho haya ocurrido en las instalaciones de la UNALM).
- **Zona 5 – Rinconada y La Planicie:** Rinconada Alta, Rinconada Baja, Rinconada del Lago, La Planicie, Huertos de La Molina y Club Campestre Las Lagunas. Este sector empieza en Molicentro y es cruzado por la Avenida La Molina y la Avenida Elías Aparicio. Conforma un 20% del distrito.
- **Zona 6 – MUSA:** Está ubicada al este del distrito. Es una zona de carácter popular, está ubicada en las últimas cuadras de la Avenida La Molina o Carretera hacia Cieneguilla. Se estima que conforma aprox. el 5% del distrito.

Como en la base de datos no tenemos especificada la zona e ocurrencia, está la hallaremos mediante las direcciones donde ocurrieron los hechos, por ejemplo si el hecho delictivo ocurrió frente a la FIA-USMP entonces sabremos que pertenece a la zona 3.

Tabla 17. Frecuencia atributo Zona

CÓDIGO	NOMBRE	FRECUENCIA
1	Camacho	526
2	La Molina Vieja	845
3	Santa Patricia	1126
4	UNALM	0
5	Rinconada y La Planicie	170
6	MUSA	67
7	NA	7

Fuente: elaboración propia

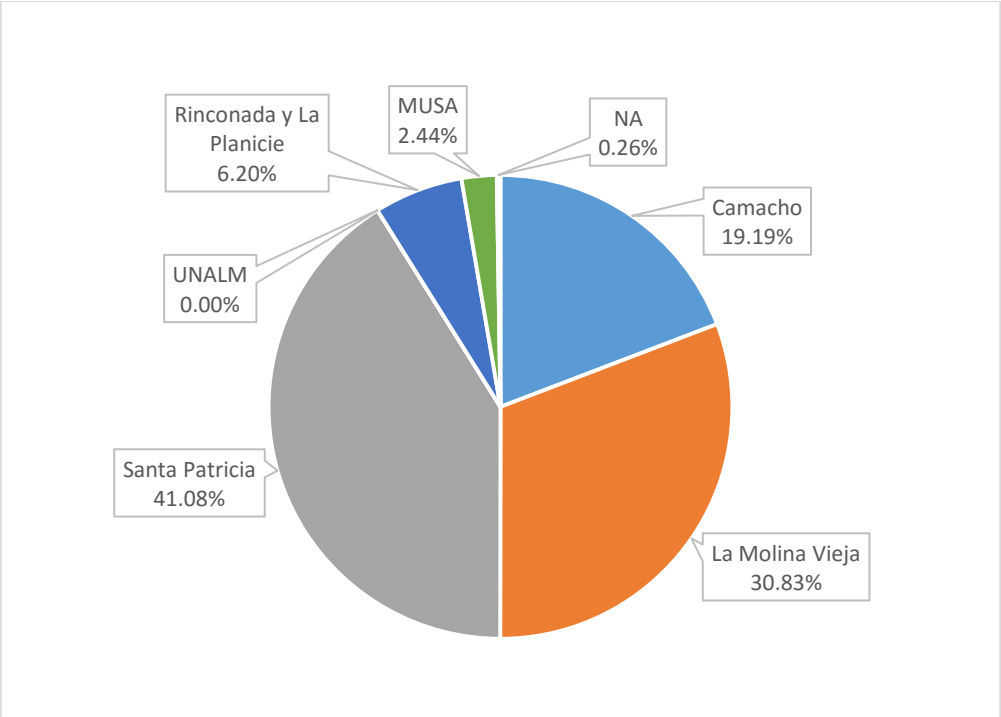


Figura 17. Distribución del atributo Zona
Fuente: Elaboración propia.

Todas las zonas presentan denuncias, excepto la zona UNALM, esto es un resultado esperado debido a que si hubiera denuncias en dicha zona, estas

serían por hechos delictivos ocurridos dentro de las instalaciones de la mencionada casa de estudios.

Mes: Este atributo describe el mes en que se cometió el hecho, considerando que solo se ha recogido data hasta el mes de agosto. La distribución es la siguiente.

CÓDIGO	MES	FRECUENCIA
1	ENERO	362
2	FEBRERO	310
3	MARZO	373
4	ABRIL	329
5	MAYO	314
6	JUNIO	360
7	JULIO	334
8	AGOSTO	359

Tabla 18. Frecuencia atributo Mes
Fuente: Elaboración Propia

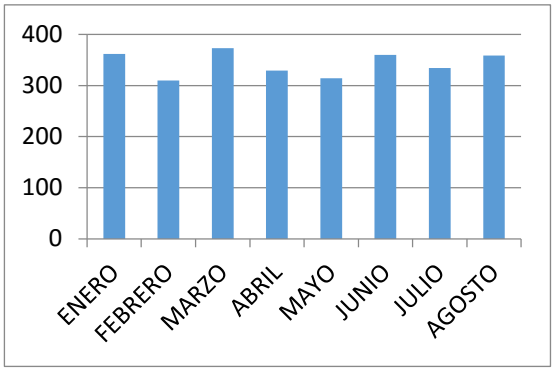


Figura 18. Distribución atributo "Mes"
Fuente: Elaboración Propia.

Día de mes: Contiene el día calendario de mes en que ocurrió el hecho. Si por ejemplo, hubo una denuncia el 1 de enero y otra el 1 de febrero, entonces la frecuencia para el día 1 será de dos.

Tabla 19. Frecuencia atributo Día de Mes.

DIA	FRECUENCIA	DIA	FRECUENCIA
1	65	17	91
2	63	18	84
3	68	19	104
4	127	20	109
5	142	21	95
6	67	22	80
7	94	23	107
8	98	24	83
9	120	25	93
10	78	26	87
11	96	27	87
12	70	28	81
13	102	29	63

14	68		30	76
15	92		31	66
16	85		**TOTAL**	**2741**

Fuente: elaboración propia

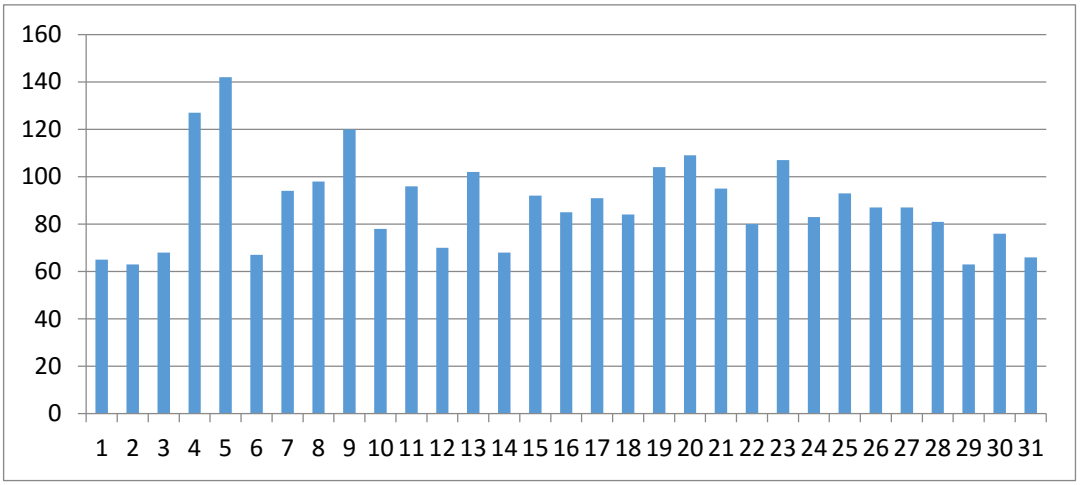


Figura 19. Distribución atributo Día de mes
Fuente: elaboración Propia.

Día de semana: Este atributo representa los días en que ocurrieron los hechos delictivos, la distribución se muestra como sigue.

Tabla 20. Frecuencia atributo Día de semana.

CÓDIGO	MODALIDAD	FRECUENCIA
1	LUNES	433
2	MARTES	379
3	MIÉRCOLES	345
4	JUEVES	413
5	VIERNES	415
6	SÁBADO	400
7	DOMINGO	356
	TOTAL	**2741**

Fuente: elaboración propia.

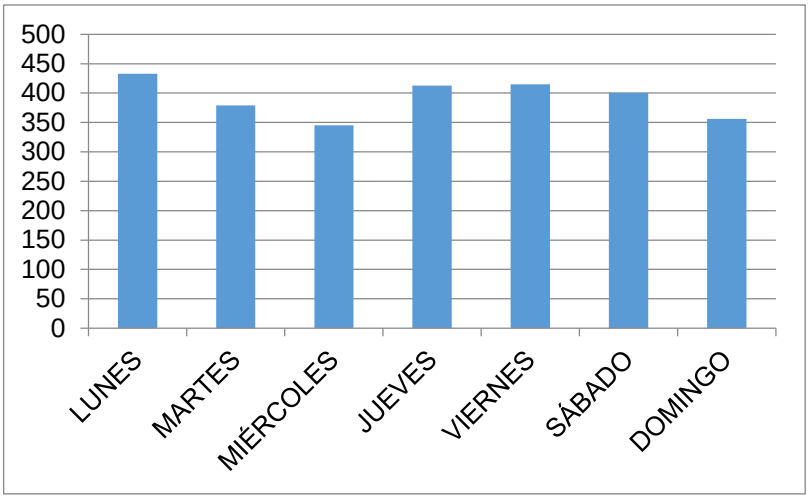


Figura 20. Distribución atributo Día de semana
Fuente: elaboración propia.

Hora del hecho: Este atributo describirá la hora exacta de ocurrencia del hecho delictivo. Por temas de modelado se ha definido rangos, que servirán para su fácil manejo al momento de introducirlo al modelo. Esto abarca:

Tabla 21. Descripción de atributo Hora del hecho

ESTADO	RANGO DE HORAS	VALOR DE ATRIBUTO	FRECUENCIA
1	Entre 0 am y 3 am	0 – 3	374
2	Entre 3 am y 6 am	3 - 6	53
3	Entre 6 am y 9 am	6 – 9	95
4	Entre 9 am y 12 m	9 – 12	404
5	Entre 12 am y 15 pm	12 – 15	433
6	Entre15 am y 18 pm	15 – 18	378
7	Entre 18 am y 21 pm	18 – 21	233
8	Entre 21 am y 24 pm	21 – 24	771

Fuente: Elaboración propia.

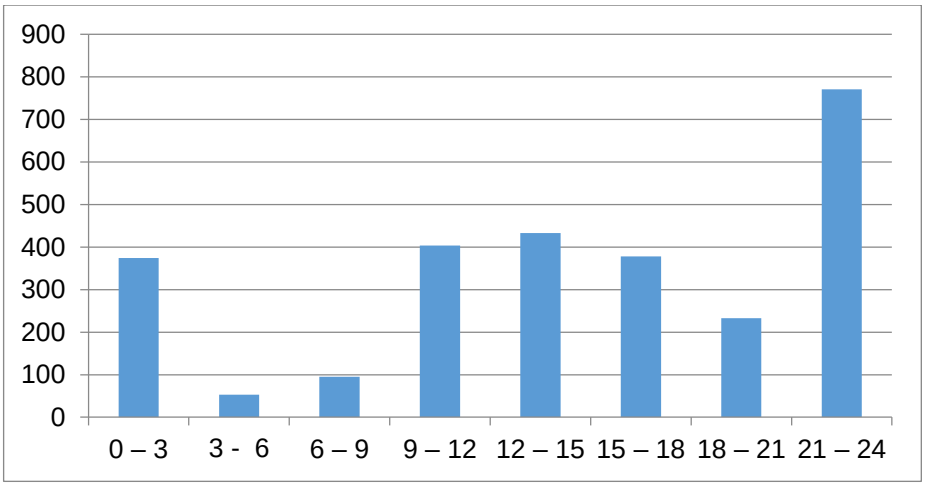


Figura 21. Distribución de atributo Hora.
Fuente: elaboración propia.

Tipo de Lugar: Contiene la clasificación del tipo de lugar donde se cometió el hecho. Los tipos son los siguientes:

Tabla 22. Descripción atributo Tipo de lugar

ID	DESCRIPCIÓN	FRECUENCIA
1	Vía pública - avenida	697
2	Vía pública - cl. o jr.	942
3	Domicilio particular	698
4	Comercio	58
5	Calzada	419
6	Otro Lugar	63

Fuente: elaboración propia.

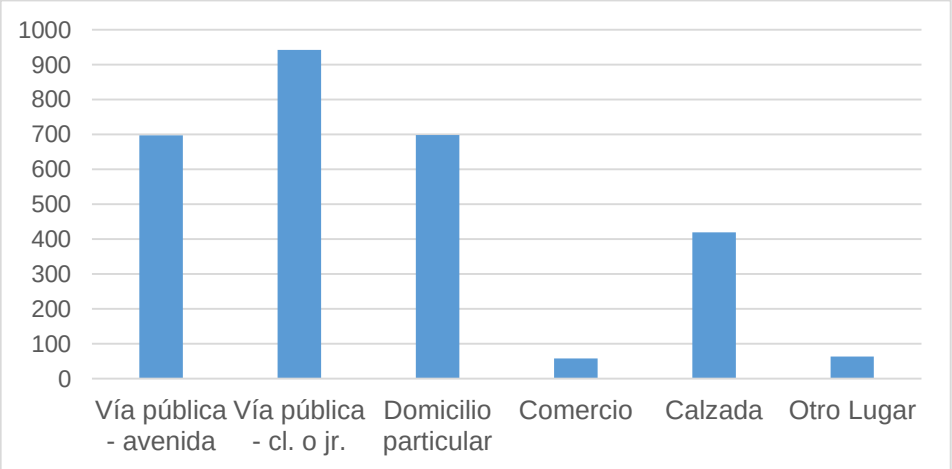


Figura 22. Distribución del atributo lugar.
Fuente: elaboración propia.

Modalidad: Este atributo corresponde a la modalidad que fue denunciada por el agraviado y que fue registrada en la comisaría correspondiente.

Tabla 23. Frecuencia atributo Modalidad.

MODALIDAD	FRECUENCIA
ABANDONO DE HOGAR	164
ABORTO	1
ACTOS CONTRA EL PUDOR	2
APROPIACION ILICITA	19
ATENTADOS CONTRA LA PATRIA POTESTAD	11
ATROPELLO	8
CHOQUE	312
DAÑO	310
DAÑOS MATERIALES	16
DELITO INFORMATICO	7
ESTAFA Y OTRAS DEFRAUDACIONES	70
EXTORSION	7
FALTAS CONTRA LAS PERSONAS	120
HOMICIDIO	1
HURTO	1122
LESIONES	85
MORDEDURA CANINA Y OTROS	17
POR FUNCIONARIOS PUBLICOS	2
ROBO	445
SECUESTRO	7
USURPACION	5
VIOLACION DE LA LIBERTAD SEXUAL	10

Fuente: elaboración Propia.

Los datos omitidos se muestran a continuación con su respectiva justificación de omisión:

Luego de tener los datos seleccionados y construidos según la herramienta seleccionada, necesitamos integrar los datos, es decir debemos tenerlos todos juntos y descartar los que no nos servirán para el modelo, en la

actividad “selección de datos” tenemos la lista de nuestros atributos relevantes, por lo tanto los siguientes atributos será descartados:

- ✓ **Comisaria:** para efectos de modelado, se omitió este atributo debido a que es indiferente si la denuncia se realizó en una comisaria u otra, lo importante es saber en qué zona ocurrió el hecho.
- ✓ **Denunciante:** Al igual que el atributo anterior, no aporta demasiado al modelo debido a que no es necesario saber quién fue víctima del hecho delictivo para armar el modelo de predicción.
- ✓ **Denunciado:** Algunos registros, según la modalidad de denuncia, contiene información acerca de la persona denunciada. Este dato se omite debido a que en su mayoría, los delitos con dolo son realizados por personas desconocidas y además no aporta significado relevante a la denuncia pues lo importante es el hecho delictivo no quien lo hizo.
- ✓ **Numero de denuncia:** No es relevante saber el identificador único de las denuncias para realizar nuestro modelo de predicción.

3.3.2 Limpieza de datos

Como se pudo ver algunos registros presentan campos incompletos, específicamente existen denuncias en las cuales no se especifica el atributo ZONA (7 registros). Para solucionar este problema, sin perder la información aportada por el resto de los campos, se adoptó el criterio sugerido por algunos especialistas de minería de datos de reemplazar la información faltante por la media o la moda del campo en cuestión, según se trate de una variable continua o categórica(Perversi, 2007).

En el caso del campo zona, este es una variable categórica; por lo tanto, se reemplazara por la moda. La modificación fue la siguiente:

- Zona: se completaron los 7 registros que no tenían especificado el tipo de lugar con la moda del campo: Zona 3.

  Por otro lado, el atributo MODALIDAD contiene registros que tienen la letra ‘ñ’, específicamente 372 registros.

Es necesario reemplazar estos caracteres para que la herramienta pueda interpretar nuestro data set sin errores.
Modalidades a reemplazar:

- Daño Físico y Psicológico
- Daño Físico
- Daño Psicológico
- Daños materiales.

3.3.3 Construir datos

En este apartado se buscó se consistentes con la terminología en común que utilizan las herramientas de minería de datos, por ello realizaron algunas modificaciones a los datos originales.

En primer lugar, las herramientas de minería de datos, por lo general, interpretan los campos con valores numéricos como atributos continuos y aquellos que presentan valores alfa-numéricos, como atributos categóricos. En nuestra base de datos, algunos de los campos seleccionados presentan valores numéricos por lo que aquellos atributos categóricos deben de ser modificados.

Por otro lado, la información temporal aportada por determinados atributos (Mes, Día de mes, Día de semana y Hora del hecho) es de naturaleza cíclica (el último estado antecede al primero). Lo que traerá problemas para su correcta interpretación por parte de algunas herramientas de minería de datos ya que a la hora de agrupar los registros, por ejemplo, se interpretaría que el lunes (día 1) está muy cerca del martes (día 2) pero muy lejos del domingo (día 7). Se modificará la escala de asignación de los valores de manera que el cambio del último al primero tenga el menor impacto posible.

Por último, los nombres de los atributos y los valores de los atributos categóricos se renombraron de forma abreviada para tener una identificación más rápida por parte de los analistas.

- **Atributo Zona:**

  Corresponde al campo Provincia. Debido a que se trata de un atributo categórico, se han reemplazado los estados originales por los siguientes:

Tabla 24. Nuevos estados del atributo Zona.

ESTADO ORIGINAL	NUEVO ESTADO	DESCRIPCIÓN
1	Cam	Camacho
2	LaMolv	La Molina Vieja
3	Spt	Santa Patricia
4	Unalm	Universidad Agraria
5	RincLaPl	Rinconada y La Planicie
6	Musa	MUSA

eFuente: Elaboración Propia.

- **Atributo Mes (mes)**

  Este atributo es continuo, pero no presenta un comportamiento estacionario. Por lo que en este caso no sería necesaria un reclasificación ya que su distribución es más aleatoria.

- **Atributo Día de mes**

  Este atributo es continuo, pero presenta un comportamiento uniforme por lo que no es necesaria su reclasificación.

- **Atributo Día de semana**

  Este atributo también es continuo, pero presenta un comportamiento uniforme por lo que no es necesaria su reclasificación.

- **Atributo Hora (hora)**

  En este caso puntual se ha decidido comenzar la escala en el intervalo 9-12 ya que existe una barrera natural entre este intervalo y el anterior (originada en el hecho de que alrededor de las 9 a.m. comienza la actividad laboral). Por esta razón se han reemplazado los estados originales por los siguientes:

Tabla 25. Nuevos estados para el atributo Hora.

ESTADO ORIGINAL	NUEVO ESTADO	DESCRIPCIÓN
1	6	Entre 0 am y 3 am
2	7	Entre 3 am y 6 am
3	8	Entre 6 am y 9 am
4	1	Entre 9 am y 12 m
5	2	Entre 12 am y 15 pm
6	3	Entre15 am y 18 pm
7	4	Entre 18 am y 21 pm
8	5	Entre 21 am y 24 pm

Fuente: Elaboración Propia.

- **Atributo tipo de lugar (lugar)**

  Corresponde al campo tipo_lugar. Debido a que se trata de un atributo categórico, se han renombrado los estados originales por los siguientes

Tabla 26. Nuevos estados para el atributo tipo_lugar.

ESTADO ORIGINAL	NUEVO ESTADO	DESCRIPCIÓN
1	Vp-Av	Vía pública - avenida
2	Vp-cl	Vía pública - cl. o jr.
3	DomPart	Domicilio particular
4	Com	Comercio
5	Calz	Calzada
6	Otro	Otro lugar

Fuente: elaboración propia.

Finalmente el data set contiene 2741 registros con 7 atributos claramente explicados y establecidos que contienen información referida a tres dimensiones: lugar de ocurrencia, circunstancia y tiempo.

Este data set será elaborado y en formato ARFF el cual es un formato propio de la herramienta WEKA para interpretaciones correctas.

3.3.4 Integrar datos

En esta actividad debemos integrar los datos construidos en la actividad anterior (“Construir datos”), si se diera el caso de que alguno de los atributos seleccionados se debe combinar con otro para formar una variable de entrada del modelo:

- Del campo DncFecha, de la tabla denuncia se obtienen tres variables que son: día de mes, día de semana y rango de horas, ya que en dicho campo se almacena la fecha y hora exacta de ocurrencia del hecho. Entonces mediante comando SQL sacamos las tres variables y le colocamos el valor que se detalló en la actividad anterior. La imagen muestra los comando sql realizados para obtener los registros con el formato elegido para nuestras variables.
- No se realizaron combinaciones de variables, ni otra actividad adicional debido a que estos atributos por si solos representan información significativa para nosotros y son las variables de entrada al modelo que estábamos buscando.

3.3.5 Formatear datos

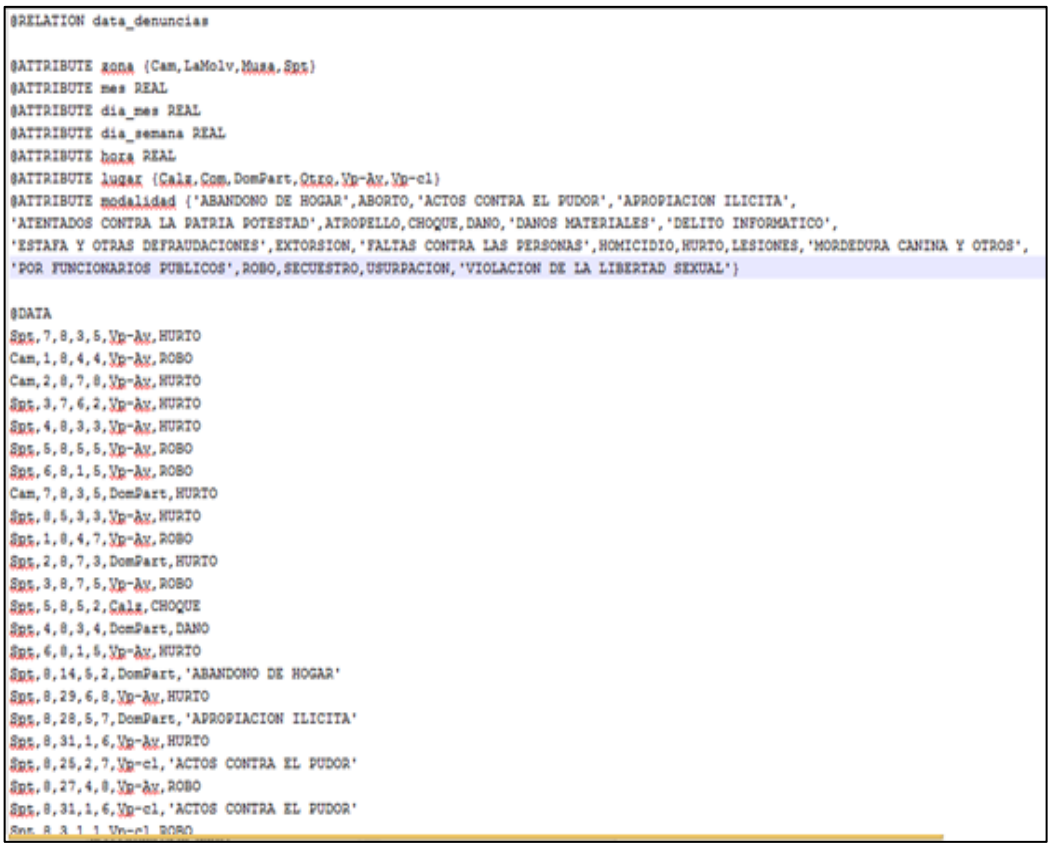
El paso siguiente, para llegar a la construcción del modelo, es la elaboración del Data Set. Se denomina Data Set al conjunto de datos a analizar con el software de minería de datos. Nuestro data set debe tener el formato que propio con el cual trabaja la herramienta WEKA, es decir un archivo con extensión .arff.

El proceso de conformación del Data set a partir de una base de datos involucra diversas etapas:

- Debemos consolidar la información en una única tabla.
- Se deben seleccionar los atributos que serán relevantes para el modelo de predicción (campos de las tablas que ya fueron identificados en actividades anteriores).
- Si fuera el caso, se deben descartar registros que tengan datos incompletos (los datos ya fueron limpiados y validados).

- Por último se debe modificar las variables de entrada en función del software a utilizar, es decir darle el orden y formato que existe para que la herramienta pueda interpretarlo de manera óptima.

Luego de los pasos realizados nuestro data set quedo de la siguiente manera:



```
@RELATION data_denuncias

@ATTRIBUTE zona {Cam,LaMolv,Musa,Spt}
@ATTRIBUTE mes REAL
@ATTRIBUTE dia_mes REAL
@ATTRIBUTE dia_semana REAL
@ATTRIBUTE hora REAL
@ATTRIBUTE lugar {Calz,Com,DomPart,Otro,Vp-Av,Vp-cl}
@ATTRIBUTE modalidad {'ABANDONO DE HOGAR',ABORTO,'ACTOS CONTRA EL PUDOR','APROPIACION ILICITA',
'ATENTADOS CONTRA LA PATRIA POTESTAD',ATROPELLO,CHOQUE,DANO,'DANOS MATERIALES','DELITO INFORMATICO',
'ESTAFA Y OTRAS DEFRAUDACIONES',EXTORSION,'FALTAS CONTRA LAS PERSONAS',HOMICIDIO,HURTO,LESIONES,'MORDEDURA CANINA Y OTROS',
'POR FUNCIONARIOS PUBLICOS',ROBO,SECUESTRO,USURPACION,'VIOLACION DE LA LIBERTAD SEXUAL'}

@DATA
Spt,7,8,3,5,Vp-Av,HURTO
Cam,1,8,4,4,Vp-Av,ROBO
Cam,2,8,7,8,Vp-Av,HURTO
Spt,3,7,6,2,Vp-Av,HURTO
Spt,4,8,3,3,Vp-Av,HURTO
Spt,5,8,5,5,Vp-Av,ROBO
Spt,6,8,1,5,Vp-Av,ROBO
Cam,7,8,3,5,DomPart,HURTO
Spt,8,5,3,3,Vp-Av,HURTO
Spt,1,8,4,7,Vp-Av,ROBO
Spt,2,8,7,3,DomPart,HURTO
Spt,3,8,7,5,Vp-Av,ROBO
Spt,5,8,5,2,Calz,CHOQUE
Spt,4,8,3,4,DomPart,DANO
Spt,6,8,1,5,Vp-Av,HURTO
Spt,8,14,5,2,DomPart,'ABANDONO DE HOGAR'
Spt,8,29,6,8,Vp-Av,HURTO
Spt,8,28,5,7,DomPart,'APROPIACION ILICITA'
Spt,8,31,1,6,Vp-Av,HURTO
Spt,8,26,2,7,Vp-cl,'ACTOS CONTRA EL PUDOR'
Spt,8,27,4,8,Vp-Av,ROBO
Spt,8,31,1,6,Vp-cl,'ACTOS CONTRA EL PUDOR'
Spt,8,3,1,1,Vp-cl,ROBO
```

Figura 23. Data de aprendizaje en formato arff
Fuente: elaboración propia

3.4 Fase 4: Modelado de datos

3.4.1 Selección de Técnicas de modelado.

Tabla 27. Comparativa de técnicas de modelado.

CRITERIO	RNA	REDES BAYESIANAS	ÁRBOLES DE DECISION	REGRESIÓN LINEAL
Probabilidad de variable(s) inicial(es)	No Necesaria	Necesaria	Necesaria	Necesaria

Independencia de variables de entrada	Cada variables es tratado como un nodo independiente	Nodos condicionados. Existen nodos padres y nodos hijos	Naturaleza condicionada.	Relaciona una variable dependiente con varias independientes
Permitir permite "n" variables de entrada	Sí, en todos los casos	En ciertos escenarios	Tiene como naturaleza construcciones lógicas basadas en premisas	Modelo insuficiente al presentar más de dos variables.
Presente en la mayoría de software	Sí	Sí	Sí	Sí

Fuente: elaboración propia

Según nuestro problema planteado y lo que queremos determinar, tendremos que utilizar un tipo de algoritmo, que cumpla los siguientes criterios:

- Algoritmo supervisado, en el cual se introducen unos valores de entrada a la red, y los valores de salida generados por esta se comparen con los valores de salida correctos. Si hay diferencias, se ajusta la red en consecuencia (se ajusta volviendo a ejecutar las entradas “n” veces o definiendo un límite de error establecido).
- Inicialmente no se tendrá probabilidades de las variables pues no se sabe que variable tiene más peso cuando se comete un hecho delictivo.
- Modelo que permita entrada de más de dos variables.
- Se sabe que las variables para la ocurrencia de un hecho delictivo no estas relacionadas unas con las otras, es decir, estas son independientes.

Con esto criterios mencionados y haciendo el análisis de los algoritmos existentes, vemos más conveniente utilizar la técnica RNA, pues cumple con lo mencionado en una medida más grande que los demás.

La técnica de modelado elegida es RNA (Redes neuronales artificiales) que son un paradigma de aprendizaje automático inspirado en la forma en que funciona el sistema nervioso de los animales. Se trata de un sistema de interconexión de neuronas que colaboran entre sí para producir un estímulo de salida.

La tipología a utilizar será el perceptrón multicapa, la cual es una red neuronal artificial (RNA) formada por múltiples capas, esto le permite resolver problemas que no son linealmente separables, lo cual es la principal limitación del perceptrón simple (otra tipología del RNA). El perceptrón multicapa puede ser totalmente o localmente conectado, en ella cada neurona de la capa "i" es entrada de una serie de neuronas (región) de la capa "i+1"

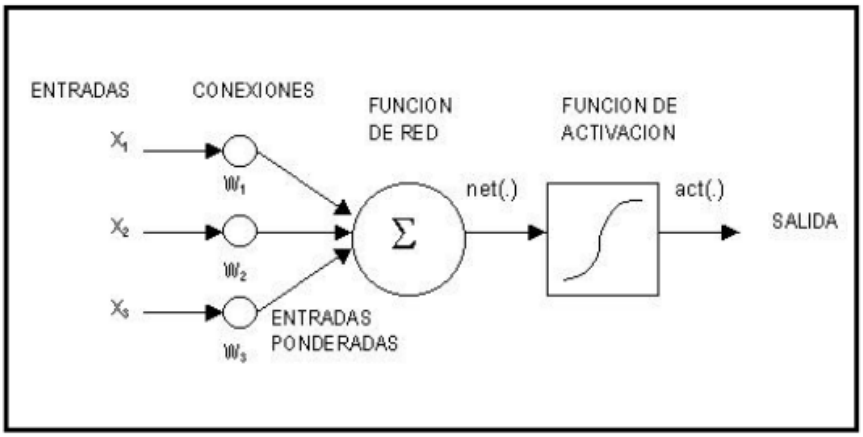


Figura 24. Perceptrón multicapa
Fuente: (Russel & Norvig, 2004)

La red está definida por neuronas, conexiones de entrada (donde cada una posee un peso que no necesariamente es el mismo por cada entrada), una función de propagación que se encarga de computar la entrada total de las conexiones, un núcleo central que activa la función de activación y una salida.

La función de propagación calcula la entrada total a la red, esta se calcula al sumar las entradas multiplicadas por sus pesos (Xi * wi). Esto equivale a la

combinación de las señales excitatorias o inhibitorias si queremos hacer una analogía con las neuronas biológicas que poseen los animales.

La función de activación es la que mejor define el comportamiento de las neuronas, esta se encarga de definir el nivel o estado de activación según las entradas totales.

Esta función debe ser una no lineal y es comúnmente conocida como la función sigmoide que define los límites de salida; es decir, cuando la suma ponderada sea muy grande, ajustara el valor de salida cercano a 1 y cuando sea muy pequeño ajustara el valor cercano a 0 (ó -1 según la función que se esté utilizando).

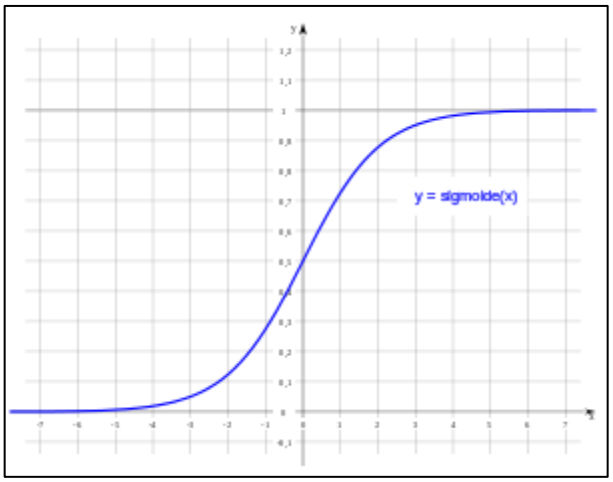


Figura 25. Función Sigmoide.
Fuente: (Russel & Norvig, 2004)

Las capas pueden clasificarse en tres tipos:

- Capa de entrada: Constituida por aquellas neuronas que introducen los patrones de entrada en la red. En estas neuronas no se produce procesamiento.
- Capas ocultas: Formada por aquellas neuronas cuyas entradas provienen de capas anteriores y cuyas salidas pasan a neuronas de capas posteriores.
- Capa de salida: Neuronas cuyos valores de salida se corresponden con las salidas de toda la red.

3.4.2 Generar Diseño de pruebas

Nuestro diseño de pruebas a realizar tendrá las siguientes consideraciones:
El 85% del total de los datos recolectados serán usados como data de test, esto para comprobar en qué medida nuestro modelo puede predecir hechos. Según los resultados, se creara la matriz de confusión para determinar que valores arrojo el modelo cuando la predicción fue errada.

También se creara la matriz de confusión para saber el porcentaje de aciertos según cada valor de la clase (individualmente.)

3.4.3 Construcción del modelo

Luego de haber hecho la comprensión y la selección de los datos, y establecimiento del algoritmo a utilizar, se procede a realizar el data set, en formato ARFF (ya que nuestra comparativa arrojó que WEKA es la herramienta que más se acomoda a nuestro proyecto).

El formato ARFF contiene tres bloques:

- El primero es el nombre de nuestro workspace que va al lado del atributo @RELATION.
- El segundo, que esta con lado del tag @ATTRIBUTE, define los atributos a utilizar y su respectivo tipo de dato.
- Finalmente bajo el tag @DATA se enumera cada registro de dato en una línea.

Para realizar el modelo se seleccionó el 85% de nuestro data set como data de entrenamiento, esta data será utilizada para crear el modelo, es decir el aprendizaje.

El 15% restante será utilizado como data para validación del modelo (Esta validación se detallara más adelante en la fase de pruebas). Antes de detallar el paso a paso del entrenamiento de los datos tenemos que explicar que es lo que ocurrirá internamente al insertar nuestros datos en la herramienta WEKA y que es lo que arrojara como resultado.

Primero: Se debe seleccionar cual de nuestros atributos será nuestra clase. En inteligencia artificial se conoce como clase a la variable que queremos predecir, esta puede ser simple (dos valores) o múltiple (varios valores).
En este caso será una multiclase pues se ha elegido como clase el atributo ZONA.

Segundo: Se debe definir las épocas. En IA se conoce como épocas a las iteraciones realizadas con la data de entrenamiento. Según () con un mínimo de 50 épocas podríamos tener una probabilidad de aceptación de más o menos 50%.

Entonces, al ingresar cada registro a la herramienta WEKA, lo que sucede internamente es lo siguiente:

- Para cada registro, cada uno de sus atributos son las neuronas de entradas que tendrán definido un número específico. Inicialmente la red intentara llegar predecir un valor de salida, pero como esta no ha aprendido nada aun entonces arrojara un resultado distinto al que se tiene. Entonces lo que se busca es hallar el error en cada caso.
- Esto inicialmente no dará ninguna predicción puesto que la data se ha ingresado una sola vez, y aquí es donde viene lo denominado **épocas.** En cada época se buscara ir reduciendo el error, ingresando el valor hallado en la época anterior y así conseguir un modelo altamente preciso.
- Luego se suma la cantidad de predicciones acertadas dividida entre la cantidad total de registros y se obtiene el porcentaje de qué tan preciso es el modelo.
- Finalmente se necesita comprobar que este ratio porcentual es verdadero, entonces aquí se mide el modelo con la data de test y se comprueba la eficacia del modelo.

Figura 26. Ventana principal WEKA.
Fuente: elaboración Propia.

Se seleccionó la opción Explorer del menú Applications

Figura 27. Opción Explorer
Fuente: elaboración Propia.

Dentro de la ventana del Explorer de Weka se visualizan distintas pestañas para realizar distintos procesos de minería de datos (Preporcess, Classify, Cluster, Associate, Select attributes y Visualize). El Explorer se inicia en la pestaña Preprocess

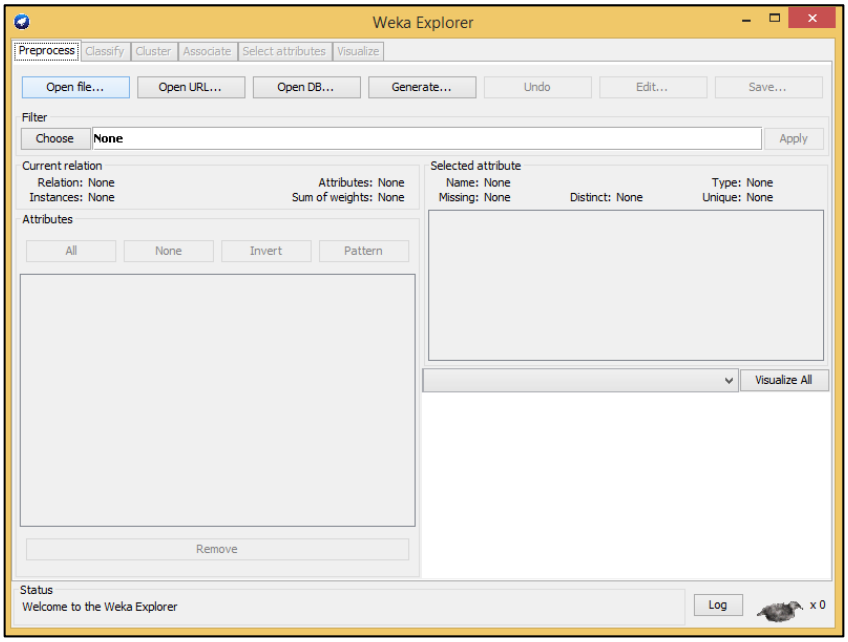


Figura 28. Ventana Explorer
Fuente: elaboración Propia.

Dentro de esta pestaña se seleccionó el botón Open file… y luego se buscó el archivo del data set en formato ARFF.

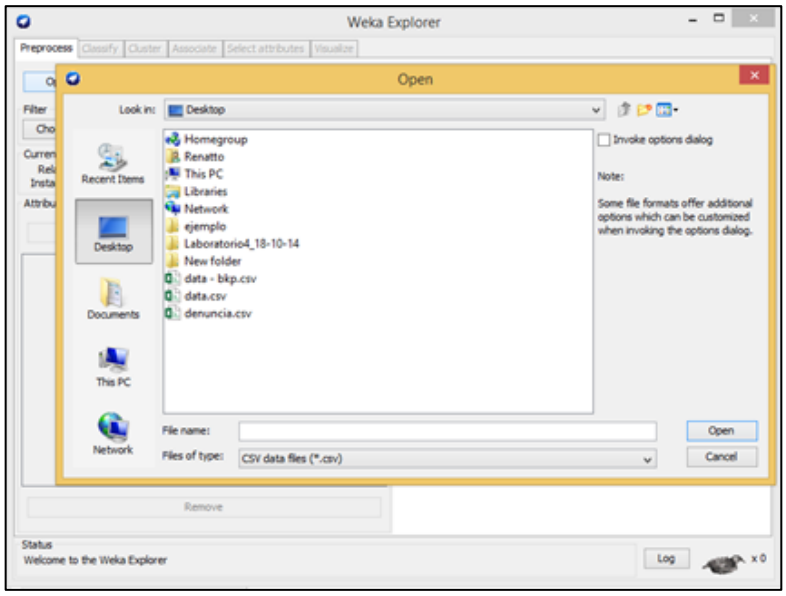


Figura 29. Carga archivo arff
Fuente: elaboración Propia.

Cuando se selecciona el archivo se pueden ver gráficos de frecuentas de los valores únicos de cada variable. Esto para tener una idea que nuestra data es consistente y ha llegado a la herramienta de la manera esperada.

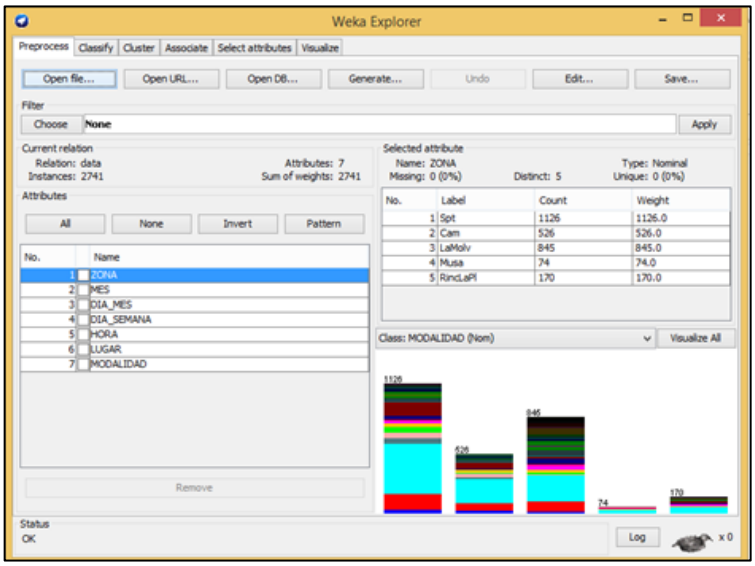


Figura 30. Data cargada a Weka.
Fuente: elaboración Propia.

Luego seleccionamos la pestaña Classify, para poder realizar el modelo, escogemos el algoritmo RNA, específicamente, Multilayer Perception.

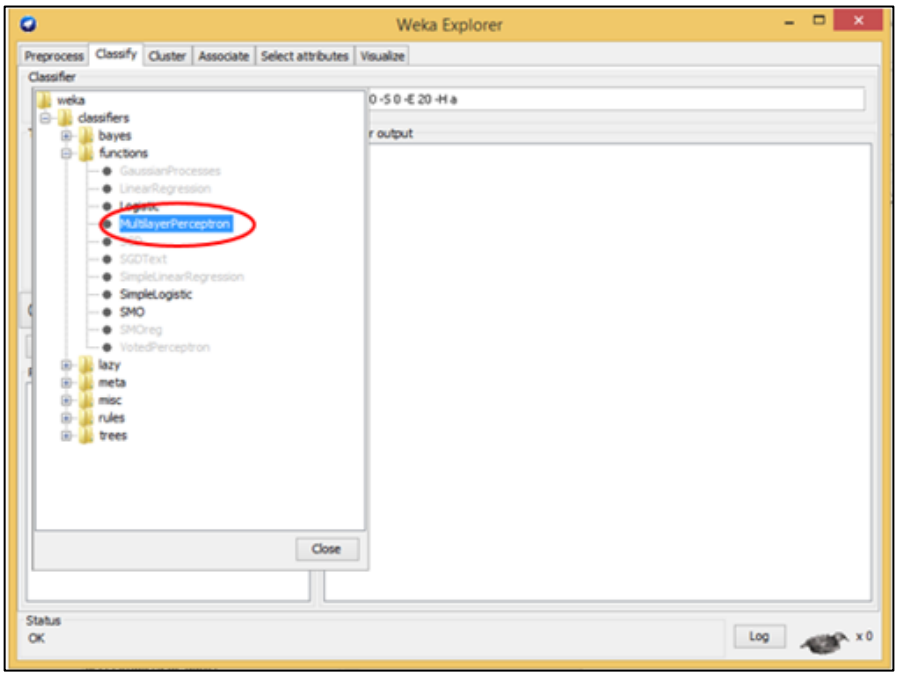


Figura 31. Elección del algoritmo RNA.
Fuente: elaboración Propia.

Luego, con la data cargada y el algoritmo seleccionado, corremos el modelo para que se realice el aprendizaje.

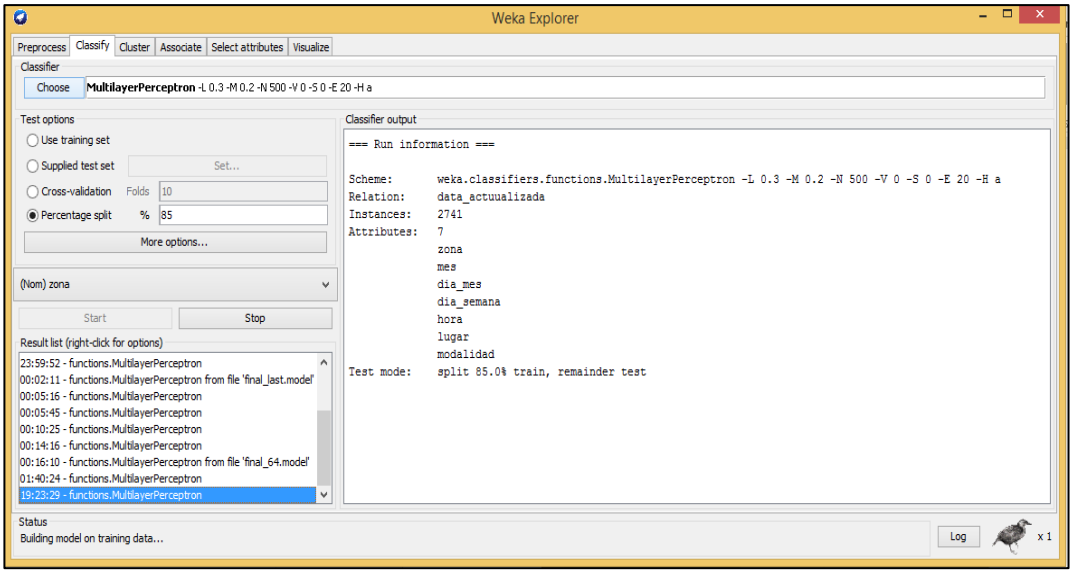


Figura 32.Corrida del algoritmo.
Fuente: elaboración Propia.

Resultado final del aprendizaje para colocar en servicios web, ser consumido por los mapas temáticos y hacer las validaciones respectivas.

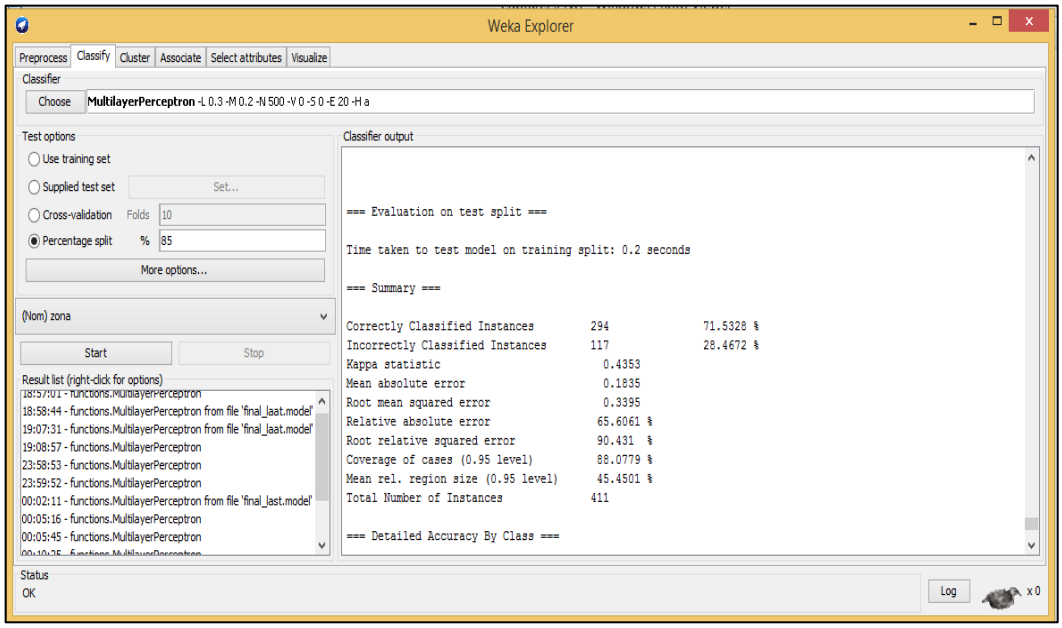


Figura 33. Resultados de la ejecución.
Fuente: elaboración Propia.

3.4.4 Evaluar el modelo

En este apartado haremos la evaluación al modelo generado según dos análisis básicos: primero evaluando los objetivos de minería de datos, específicamente lo criterios claves de éxito planteados en la fase 1 y

segundo, aplicando el criterio de evaluación de modelos de MD Según Witten, Frank y Hall (2011).

Según los objetivos de MD, la evaluación es la siguiente:

- Obtener un mínimo de 60% de probabilidad de acierto en las predicciones, viendo que se obtuvo más de un 70% de probabilidad de acierto del modelo creado, podemos decir que esta evaluación es satisfactoria pues se pasó la probabilidad mínima establecida.

- Tener un detalle de las zonas con los delitos más comunes con su probabilidad de ocurrencia, los delitos más comunes para las predicciones fueron obtenidos según las frecuencias de cada uno encontrada en la data histórica. Adicionalmente, las predicciones arrojan la probabilidad de ocurrencia estimada. La evaluación de este criterio se ve favorable al igual que el anterior.

Al evaluar estos criterios, tenemos el primer indicador que el modelo creado es cumple los criterios establecidos inicialmente.

Por otro lado, según Witten, Frank y Hall (2011), el resultado del algoritmo ZeroR debe ser superado por cualquier aplicación de otro algoritmo en un modelo de minería de datos". Esta afirmación se sustenta básicamente debido a que "el ZeroR clasifica todos los datos con la clase mayoritaria; es decir, si el 90% de los datos son negativos y el otro 10% son positivos, entonces el algoritmo clasificará todo como negativos.

Aplicando ZeroR podemos ver el resultado que se muestra a continuación:

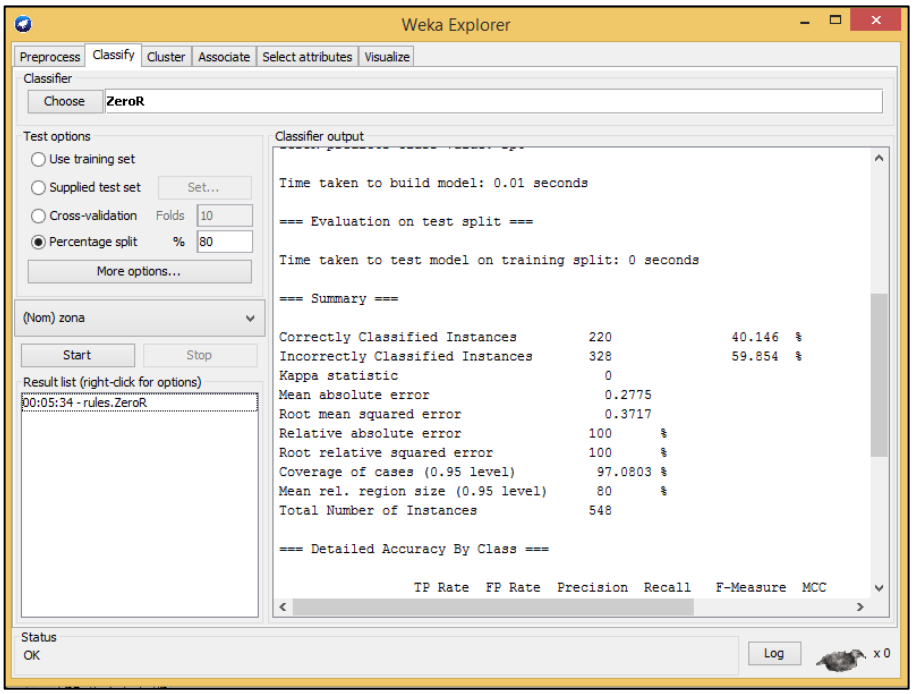


Figura 34. Aplicación de algoritmo ZeroR
Fuente: Herramienta Weka

Como se puede ver, el porcentaje de acierto es menos al obtenido por nuestro modelo de redes neuronales (40.146% contra 71.53%), esto quiere decir que nuestro modelo supera el porcentaje mínimo aceptado y los resultados pueden ser aceptados para un análisis posterior en la siguiente fase de la metodología en donde se realiza otra evaluación pero en ese caso es realizada a nivel de resultados.

3.5 Fase 5 Evaluación

3.5.1 Evaluar los resultados

Los resultados de nuestro modelo de predicción tienen una más alta probabilidad de fallo según diversos criterios, los cuales son:

- Si los datos no son representativos, es decir la muestra no abarca representativamente todos los posibles estados del problema a evaluar.
- Si el número de épocas no es lo suficientemente grande para entrenar los datos.
- Si se excede en el número de capas ocultas (las que procesan la información). Esto debido a que si se tiene más capas ocultas para afinar la probabilidad de acierto, entonces el modelo solo responderá y

predecirá de manera correcta para datos iguales a los que se ha entrenad mas no para datos nuevos diferentes.

Los resultados se han disgregado para analizar cada valor individual de la clase llamada zona, esto nos servirá para poder ver en qué zona específica nuestro modelo ha cometido más errores. Como ya se sabe el 15% de la data se utilizó para validar el modelo, es decir de test. La tabla de confusión muestra los las predicciones correctas (en cantidades y porcentaje) y, adicionalmente muestra hacia qué zona se dio como resultado cuando la predicción no fue la correcta.

Tabla 28. Matriz de confusión

		CLASES VERDADERAS					TOTAL
		Cam	LaMolV	Musa	Spt	RincyLp	
CLASIFICADO COMO	Cam	**73**	9	1	3	0	86
	LaMolV	14	**71**	0	39	2	126
	Musa	2	0	**5**	4	1	12
	Spt	5	16	1	**137**	2	161
	RincyLp	1	7	1	9	**8**	26
TOTAL		95	103	8	192	13	**411**
Precision (%)		0.8488	0.5635	0.4167	0.8509	0.3077	**0.7153**

Fuente: elaboración propia

Haciendo un análisis a al resultado que arroja nuestro modelo podemos ver que en la zona de Rinconada y La Planicie y Musa no se ha tenido una gran cantidad de aciertos.

Esto a prior puede ser por causa de dos factores fundamentales:

- Debido a que la cantidad de denuncias en este año 2015 en dichas zonas no han sido muy considerable.
- El modelo ha aprendido específicamente para esos limitados casos, haciendo más complicado acertar en nuevos escenarios.

Este resultado es esperado, debido a que el modelo es una abstracción de la realidad y por tal motivo contiene variables limitadas de entrada, entonces teniendo un 70% de acierto haciendo la sumatoria de todas las zonas habríamos alcanzado un resultado aceptable.

Debido a que la zona de “rinconada y la planicie” presenta menos denuncias que santa patricia y Camacho, entonces podemos decir que las dos últimas zonas están más propensas a tener hechos delictivos, y nuestro modelo tiene una probabilidad más grande de aciertos en dichas zonas (más del 80%).

3.5.2 Revisar el proceso

Según los resultados obtenidos, podemos decir que nuestro proceso fue realizado correctamente, más del 70% de las predicciones fueron correctas, lo que da una alta tasa de aciertos, determinando así que podríamos decir que la selección de variables y clase fueron las adecuadas. Cabe señalar que es poco probable que un modelo dé la predicción del 100% debido a que s esto ocurriera, la data entrenada solo estaría preparada para predecir casos en los que son similares a la data que se usó de training.

3.5.3 Determinar los siguientes pasos

- Los siguientes pasos luego de la creación del modelo de predicción son:
- Realizar un análisis para determinar cómo mostrar la información.
- Determinar tecnologías a utilizar para mostrar información precedida en mapas temáticos.
- Estimar tiempos de implementación de visor web y móvil del sistema de predicción.
- Desarrollar el sistema de predicción apto para web y móvil.
- Realizar un plan de pruebas para asegurar la calidad del sistema, mostrando a información predicha.
- Entregar el sistema al usuario final.

3.6 Despliegue

Para la fase de despliegue se ha optado por mostrar los datos recogidos en el modelo de predicción en a través de dos plataformas: Web y móvil, por ende SISTPRE poseerá dos líneas de desarrollo, que tendrán las mismas funcionalidades sin embargo la parte de navegabilidad e Interfaces de Usuario variarán.

3.6.1 Plan de despliegue

El plan de despliegue describe las actividades necesarias para que nuestro modelo de predicción sea plasmado en un sistema (móvil y web), tal como lo definimos en el alcance del proyecto en capítulos anteriores.

El Anexo 6 muestra el documento que se generado en esta actividad que tiene el mismo nombre de la actividad: “Plan de despliegue”.

3.6.1.1 Desarrollo del Sistema de Predicción

Para especificar el desarrollo del sistema se realizaran solamente cuatro fases principales

- **Captación de Requerimientos**

Tabla 29. Requerimientos para el sistema de predicción

N	REQUERIMIENTOS	WEB	MÓVIL
1	Ver el número de denuncias registradas por meses	✓	✓
2	Ver el número de denuncias por tipo y Comisaria	✓	✓
3	Ver el número de denuncias por tipo y Zona	✓	✓
4	Ver zonas de La Molina en mapa geográfico	✓	✓
5	Ver Zonas de Santa Felicia	✓	x
6	Ver Predicciones de La Molina	✓	✓
7	Ver Predicciones de Santa Felicia	✓	x
8	Refrescar Mapa	x	✓
9	Ver Portal Informativo	✓	x
10	Ver modalidades de denuncia por comisaria	✓	✓

Fuente: elaboración propia

Esta tabla nos muestra las distintas funcionalidades que poseerá el SISTPRE indicando cuales se encontraran disponibles en cada tipo de plataforma a desarrollar, además nos sirve para tener en claro las primeras actividades a desarrollar.

- **Elección de Tecnologías**

Para la elección de las herramientas de desarrollo se ha tenido en cuenta dos aspectos importantes que son:
Cuota de Mercado y Costo de Equipos: Este elemento fue primordial en la elección de la versión móvil de SISTPRE, ya que Android posee la mayor cuota del mercado a nivel Mundial y además posee una gran cantidad de equipos a precios mucho más accesibles que IOS.

La siguiente figura muestra los sistemas operativos móviles más usados, donde se ve que Android es el que domina el mercado, por lo cual será elegida para el desarrollo.

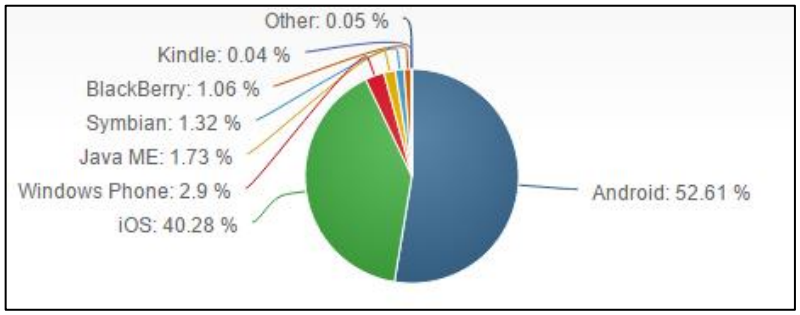


Figura 35. Cuota de mercado de SO móviles hasta Octubre 2015.
Fuente: TechMarket Share

Para la versión móvil no se tuvo mucha discusión ya que toda la web funciona en base a tres componentes universales que son HTML, CSS y Javascript, por ello en el lado del cliente (Frontend) se están utilizando estas tecnologías.

Flexibilidad: Con Flexibilidad nos referimos al manejo del intercambio de información que ocurren entre cliente-sistema, el nexo entre ambas en programación se le conoce como Backend o lenguaje del lado del servidor y

este se encarga además de mantener actualizada la base de datos donde persisten las transacciones realizadas.

Por ella para la versión móvil se ha optado por un desarrollo nativo en Java 7 y por el lado de la versión Web se usara NodeJS y Express que son librerías hechas de Javascript.

NOTA: Ambas aplicaciones se conectan a una Base de Datos Mysql 5.6 donde se almacena el resultado del modelo predictivo arrojado por Weka.

- **Diseño y Prototipos**

  Para la fase de diseño lo primero que se realizó fue en la creación de Mockups (Prototipos Iniciales) del Sistema de Predicción, para ello se hizo uso de la versión trial de Balzamiq Mockups y MockFlow Wireframe. A continuación se mostrarán los distintos prototipos iniciales del SISTPRE.

  Versión Móvil:

Figura 36. MockUp mapa predictivo – Versión móvil
Fuente: elaboración Propia

Figura 37. MockUp reportes – Versión móvil
Fuente: elaboración Propia

Versión Web

- **Portada**

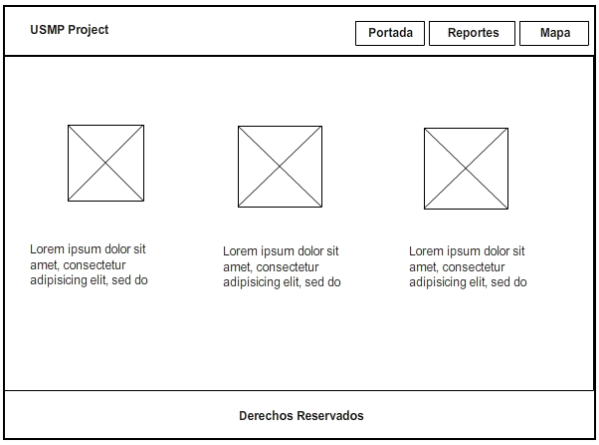


Figura 38. MockUp Portada – Versión Web
Fuente: elaboración Propia

- **Reportes**

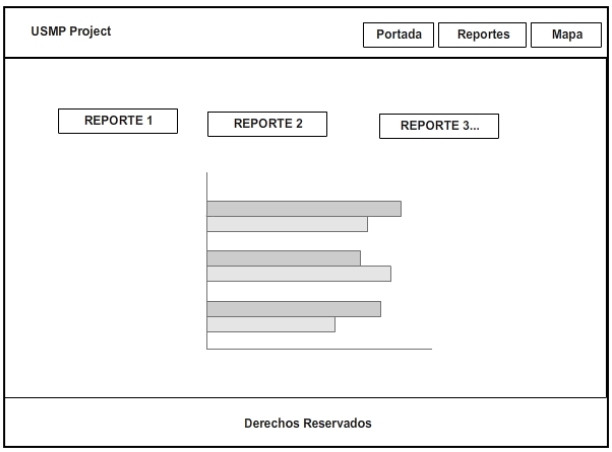


Figura 39. MockUp reportes – Versión Web
Fuente: elaboración Propia

- **Mapa**

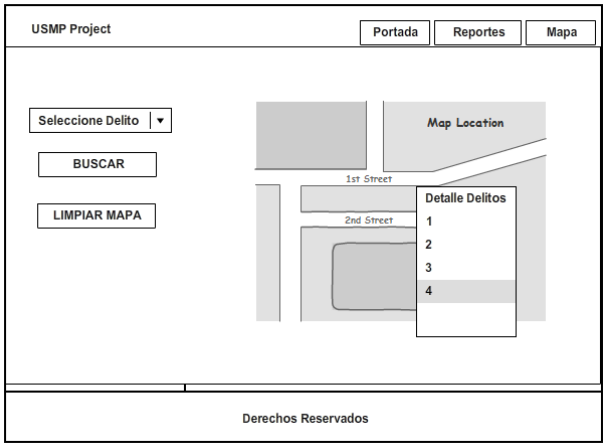


Figura 40. MockUp Mapa – Versión Web
Fuente: elaboración Propia

- **Historias de Usuario**

Las historias de usuario mostradas

HISTORIA DE USUARIO		
Número de Historia	9	Usuario: Policía
Nombre de Historia	Visualizar Portal Informativo	
Importancia	40	
Responsable	Jaulis Rua Jorge Julio	
Descripción		
Esta funcionalidad permite mostrar información genérica perteneciente a la municipalidad de La Molina donde se hace énfasis en la seguridad ciudadana para ello este cambia a determinadas horas del día para mostrar distinta información, para el desarrollo de este se ha utilizado un diseño responsivo para que se pueda ajustar a cualquier tamaño de pantalla.		
Validación		
El usuario será capaz de visualizar información genérica de las novedades que ocurren en el distrito de La Molina y servirá para mantener informado al usuario de los últimos avances que ocurren a lo largo de todas las comisarías existentes.		
Interfaz Web		
		

Figura 41. Historia de usuario visualizar portal inf.
Fuente: elaboración propia

HISTORIA DE USUARIO		
Número de Historia	1,2,3,10	Usuario: Policía
Nombre de Historia	Generar Reportes	
Importancia	80	
Responsable	Jaulis Rua Jorge Julio	
Descripción		
Esta historia de usuario permite visualizar una cantidad determinada de reportes, cada uno tiene distintos tipos de filtros que va permitir hacer más dinámico el momento de la visualización de los datos, cabe resaltar que los datos mostrados se han realizado en base a las denuncias registradas por cada una de las distintas comisarías del distrito de La Molina, actualmente la data histórica datan desde el mes de enero, hasta el mes de agosto de este año.		
Validación		
El usuario tendrá acceso al menú de Reportes donde visualizará una cantidad de reportes, en donde podrá filtrar la información a través de parámetros de entrada el cual determinara la información a mostrar por dicho reporte		
Interfaz Web		

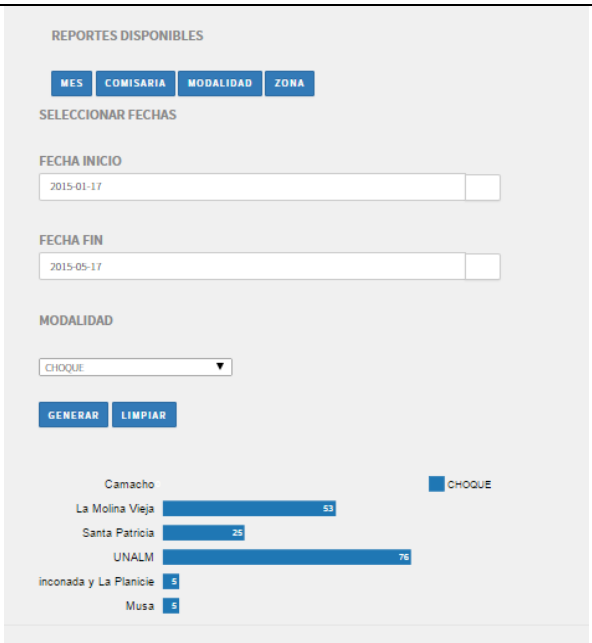


Figura 42. Historia de usuario Generar Reportes.
Fuente: elaboración propia

HISTORIA DE USUARIO		
Número de Historia	6,7	Usuario: Policía
Nombre de Historia	Visualizar Mapa Predictivo	
Importancia	90	
Responsable	Jaulis Rua Jorge Julio	
Descripción		
Esta función se encuentra en el Menú Mapa predictivo, donde muestra distintas zonas de alto riesgo para ello, se sombrea en el mapa con un color rojo, dicha zona también se podrá ver un detalle histórico donde se verán las denuncias.		
Validación		
El usuario será capaz visualizar las zonas propensas a ocurrencia donde exista gran probabilidad de riesgo a que ocurra un evento delictivo esto servirá para que las comisarías puedan realizar distintas medidas preventivas.		
Interfaz Web		

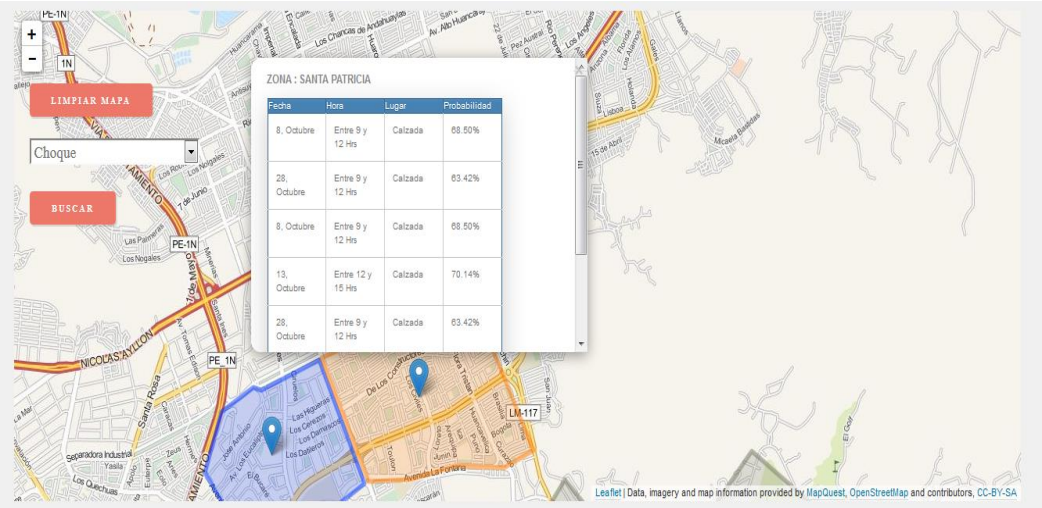


Figura 43. Historia de usuario Visualizar Mapa Predictivo
Fuente: elaboración propia.

3.6.2 Plan de monitoreo y mantenimiento

Este plan detalla que actividades a realizar para que el sistema implementado no deje de tener valor en el tiempo; es decir, que siga siendo clave para el proceso de prevención del delito. Esto incluye dar soporte a la solución, mantener la tecnología operativa (Base de datos, hosting, etc.).
El **Anexo 7** muestra el plan de monitoreo y mantenimiento de nuestro sistema de predicción de hechos delictivos.

3.6.3 Producir reporte final

Como se ha mencionado anteriormente en base a la creación del modelo de predicción mostraremos los hechos delictivos con sus zonas predichas a través de un mapa temático que se visualizara tanto en tecnología web como en móvil.

Además se tendrá un apartado de reportes donde se podrán observar los distintos gráficos de barras de acuerdo a la información obtenida en las distintas comisarías, esta información es obtenida a partir de la data histórica recolectada.

Reporte final

La funcionalidad de los reportes, esta se ha hecho tomando la información del registro de denuncias pertenecientes a las distintas comisarías de la Molina (Santa Patricia, La Molina, Las Praderas). A continuación las siguientes capturas de pantalla pertenecen al SISTPRE versión Web. En todos los casos se tiene filtros de fechas.

- Reporte 1 Ver delitos por meses.
- Reporte 2: Modalidades de denuncia por comisarias
- Reporte 3: Modalidades de denuncias por zonas

Reporte 01: Registro de delitos mensuales

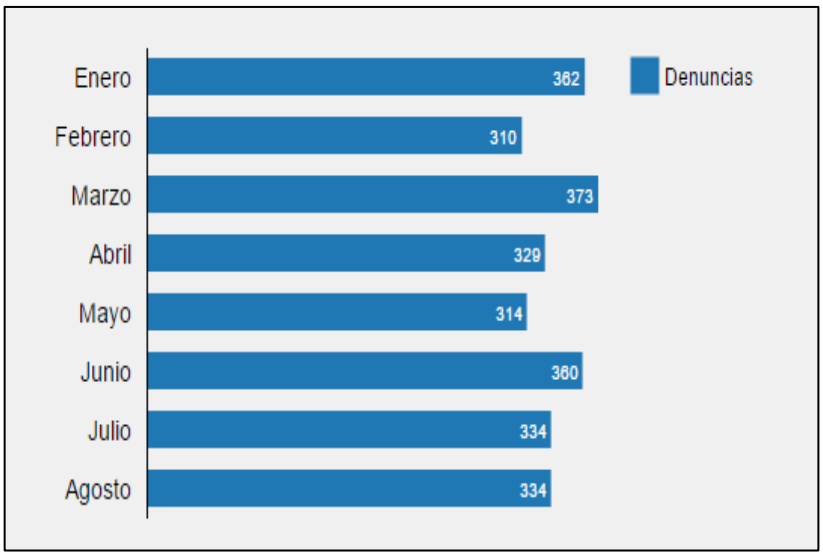


Figura 44. Reporte de delitos por mes.
Fuente: elaboración propia.

Este primer reporte muestra las denuncias registradas hasta el mes de Agosto. Desde un punto de vista general, en base a este gráfico podemos ver las denuncias registradas siempre son mayores a 300 como mínimo en cada mes.

Reporte 02: Modalidad de Denuncia Por Comisaría

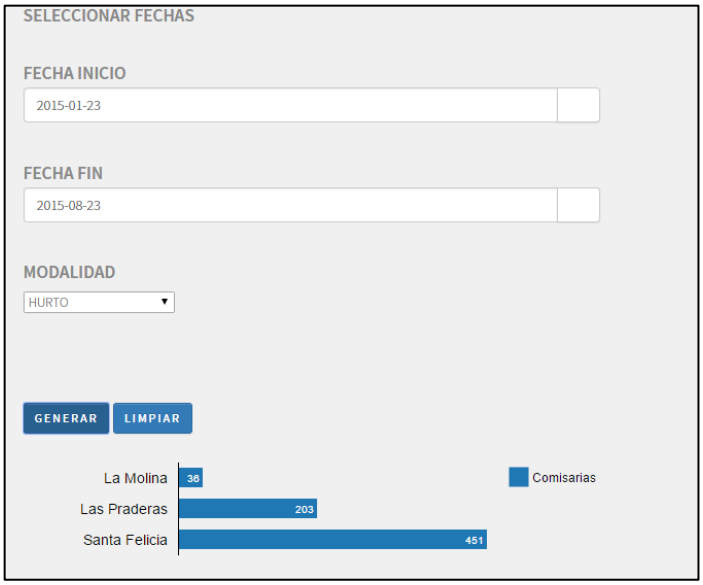


Figura 45. Modalidad de denuncia por comisaría
Fuente: elaboración propia.

A diferencia del reporte anterior, este nos muestra dos tipos de filtros:

- **Filtro 1 Fecha de Inicio:** Este filtro sirve para poner una fecha de inicio donde haya ocurrido una denuncia policial, siendo un tipo de dato Date
- **Filtro 2 Fecha Fin:** Este filtro sirve para poner una fecha de Final donde haya ocurrido una denuncia policial, siendo un tipo de dato Date
- **Filtro 3 Modalidad** Sirve para filtrar por el tipo de modalidad del hecho delictivo entre los cuales pueden ser: robo, abandono de hogar, atropello, choque, Daño o Hurto.

Como resultado se obtiene la cantidad de denuncias registradas donde se hayan registrado en determinada Comisaría.

Reporte 03: Tipo de Denuncia Por Zona

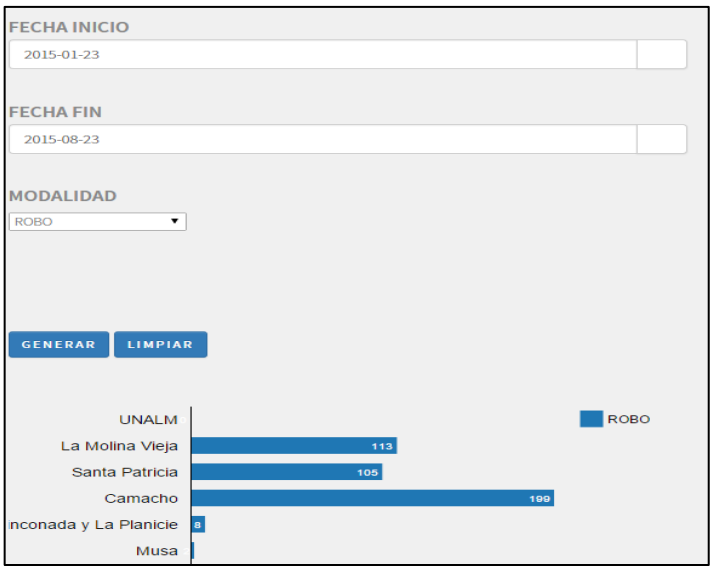


Figura 46. Reporte Tipo de Denuncia Por Zona
Fuente: elaboración propia.

- **Filtro 1: Fecha de Inicio**
  Este filtro sirve para poner una fecha de inicio donde haya ocurrido una denuncia policial, siendo un tipo de dato Date
- **Filtro 2: Fecha Fin**
  Este filtro sirve para poner una fecha de Final donde haya ocurrido una denuncia policial, siendo un tipo de dato Date

- **Filtro 3: Modalidad**

  Sirve para filtrar por el tipo de modalidad del hecho delictivo entre los cuales pueden ser: robo, abandono de hogar, atropello, choque, Daño o Hurto.

Como resultado se obtiene la cantidad de denuncias registradas donde se hayan registrados y que están categorizadas por zonas donde podemos ver que la mayor cantidad de robos se han realizado en Camacho.

Presentación final (sistema)

A continuación se muestran imágenes del sistema en funcionamiento (basad en las historias de usuario) a manera de detalle de uso del mismo.

Parte web.

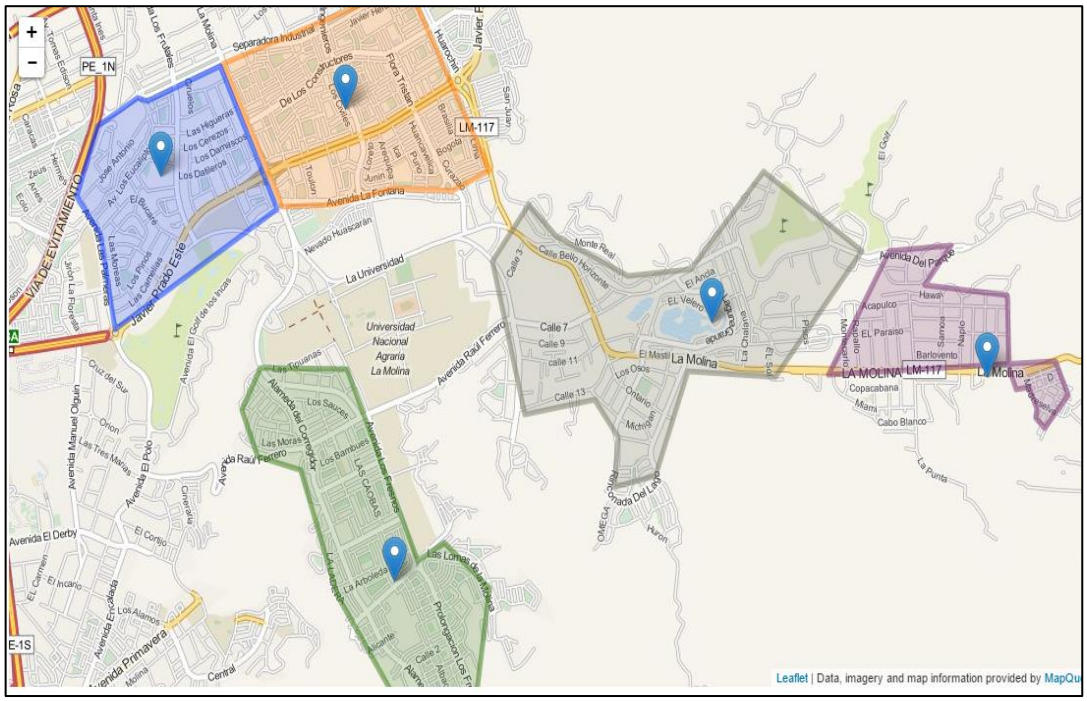


Figura 47. Mapa de La Molina zonificado.
Fuente: elaboración propia.

Parte Móvil

A continuación se muestra la parte móvil de sistema, mostrando sus dos funcionalidades principales: el mapa predictivo y los reportes.

En lo referido al mapa predictivos, al ingresar se muestra el mapa del distrito con las zonas que coinciden con el criterio de búsqueda, en este caso es “Hurto”.

Figura 48. Mapa predictivo separado por zonas.
Fuente: elaboración propia.

Al entrar al detalle de una zona se ven las predicciones tal como se muestra en la imagen siguiente.

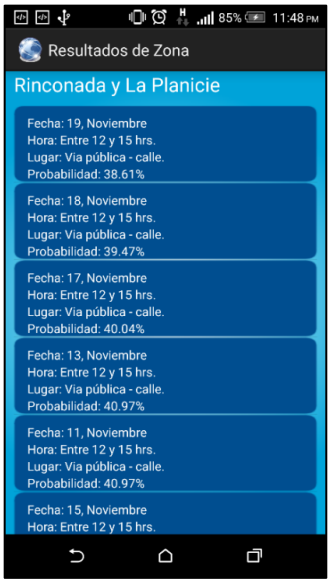


Figura 49. Detalle de la predicción según zona
Fuente: elaboración propia.

.

En lo referido a la **funcionalidad de los reportes**, se tiene 4 reportes fundamentales, a continuación se muestran las capturas de dichos reportes:

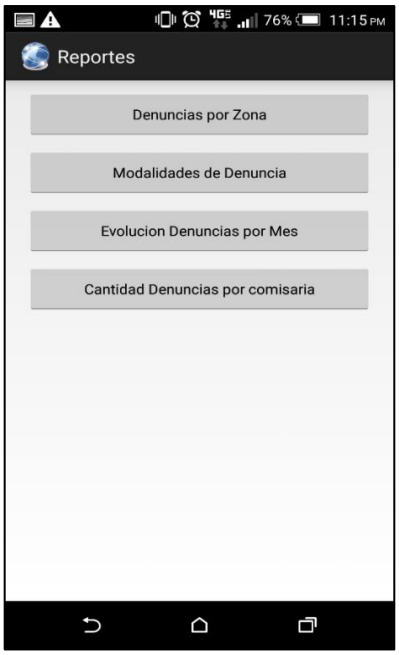


Figura 50. Listado de reportes en la parte móvil
Fuente: elaboración propia

El detalle con la información del reporte se muestra en el sistema de la siguiente manera:

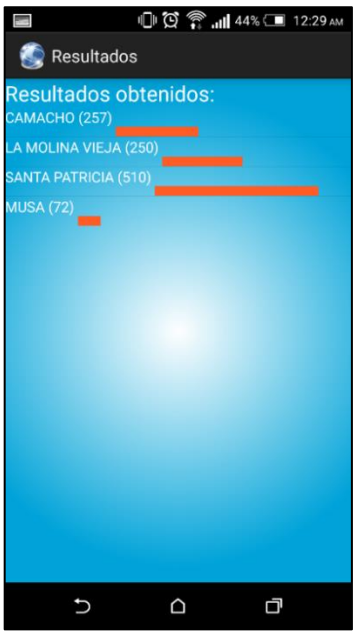


Figura 51. Detalle del reporte denuncias por zona.
Fuente: elaboración propia

CAPÍTULO IV
PRUEBAS Y RESULTADOS

4.1 Pruebas

4.1.1 Plan de pruebas

En este apartado se definirán y detallarán los pasos para realizar las pruebas, las mismas que se dividirán en dos bloques validación del modelo y pruebas del sistema, con el fin de comprobar y asegurar que nuestro sistema cumple los requerimientos y objetivos trazados al inicio.

Referido a validación del modelo, se probará y validará el modelo de predicción desarrollado, verificado con data nueva la probabilidad de éxito o error en la predicción de hechos delictivos.

Por otro lado, en lo referido a las pruebas del sistema, aquí habrán unas pruebas de aceptación conde el usuario final da el visto bueno de que el sistema cumple con lo que el negocio requería en un principio.

Adicionalmente las pruebas funcionales y de stress serán planteadas para verificar que el sistema es de calidad. El Anexo 8 muestra el documento de plan de pruebas de validación.

4.3 Pruebas de Stress

Uno de los Test a realizar es el de performance o rendimiento, comúnmente denominado como Stress Test, donde se realizan diversos análisis, entre los principales se encuentran: simulación de cantidad conexiones, tiempo de latencia que toma en cargar una página web, cantidad de datos transmitidos en base a una métrica, uso de bando de ancho entre otros.

Para ello se ha hecho uso de Load Impact que es una herramienta que nos permite realizar todo lo antes mencionado y observarlo a través de un dashboard, actualmente se ha optado por la subscripción free ya que con la información que se recopila no hace falta de la adquisición otra versión.

La Url a analizar es la siguiente, ya que ahí se encuentra alojada el SISTPRE:

www.usmpprojectsistpre-skad357.c9.io

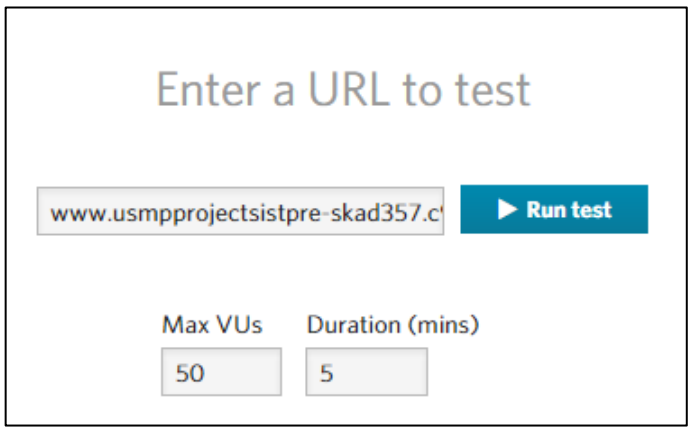


Figura 52. Inicio de las pruebas de stress
Fuente: LoadImpact, https://loadimpact.com/

Lo primero que realiza la herramienta es capturar todas las urls internas que maneja el SISTPRE, la siguiente es una muestra de algunos elementos que forman parte de la aplicación y que serán considerados en la prueba de stress.

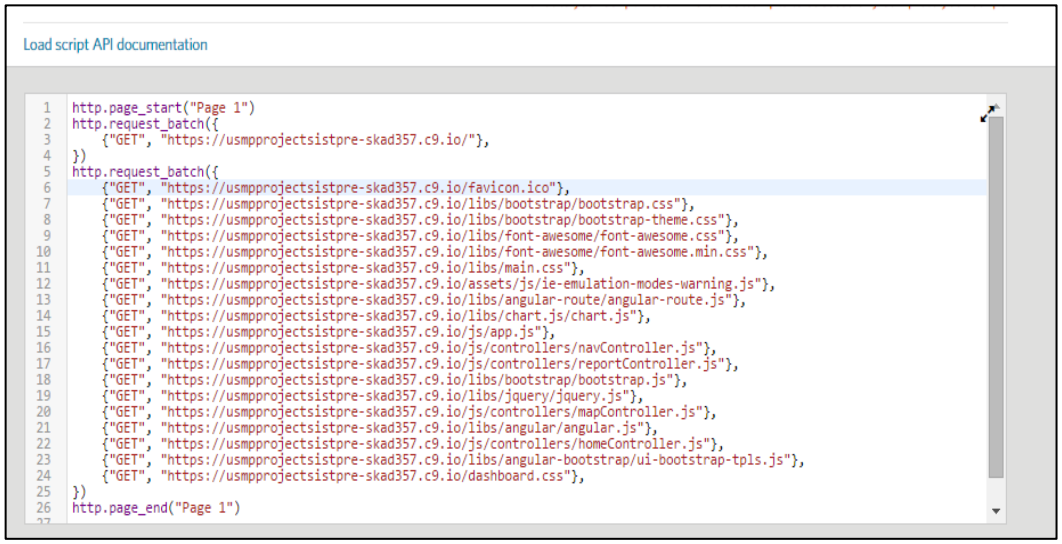


Figura 53. Archivos considerados en prueba de stress.
Fuente: elaboración propia.

En el siguiente cuadro podemos ver el detalle de los segundos y microsegundos que algunos de los componentes toman en cargar a partir de 2248 peticiones al servidor se tiene el siguiente resultado.

URL	Load zone	User scenario	Successful	Failed	Last avg
usmpprojectsistpre-skad357.c9.io/../main.css	Ashburn, US (Amazon)	Auto generated from usmpprojectsistpre-skad357.c9.io	113	0	3.76s
usmpprojectsistpre-skad357.c9.io	Ashburn, US (Amazon)	Auto generated from usmpprojectsistpre-skad357.c9.io	114	0	548.86ms
usmpprojectsistpre-skad357.c9.io/../chart.js	Ashburn, US (Amazon)	Auto generated from usmpprojectsistpre-skad357.c9.io	113	0	2.56s
usmpprojectsistpre-skad357.c9.io/../font-awesome.min.css	Ashburn, US (Amazon)	Auto generated from usmpprojectsistpre-skad357.c9.io	109	0	2.11s
usmpprojectsistpre-skad357.c9.io/../bootstrap.css	Ashburn, US (Amazon)	Auto generated from usmpprojectsistpre-skad357.c9.io	114	0	8.68s
usmpprojectsistpre-skad357.c9.io/../reportController.js	Ashburn, US (Amazon)	Auto generated from usmpprojectsistpre-skad357.c9.io	113	0	468.98ms
usmpprojectsistpre-skad357.c9.io/../app.js	Ashburn, US (Amazon)	Auto generated from usmpprojectsistpre-skad357.c9.io	113	0	3.81s
usmpprojectsistpre-skad357.c9.io/../font-awesome.css	Ashburn, US (Amazon)	Auto generated from usmpprojectsistpre-skad357.c9.io	114	0	3.45s
usmpprojectsistpre-skad357.c9.io/../homeController.js	Ashburn, US (Amazon)	Auto generated from usmpprojectsistpre-skad357.c9.io	110	0	384.91ms
usmpprojectsistpre-skad357.c9.io/../mapController.js	Ashburn, US (Amazon)	Auto generated from usmpprojectsistpre-skad357.c9.io	111	0	382.84ms
usmpprojectsistpre-skad357.c9.io/../jquery.js	Ashburn, US (Amazon)	Auto generated from usmpprojectsistpre-skad357.c9.io	113	0	13.5s
usmpprojectsistpre-skad357.c9.io/../navController.js	Ashburn, US (Amazon)	Auto generated from usmpprojectsistpre-skad357.c9.io	113	0	388.82ms
usmpprojectsistpre-skad357.c9.io/../bootstrap.js	Ashburn, US (Amazon)	Auto generated from usmpprojectsistpre-skad357.c9.io	113	0	14.47s
usmpprojectsistpre-skad357.c9.io/../angular-route.js	Ashburn, US (Amazon)	Auto generated from usmpprojectsistpre-skad357.c9.io	113	0	2.79s
usmpprojectsistpre-skad357.c9.io/../bootstrap-theme.css	Ashburn, US (Amazon)	Auto generated from usmpprojectsistpre-skad357.c9.io	114	0	2.52s
usmpprojectsistpre-skad357.c9.io/../angular.js	Ashburn, US (Amazon)	Auto generated from usmpprojectsistpre-skad357.c9.io	111	0	35.34s
usmpprojectsistpre-skad357.c9.io/../ie-emulation-modes-warning.js	Ashburn, US (Amazon)	Auto generated from usmpprojectsistpre-skad357.c9.io	0	113	1.48s
usmpprojectsistpre-skad357.c9.io/../ui-bootstrap-tpls.js	Ashburn, US (Amazon)	Auto generated from usmpprojectsistpre-skad357.c9.io	110	0	14.78s

Tabla 30. Lista de latencia de componentes del sistema
Fuente: Elaboración propia.

Como podemos ver no tenemos ningún archivo caído, y el máximo tiempo de latencia es de 14.47s, y hace referencia al archivo donde se ejecuta la carga del mapa, esta parte es un poco pesada entre comillas ya que levanta una gran cantidad información de un solo golpe, se planea mejorar este performance, sin embargo este impacto está considerado como bajo.

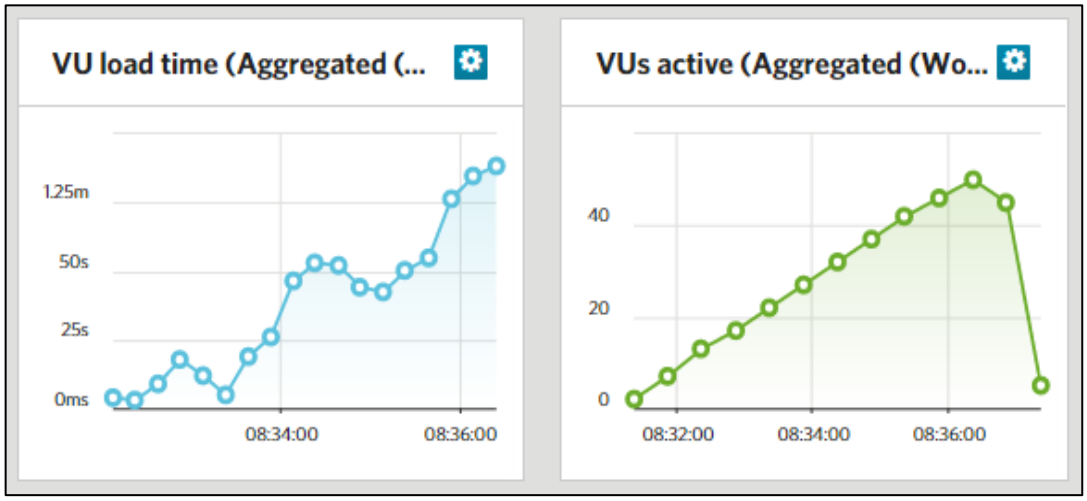


Figura 54. Tiempo de carga de la aplicación web.
Fuente: elaboración propia.

Estas dos imágenes nos muestran lo siguiente; en el primer caso se puede observar que el tiempo de carga de la página es de 1 minuto y 25 segundos, sin embargo este hace referencia a todo el flujo de la aplicación, específicamente de todo el flujo principal.

En el segundo caso se pueden observar VU (Usuarios virtuales) simulados durante el tiempo que se hacía el test; esto quiere decir que todo ocurrió con normalidad ya que si no se hubiera dado una caída en la aplicación y se mostraría puntos rojos, se han simulado solo 45 usuarios ya que es el promedio de usuarios que se piensa que estará conectado al sistema simultáneamente, además que el acceso a la aplicación está restringido y no está público para todos.

4.2 Resultados

El modelo de predicción fue elaborado en la última quincena del mes de setiembre, es por ello que no se tomó información de denuncias en dicho

mes, solo hasta agosto del 2015. Por ese motivo se empezó con la predicción de casos en primera instancia para el mes de octubre y poner el sistema a funcionar en dicho mes.

Para la predicción de hechos delictivos que pueden venir en el mes de octubre se hicieron los siguientes pasos:
Primero se determinó la moda en cada mes de la variable de “modalidad de delito”. Como ya se sabe, nuestra información recolectada tiene información de Enero a Agosto del 2015.

Para los meses de enero a agosto la moda es hurto, por ello, para obtener más modalidades a predecir, se incluirán las modalidades más registradas cada mes pero sin considerar repeticiones, es decir si una modalidad ya la consideramos para un mes entonces ya no será contada para el siguiente, siendo de la siguiente forma: en enero se tomará en cuenta el hurto; en febrero el robo; en marzo el choque; en abril el daño; en mayo el abandono de hogar; para junio las lesiones; en julio estafay otras defraudaciones y finalmente en agosto las faltas contra las personas.

Segundo, según los datos recolectados en la fase 2 del desarrollo del proyecto, se pudo conocer las horas que los efectivos poseen la logística necesaria para hacer patrullajes, con ello obtuvimos las horas.

Por último conocimos los tipos de lugares donde la policía puede ir de manera más rápida a efectuar una intervención o patrullaje, con esto determinamos la variable llamada tipo lugar.

Con todo lo mencionado responderíamos al supuesto dicho por un efectivo policial: “Mañana viernes dos de octubre, un hecho delictivo de hurto simple entre las 9pm y 12am en la vía-publica.- calle ¿dónde ocurrirá?”. El modelo arroja la zona predicha con un porcentaje de ocurrencia.

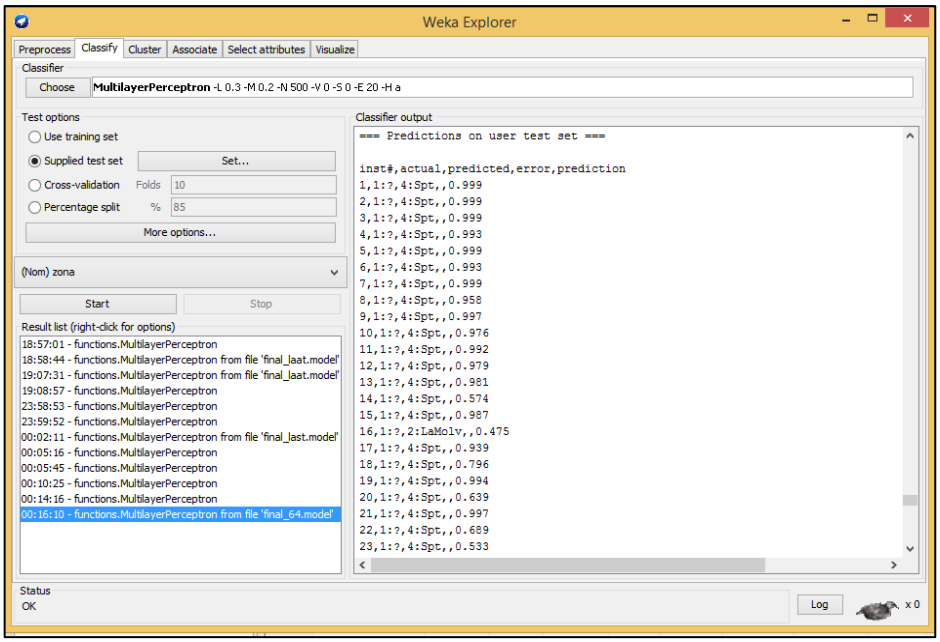


Figura 55. Resultado de la predicción.
Fuente: elaboración propia.

Los resultados totales de las predicciones se muestran a continuación.

Es importante mencionar que el modelo tiene una tasa de acierto de 72,09%, es decir de cada una de las predicciones serán ciertas con dicha probabilidad. Por ejemplo para el caso: “Robo agravado en Santa Patricia con una probabilidad de ocurrencia del 86.05%” esto quiere decir 0.8505 * 0.7209 = 61,31 % de que el hecho delictivo ocurra en esa zona con el escenario especificado.

CAPÍTULO V
DISCUSIÓN Y APLICACIÓN

5.1 Discusión

La presente investigación planteó una hipótesis general netamente involucrada a la mejora del proceso de prevención del delito. Dicha hipótesis será validada al corroborar las tres hipótesis específicas; es decir, si se logra comprobar las hipótesis específicas se podrá decir que se ha probado la hipótesis general.

El proyecto, que implementa un sistema de predicción de hechos delictivos en el distrito de La Molina, trajo consigo una mejora en dicho proceso.

Entonces, en este apartado se tratará de contrastar las hipótesis planteadas al inicio de la investigación.

Es necesario aclarar que el sistema fue implementado a inicios de octubre, es por ello que a partir del 20 de octubre del 2015 se recogerán las muestras para los datos después del sistema.

Con respecto a la muestra considerada como antes del sistema, sí se tiene la información de cómo marchaba la situación antes del sistema.

Entonces, el mes de octubre es el que se describe como: “luego de implementado el sistema”, cualquier otro mes indica que es “antes de la implementación del sistema”. El detalle de la trazabilidad de los objetivos con las hipótesis del proyecto a medir se puede encontrar en el ANEXO 9.

Para medir las hipótesis relacionadas a la mejora o disminución de algún factor luego de la implementación del sistema podemos utilizar dos estadísticos:

- **Prueba T - Student para datos relacionados o apareados**: se utiliza cuando las muestras (antes y después de algún factor) presentan una distribución normal.
- **Prueba de Wilconxon:** Es la variante d la prueba de T - Student cuando no se puede suponer la normalidad de las muestras.

Para realizar estas tareas se utilizará la herramienta SPSS de IBM y por último se debe mencionar que la implementación tuvo como piloto la comisaria de santa Felicia, que es donde se recogerán los resultados (Como toda implementación de sistemas siempre se debe hacer por divisiones o áreas, en este caso lo haremos por divisiones que son las comisarias).

Hipótesis específica 1: La implementación de una aplicación móvil que muestre un mapa demarcado por zonas donde se pueda apreciar los posibles delitos a ocurrir, podrá mejorar los tiempos tiempo de acción de los policías al poder identificar zonas de riesgo de una manera más rápida.
Aquí las variables a medir son las siguientes:

- T1 = Tiempo que se tomaba en identificar una zona de riesgo
- T2 = Tiempo que se toma actualmente en identificar una zona de riesgo.

La obtención de los datos para medir este objetivo se ha realizado mediante encuestas hechas a 9 efectivos policiales pertenecientes a la comisaria de Santa Felicia.

Estas encuestas han sido realizadas a los mismos efectivos policiales en dos escenarios distintos (T1 y T2 descritas anteriormente). La pregunta de la encuesta que nos interesa para este objetivo es: ¿Cuánto tiempo invierte Ud. en identificar una zona de riesgo?

Estas dos premisas nos dan el primer acercamiento de que debemos utilizar una comprobación estadística de T-Student con muestras relacionadas debido a las siguientes razones:

- No se puede determinar a priori si los tiempos de acción han sido reducidos o no y si estos representan una reducción significativa.
- Se evalúa una muestra en base a dos escenarios (antes y después de la ocurrencia de un hecho, en este caso el sistema).
- Nuestras variables a medir son cuantitativas.

La tabla que se muestra continuación muestra el resultado a la pregunta tiempo empleado antes del sistema. (Se conoce como tiempo de acción al tiempo que tarda un efectivo policial desde que empieza a identificar una zona de riesgo hasta que lo consigue y toma una decisión).

Tabla 31. Tiempo en identificar una zona de riesgo AS

EFECTIVO POLICIAL	TIEMPO ANTES DEL SISTEMA	EFECTIVO POLICIAL	TIEMPO ANTES DEL SISTEMA
1	4 min.	11	4 min.
2	5 min.	12	7 min.
3	5 min.	13	4 min.
4	6 min.	14	9 min.
5	8 min.	15	8 min.
6	7 min.	16	7 min.
7	10 min.	17	12 min.
8	9 min.	18	9 min.
9	30 min.	19	10 min.
10	7 min.	20	4 min.

Fuente: elaboración propia

Ahora es necesario tener la información de los mismo policías encuestados anteriormente, pero ahora recogiendo el tiempo que demoran después de implementado el sistema.

La tabla que se muestra a continuación muestra los tiempos empleados por los mismos policías en identificar zonas de riesgo después de la implementación del sistema.

Tabla 32. Tiempo en identificar zona de riesgo DS

EFECTIVO POLICIAL	TIEMPO DESPUES DEL SISTEMA	EFECTIVO POLICIAL	TIEMPO DESPUES DEL SISTEMA
1	3 min.	11	6 min.
2	3 min.	12	3 min.
3	3 min.	13	4 min.
4	5 min	14	5 min
5	4 min	15	4 min
6	3 min	16	7 min.
7	3 min.	17	3 min.
8	3 min.	18	5 min.
9	3 min.	19	4 min.
10	4 min.	20	4 min.

Fuente: elaboración propia.

Para poder usar la distribución de T-Student con muestras relacionadas debemos en primer lugar comprobar que las muestras obtenidas tienen una distribución normal. SI no se logra comprobar que se tiene una distribución normal, entonces utilizaremos la prueba estadística de Wilcoxon.

Entonces los pasos para comprobar nuestro objetivo son los siguientes:

- **Paso 1: Redactar la Hipótesis Nula (H_0) y la esperada (H_1)**

  H_0: La implementación de una aplicación móvil que muestre un mapa demarcado por zonas donde se pueda apreciar los posibles delitos a ocurrir, **NO** mejora los tiempos tiempo de acción de los policías al poder identificar zonas de riesgo de una manera más rápida.

  H_1: La implementación de una aplicación móvil que muestre un mapa demarcado por zonas donde se pueda apreciar los posibles delitos a ocurrir, mejora los tiempos tiempo de acción de los policías al poder identificar zonas de riesgo de una manera más rápida.

- **Paso 2: Definir α**

  Siendo **α** el margen de error aceptado. Para nuestro caso es 0.05 o 5%.

- **Paso 3: Verificación de la Normalidad de los variables**

  Para analizar la normalidad de las variables tenemos que hacerlo a través de su nivel de significancia, para nuestro caso, donde nuestra

muestra es menor a 30 individuos se usa la prueba de Normalidad de Shapiro-Wilk.

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
AS	.278	20	.000	.621	20	.000
DS	.246	20	.002	.802	20	.001

a. Lilliefors Significance Correction

Figura 56. Prueba de normalidad en tiempos de acción
Fuente: Herramienta SPSS

Dónde:

- ✓ Nivel de significancia de T1 = 0.00
- ✓ Nivel de significancia de T2 = 0.001

Los niveles de significancia deben ser mayores a alfa (0.05), y en ambos casos los niveles son menores a 0.05, esto quiere decir que las variables no tienen una distribución normal, por lo que sebe usar la prueba de Wilcoxon para la muestra.

- **Paso 4: Aplicar prueba de Wilcoxon**

Para comprobar que si existe una diferencia significativa entre los tiempos de acción de los efectivos policiales antes de la implementación del sistema y los tiempos de acción luego de implementado el sistema, se debe aplicar Wilcoxon:

Wilcoxon Signed Ranks Test

Ranks

		N	Mean Rank	Sum of Ranks
DS - AS	Negative Ranks	16[a]	9.31	149.00
	Positive Ranks	1[b]	4.00	4.00
	Ties	3[c]		
	Total	20		

a. DS < AS
b. DS > AS
c. DS = AS

Test Statistics[a]

	DS - AS
Z	-3.452[b]
Asymp. Sig. (2-tailed)	.001

Figura 57. Prueba Wilcoxon para tiempos de acción
Fuente: Herramienta SPSS

Analizando esta tabla, podemos ver que nuestro nivel de significancia es de 0.001.

Para este tipo de análisis, los criterios de aceptación o rechazo son:

- Si el nivel de significancia obtenido es mayor a α**,** entonces se acepta H0.
- Si el nivel de significancia obtenido es menor a α entonces se acepta H1.

Nuestro nivel de significancia 0.001 < 0.05 lo que quiere decir que se acepta H1.

Entonces se puede decir que La implementación de una aplicación móvil que muestre un mapa demarcado por zonas donde se pueda apreciar los posibles delitos a ocurrir, mejora los tiempos tiempo de acción de los policías al poder identificar zonas de riesgo de una manera más rápida.

Hipótesis específica 2: La implementación de una herramienta de predicción de hechos delictivos en el distrito de La Molina logrará mejorar el servicio policial en base a la realización de acciones preventivas por zonas.

Esta hipótesis es medida analizando la variable: Acciones preventivas por zonas. Considerando que las acciones preventivas comprenden los patrullajes, reglajes, operativos y redadas.

De la misma manera se debe comparar dos situaciones: Cómo estaba antes y cómo está después del sistema implementado y así obtener el porcentaje de mejora. Estas acciones preventivas comprenden rondas, batidas, organización de operativos, etc.

La información que servirá como “antes del sistema” fue recolectada fue en base a una encuesta realizada a los efectivos policiales, donde se recogió la respuesta a: Semanalmente, ¿En cuántas acciones preventivas participa? Arrojando como resultado la siguiente tabla:

Tabla 33. Acciones preventivas por policía.

Nro. Policía	Nro. Acciones preventivas	Tipos	Nro. Policía	Nro., Acciones preventivas	TIPOS
1	18	Patrullaje	11	16	Intervenciones
2	20	Patrullaje	12	15	Intervenciones
3	15	Intervenciones	13	17	Patrullaje
4	15	Patrullaje	14	20	Intervenciones
5	20	Intervenciones	15	17	Patrullaje
6	17	Patrullaje	16	15	Patrullaje
7	15	Patrullaje	17	19	Patrullaje
8	20	Patrullaje	18	14	Patrullaje
9	16	Patrullaje	19	16	Patrullaje
10	18	Patrullaje	20	16	Patrullaje

Fuente: elaboración propia

De la misma manera, se planteó la misma pregunta a los mismos efectivos policiales luego de 2 semanas de implementado el sistema. El resultado de dicha encuesta se muestra a continuación:

Tabla 34. Acciones preventivas por policía

EFECTIVO POLICIAL	NRO. DE ACCIONES PREVENTIVAS	TIPOS DE ACCIONES PREVENTIVAS
1	22	Patrullaje
2	19	Patrullaje
3	18	Intervenciones
4	17	Patrullaje
5	21	Intervenciones
6	20	Patrullaje
7	17	Patrullaje
8	18	Patrullaje
9	20	Patrullaje
10	16	Patrullaje
11	20	Patrullaje
12	14	Intervenciones
13	15	Patrullaje
14	19	Intervenciones

15	19	Patrullaje
16	17	Patrullaje
17	18	Patrullaje
18	20	Patrullaje
19	21	Patrullaje
20	18	Patrullaje

Fuente: elaboración propia.

Ahora con la ayuda del SPSS haremos el análisis estadístico para comprobar si existe una mejora.

Para poder usar la distribución de T-Student con muestras relacionadas debemos en primer lugar comprobar que las muestras obtenidas tienen una distribución normal. SI no se logra comprobar que se tiene una distribución normal, entonces utilizaremos la prueba estadística de Wilcoxon.

Entonces los pasos para comprobar nuestro objetivo son los siguientes:

- **Paso 1: Redactar la Hipótesis Nula (H_0) y la esperada (H_1)**

  H_0: La implementación de una herramienta de predicción de hechos delictivos en el distrito de La Molina no logra mejorar el servicio policial en base a los tiempos empleados en la realización de acciones preventivas y asignación de recursos por zonas.

  H_1: La implementación de una herramienta de predicción de hechos delictivos en el distrito de La Molina logra mejorar el servicio policial en base a los tiempos empleados en la realización de acciones preventivas y asignación de recursos por zonas.

- **Paso 2: Definir α**

Siendo **α** el margen de error aceptado. Para nuestro caso es 0.05 o 5%.

- **Paso 3: Verificación de la Normalidad de los variables**

Para analizar la normalidad de las variables se tiene que hacerlo a través de su **nivel de significancia**, para nuestro caso, donde nuestra muestra es menor a 30 individuos se usa la prueba de Normalidad de Shapiro-Wilk.

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
AS	.184	20	.075	.892	20	.029
DS	.124	20	.200*	.969	20	.739

Figura 58. Test de normalidad - acciones preventivas
Fuente: Herramienta SPSS

Donde:

- ✓ Nivel de significancia de A1 = 0.029
- ✓ Nivel de significancia de T2 = 0.739

Los niveles de significancia deben ser mayores a nuestro alfa (0.05) para que sea una distribución normal. En el primer caso el nivel es menor a 0.05, esto quiere decir que las variables no tienen una distribución normal, por lo que sebe usar la prueba de Wilcoxon para nuestra muestra.

- **Paso 4: Aplicar prueba de Wilcoxon**

Para comprobar que si existe una diferencia significativa entre los tiempos de acción de los efectivos policiales antes de la implementación del sistema y los tiempos de acción luego de implementado el sistema, debemos aplicar T – Student. La aplicación de la prueba estadística mediante la herramienta IBM SPSS se muestra en la siguiente figura:

Wilcoxon Signed Ranks Test

Ranks

		N	Mean Rank	Sum of Ranks
TD - TA	Negative Ranks	7[a]	5.79	40.50
	Positive Ranks	13[b]	13.04	169.50
	Ties	0[c]		
	Total	20		

a. TD < TA
b. TD > TA
c. TD = TA

Test Statistics[a]

	TD - TA
Z	-2.431[b]
Asymp. Sig. (2-tailed)	.015

a. Wilcoxon Signed Ranks Test
b. Based on negative ranks.

Figura 59. Prueba Wilcoxon para acciones preventivas
Fuente: Herramienta SPSS

Analizando esta tabla, podemos ver que nuestro nivel de significancia es de 0.015.

Para este tipo de análisis, los criterios de aceptación o rechazo son:

- Si el nivel de significancia obtenido es mayor a α, entonces se acepta H0.
- Si el nivel de significancia obtenido es menor a α entonces se acepta H1.

Nuestro nivel de significancia 0.015 > 0.05 lo que quiere decir que se acepta H1. Entonces se puede decir con que la implementación de una herramienta de predicción de hechos delictivos en el distrito de La Molina mejora el servicio policial en base a la realización de acciones preventivas por zonas.

Esta conclusión se llega luego de hacer un aumento en la muestra, pues se tiene más precisión.

Hipótesis específica 3: La unificación y estructuración de la información de las distintas comisarías de la Molina permite tener una mejor organización de la información de las denuncias del distrito.

La medición del cumplimiento de este objetivo está relacionada directamente con el diseño e implementación de la base de datos, donde se pueda tener la información consolidada de las tres comisarías (Santa Felicia, Las Praderas, La Molina).

En la base de datos construida, se cargaron las denuncias hasta agosto del 2015. No es necesario detallar más en esto debido a que en el capítulo 3 se tiene el detalle del modelado de la base de datos y de los procesos ETLs construidos para poder tener la información automatizada.

Adicionalmente, se realizó una encuesta de percepción que fue aplicada a los usuarios finales, la cual podemos utilizar como medida adicional para comprar si el proceso de prevención del delito tuvo una mejora.

Hipótesis general: Mejorar el rendimiento del proceso de prevención del delito de las comisarías de La Molina utilizando minería de datos.

Este objetivo será medido mediante encuestas de percepción a los partícipes del proceso.

Se hará un análisis según la encuesta de percepción realizada, para ver si el proceso ha mejorado. Para poder analizar los resultados, esta encuesta se realizará mediante la escala de Likert. El día miércoles 4 de noviembre se hizo una nueva visita a la comisaria piloto (Santa Felicia).

La tabla que se muestra a continuación detalla las respuestas de 7 efectivos policiales (mostrados con sus nombres) en la comisaría de Santa Felicia. El detalle de esta encuesta y de otras encuestas adicionales realizadas se puede observar en el ANEXO 10.

Los enunciados fueron:

a. Con el sistema se pudo realizar mejores controles preventivos
b. El sistema me ayudó a identificar zonas de riesgo más rápido que con el método habitual
c. El sistema implantado me ayuda a prevenir delitos con dolo en la jurisdicción tales como robo, hurto, lesiones, daños, etc.
d. El sistema implementado sirve como herramienta de ayuda para tener una estadística adicional que colabora con la actual que posee la comisaria.

Las alternativas de respuestas fueron:

1: Totalmente de acuerdo. 2: De acuerdo. 3: Ni de acuerdo ni en desacuerdo 4: En desacuerdo. 5: Totalmente en desacuerdo

Tabla 35. Resultado de encuesta de percepción.

ENCUESTADO	PREG. 1	PREG. 2	PREG. 3	PREG. 4
SO.S PNP Sotelo Olivera	3	4	1	2
S.O.T2 PNP Caldas Guerra	4	4	4	3
SO.B PNP Sánchez Silva	5	5	5	2
SO 3 PNP Aguirre Pajuelo Edyson	2	3	4	3
SO.S Alarcón Quispe Adin	3	4	3	4
S.O.S PNP Rondinel Peralta	4	4	4	3
SO.B Soto Santiago Juan	2	5	4	5

Fuente: elaboración propia.

Haciendo el análisis de estadística descriptiva, por ser una encuesta en escala de Likert, se tiene que para cada pregunta las frecuencias acumuladas son:

Tabla 36. Frecuencia acumulada de los resultados.

	PREG. 1	PREG. 2	PREG. 3	PREG. 4
Totalmente de acuerdo	1	2	1	1
De acuerdo	2	4	4	1
Ni de acuerdo ni en desacuerdo	2	1	1	3
En desacuerdo	2	0	0	2
Totalmente en desacuerdo	0	0	1	0

Fuente: elaboración propia.

Tabla 37. Resultados de la encuesta de percepción.

	FA	PORCENTAJE
Totalmente de acuerdo	5	17.86%
De acuerdo	11	39.29%
Ni de acuerdo ni en desacuerdo	7	25.00%
En desacuerdo	4	14.29%
Totalmente en desacuerdo	1	3.57%
TOTAL	28	100.00%

Fuente: elaboración propia.

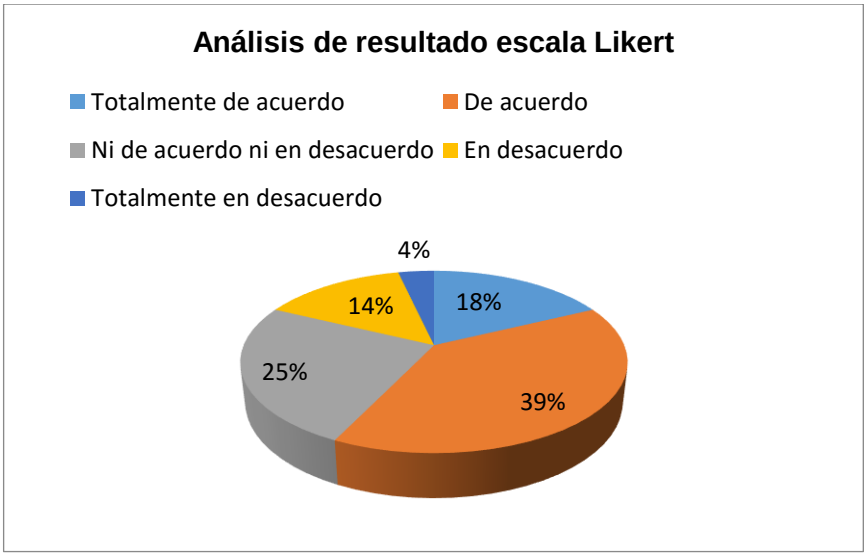


Figura 60. Muestra de resultados de la encuesta.
Fuente: Elaboración propia.

Si bien en el 39% de los casos, las respuestas fueron “Totalmente de acuerdo”, no se puede decir con exactitud que el proceso ha mejorado pues no hay mayoría, pero si podemos decir que el sistema no es indiferente, es decir que trae un beneficio a la comisaria, el cual se podría traducir en mejora del proceso en un futuro. Entonces, este mismo ejercicio se debería volver a realizar pero cuando el sistema tenga más tiempo de ejecución para sacar una conclusión más exacta sobre una mejora.

5.1.1 Cambio en la organización

Gracias a la implementación del sistema de predicción, los efectivos vieron automatizadas las siguientes tareas manuales:

- Identificación de zonas de riesgo: el proceso de identificar zonas de riesgo anteriormente abarcaba la tarea manual de ubicar en su mapa físico las zonas y colocar ahí las que tienen la mayor cantidad de denuncias de delitos. Esta actividad perteneciente al proceso de prevención del delito se realizaba cada vez que se necesitaba asignar recursos para realizar patrullajes o intervenciones.

Con el sistema esto se reduce a cargar el mapa y hacer el filtrado por delitos.

- De la misma manera, la cantidad de acciones preventivas realizadas no era óptima y eran pocas, debido a que no se tenían claras las zonas donde

realizar estas acciones y se logre evitar posibles delitos. El sistema ayuda a tener las zonas demarcadas con la probabilidad de ocurrencias de delitos para que así se realicen las rondas o patrullajes necesarias que ayuden a prevenir posibles hechos delictivos.

Estos cambios de tareas manuales a tareas automatizadas se pudieron lograr mediante la capacitación a los efectivos policiales y a la entrega de un manual de usuario que se aprecia en el Anexo 11.

5.2 Aplicación

En este apartado se detallará otras aplicaciones o mejoras que se podrían agregar a nuestro sistema para que amplíe su alcance y así abarcar más usuarios o pueda dar más precisión en las predicciones.

- Nuestra propuesta puede dar información no solo a nivel operativo que es como se plantea, sino también hacia mandos superiores. Donde ellos puedan tener información sobre delitos que posiblemente pueden venir y según esto modificar su logística para asignar más recursos en jurisdicciones más vulnerables.

- **Localizar hechos delictivo en el lugar preciso:**
  Esta aplicación está referida básicamente a la posibilidad de utilizar los dispositivos GPS que poseen los sistemas para conocer la ubicación exacta de los efectivos, y puedan ser guiados o dirigidos hacia zonas (calles, avenidas) donde se realizan patrullajes.

- **Control de efectivos policiales.**
  Aparte de ser un sistema de predicción, si el dispositivo donde está la aplicación tiene GSP, entonces se puede tener una funcionalidad adicional que sería el rastreo del dispositivo. Donde se tendría ubicaciones exactas para saber si los efectivos policiales están realizando sus tareas asignadas. Es decir podría servir además servir como un sistema de control

- **Registrar denuncias.**

  La solución presenta una sección de reportes, donde se muestra la evolución de denuncias según diversos criterios de búsqueda. Una aplicación adicional relacionada con esta funcionalidad es el de registro de denuncias; con esto, nuestro sistema abarcaría un poco más la parte transaccional de las incidencias diarias que llegan a las comisarías.

CONCLUSIONES

Primera: Se logró una relativa mejora del proceso de prevención de delitos que se realizan en la comisaría de Santa Felicia, comprobándose que los efectivos policiales ven la simplificación de sus actividades, esto gracias a la implementación del sistema de predicción. Lo que ayuda a los efectivos policiales a tomar mejores decisiones con respecto a las zonas donde podría ocurrir un hecho delictivo con más probabilidad.

Segunda: Los tiempos de acción de los efectivos policiales en detectar zonas de riesgos se vieron reducidos de una manera considerable, debido a que sus tareas que normalmente son manuales, serian reemplazadas periódicamente a la utilización del sistema, donde sólo se necesite ejecutar y analizar los resultados.

Tercera: La eficiencia del servicio policial se vio mejorada, debido a que las acciones preventivas aumentaron, esto debido a que el sistema da la iniciativa de hacia dónde se debe centrar la prevención del delito y mejorando su asignación de personal.

Cuarta: La información de las tres comisarías del distrito describe todos los hechos delictivos de la zona, por ello se logró que esta esté unificada y consolidada para poder agilizar el proceso de prevención de delitos, aparte de tener una mayor disponibilidad del acceso a la información.

RECOMENDACIONES

Primera: El proceso de prevención del delito en el distrito de La Molina podría tener controles y métricas relacionadas a la mejora que se obtiene según nuevas medidas adoptadas, en este caso, el sistema de predicción.

Segunda: El sistema podría permitir la posibilidad de establecer denuncias en la misma aplicación, debido a que en ocasiones los efectivos policiales pueden estar en el momento exacto de la ocurrencia del hecho delictivo.

Tercera: El sistema debería conectarse a la base de datos de la RENIEC y sistema de antecedentes penales para poder obtener la información de las personas, y en ciertos hechos delictivos, poder obtener los antecedentes policiales de los denunciados y establecer así un estado crítico de la denuncia si fuera el caso.

Cuarta: Se podría agregar más data histórica al modelo proveniente de otras fuentes de información tales como sistemas judiciales, encuestas de percepción pública, etc. para obtener una mayor certeza en las predicciones.

FUENTES DE INFORMACIÓN

Agrawall, R., & Shim, K. (3 de Setiembre de 2015). *Citeseerx.* Obtenido de Develping Tightly - Coupled Data Mining Applications on a Relational Database System: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.88.8047&rep=rep1&type=pdf

Basombrío, I. C. (2012). *Lima y otras Ciudades del Perú comparadas con América Latina.* Lima: Bellido Ediciones.

Cadena, L. L. (2011). *Aplicando minería de datos al marketing educativo.* Bogota: Universidad Sergio Arboleda.

Caridad, J. (2014). *La Mineria de Datos: Analisis de Bases de Datos en la Empresa.* Mexico: Ocerin.

Concytec. (2003). *Perú ante la sociedad del conocimiento.Indicadores de ciencia, tecnología e innovación.* Lima: Concytec.

El Comercio. (22 de Abril de 2015). Perú tiene la más alta tasa de delincuencia en Latinoamérica. *Perú tiene la más alta tasa de delincuencia en Latinoamérica*.

Engels, R., & Theusinger, C. (4 de Septiembre de 2015). *citeseerx.* Obtenido de Using RNA algorithms for voicing advise in Data Mining Aplications: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.56.7414&rep=rep1&type=pdf

Fatih, T., & Bekir, C. (2015). *Police Use of Technology To Fight Against Crime.* Ankara: European Scientific Journal.

Fayad, U., & Uthurusamy, R. (27 de Agosto de 2015). *Evolving Data Mining into Solutions for Insights.* Obtenido de Evolving Data Mining into Solutions for Insights: http://sceweb.uhcl.edu/boetticher/ML_DataMining/p28-fayyad.pdf

FayerWayer. (10 de 03 de 2012). *Fayer Wayer*. Obtenido de https://www.fayerwayer.com/2012/03/mexico-lo-que-debes-saber-sobre-la-ley-de-geolocalizacion/

Fernandez, L., & Rivas, P. (25 de 11 de 2015). *EL COMERCIO.* Obtenido de http://elcomercio.pe/lima/ciudad/nueve-cada-diez-limenos-no-se-siente-seguro-lima-noticia-1820680

Gibbs, M. (2014). *Predicting crime with big data.* Londres.

Hall, M. (01 de Setiembre de 2015). *SIGKDD.* Obtenido de The Wka Data Mining Software: An Update: http://www.sigkdd.org/sites/default/files/issues/11-1-2009-07/p2V11n1.pdf

Henderson, M. T., Wolfers, J., & Zitzewitz, E. (2008). *Predicting Crime.* Chicago.

Hernández Sampieri, R., Fernández Collado, C., & Baptista Lucio, P. (1997). *Metodología de la Investigación.* Atlacomulco: McGraw - Hill Interamericana de Mexico.

INEI. (2015). *Informe Técnico de Seguridad ciudadana.* Lima.

Instituto Nacional de Estadística e Informática. (2014). *Gestion de la Seguridad Ciudadana.* Lima.

Instituto Nacional de Estadística e Informática. (26 de Agosto de 2015). *INEI.* Obtenido de INEI: http://www.inei.gob.pe/media/MenuRecursivo/boletines/boletin-seguridad_web.pdf

Lastre Valdes, M., & García Vidal, G. (2014). *Redes neuronales artificiales en la predicción de insolvencia.* Ecuador.

Martínez, V. M. (2009). *Caracterización y predicción espacio-temporal de patrones delictivos mediante modelos lógico-combinatorios.* México D.F.

Microsoft Corporation. (1 de Setiembre de 2015). *Microsoft*. Obtenido de Conceptos de Minería de Datos: http://msdn.microsoft.com/es-es/library/ms174949.aspx

Ministerio del Interior - Perú. (28 de Agosto de 2015). *Presidencia del Consejo de Ministros.* Obtenido de Plan Nacional de Seguridad Ciudadana 2013-2018: http://www.pcm.gob.pe/seguridadciudadana/wp-content/uploads/2013/05/Plan.Nacional.Seguridad.Ciudadana.2013-2018.pdf

Mitchell, T. (02 de Setiembre de 2015). *Villanova University.* Obtenido de Machine Learning and Data Mining : http://what.csc.villanova.edu/~cassel/2500/S2007/p30-mitchell.pdf

Murazzo Carrillo, F. (2014). *Reflexiones sobre la seguridad ciudadana en el Perú.* Lima: Ediciones Nova Print S.A.C.

Norvig, P., & Russell, S. (2004). Inteligencia Artificial. Un enfoqu moderno. En S. Russell, & P. Norvig, *Inteligencia Artificial. Un enfoqu moderno.* (pág. 500). Madrid: Prentice-Hall.

Orlich, J. M. (30 de Setiembre de 2014). *Universidad para la cooperación internacional*. Obtenido de http://www.uci.ac.cr/descargas/AE/FODA(SWOT).pdf

Palacio, J., & Ruata, C. (2014). Scrum Manager Gestión de Proyectos. En J. Palacio, & C. Ruata, *Scrum Manager Gestión de Proyectos* (pág. 98). Barcelona.

Pérez López, C. (2007). Minería de datos: técnicas y herramientas. En C. Pérez López, *Minería de datos: técnicas y herramientas* (pág. 503). Madrid: Thomson ediciones.

Perversi, I. (2007). *Aplicación de minería de datos para la exploración y detección de patrones delictivos en Argentina .* Buenos Aires.

PMBOK Fifth Edition. (2014). *Purpose Of PMBOK Guide.*

PMI. (2015). *PMBOK® Guide.*

Salman, B. A. (1 de Setiembre de 2015). *Science Direct.* Obtenido de Application of Data Mining : Diabetes health care in young and old patients: http://www.sciencedirect.com/science/article/pii/S1319157812000390

SAS - España. (2006). *Minería de Datos en el Sector Seguros.* Madrid.

Valera, F. (2012). *Análisis delictual: técnicas y metodologías para la reducción del delito .* Santiago de Chile: Fundacion Paz Ciudadana.

Witten, H., Frank, E., & Hall, M. (2011). *Data Mining: Practical Machine Learning Tools and Techniques.* Miami: Morgan Kaufmann.

Woods, R. (12 de Noviembre de 2015). *ViewPoints and Perspectives.* Obtenido de Software Systems Architecture: http://www.viewpoints-and-perspectives.info/home/perspectives/security/

ANEXOS

ANEXO 1

CRONOGRAMA DEL PROYECTO

El desarrollo del Proyecto se rige en base a la Metodología del PMBOK (5TA Edición), por ende establecemos las mismas fases que esta comprende (Inicio, Organización y Preparación, Ejecución y Cierre). Que engloban el conjunto de actividades realizadas.

CRONOGRAMA DE ACTIVIDADES			
	Duración	**Comienzo**	**Fin**
SISTPRE (SISTEMA DE PREDICCION DE HECHOS DELICTIVOS)	**118 días**	**vie 08/08/15**	**vie 04/12/15**
Fase I : Inicio del Proyecto	**7 días**	**mie 12/08/15**	**mie 19/08/15**
Realización del Acta de Constitución de Proyecto (Anexo 2)	5 días	mie 12/08/15	dom 16/08/15
Definición de la Situación Problemática	2 días	jue 13/08/15	vie 14/08/15
Determinación de los Problemas	2 días	vie 14/08/15	sab 15/08/15
Determinación de los Objetivos	4 días	sab 15/08/15	mar 18/08/15
Fase II : Organización y Preparación	**69 días**	**vie 21/08/15**	**mié 25/11/15**
Realización del Plan del Proyecto	2 días	sáb 21/11/15	lun 23/11/15
Elección de las Gestiones de Conocimiento adaptadas al proyecto	2 días	sáb 21/11/15	lun 23/11/15
Realización de	**68 días**	**vie 21/08/15**	**mar 24/11/15**

Entregables de cada Gestión			
Plan de Gestión de Integración	2 días	sáb 21/11/15	lun 23/11/15
Plan de Gestión de Riesgos	2 días	sáb 21/11/15	lun 23/11/15
Plan de Gestión de Interesados	2 días	sáb 21/11/15	lun 23/11/15
Fase III : Ejecución del Trabajo	**45 días**	**vie 21/08/15**	**jue 22/10/15**
Metodología de Minería de Datos	**45 días**	**vie 21/08/15**	**jue 22/10/15**
Fase 1 : Comprensión del Negocio	**11 días**	**vie 21/08/15**	**vie 04/09/15**
Determinar los objetivos de Negocio	**3 días**	**vie 21/08/15**	**mar 25/08/15**
Trasfondo	1 día	vie 21/08/15	vie 21/08/15
Objetivos de Negocio	3 días	sáb 22/08/15	mar 25/08/15
Criterio de Éxito de Negocio	3 días	sáb 22/08/15	mar 25/08/15
Evaluar Situación	**5 días**	**vie 21/08/15**	**jue 27/08/15**
Inventario de Recursos	**4 días**	**vie 21/08/15**	**mié 26/08/15**
Terminología	3 días	**vie 21/08/15**	mar 25/08/15
Costos y Beneficios	3 días	**vie 21/08/15**	mar 25/08/15
Determinar los objetivos de Minería de Datos	**7 días**	**mié 26/08/15**	**jue 03/09/15**
Objetivos de Minería de Datos	2 días	**jue 27/08/15**	vie 28/08/15
Criterios de Éxito de Minería de Datos	4 días	**mié 26/08/15**	lun 31/08/15
Producción de Plan de Proyecto	**9 días**	**vie 21/08/15**	**mié 02/09/15**
Plan de Proyecto	3 días	**vie 21/08/15**	mar 25/08/15
Evaluación Inicial de	4 días	**vie 21/08/15**	mié 26/08/15

Herramientas y Técnicas			
Fase 2 : Comprensión de los Datos	**16 días**	**mar 01/09/15**	**mar 22/09/15**
Colección Inicial de los Datos	**13 días**	**mar 01/09/15**	**jue 17/09/15**
Informe de Colección de Datos Iniciales	2 días	jue 03/09/15	vie 04/09/15
Descripción de los datos	**6 días**	**mar 01/09/15**	**mar 08/09/15**
Informe de Descripción de los Datos	3 días	vie 04/09/15	mar 08/09/15
Exploración de los Datos	**8 días**	**mar 01/09/15**	**jue 10/09/15**
Informe de Exploración de Datos	3 días	mar 08/09/15	jue 10/09/15
Verificación de la Calidad de los Datos	**15 días**	**mar 01/09/15**	**lun 21/09/15**
Informe de Calidad de Datos	4 días	mié 16/09/15	lun 21/09/15
Fase 3 :Preparación de Datos	**15 días**	**vie 04/09/15**	**jue 24/09/15**
Selección de los Datos	**8 días**	**mar 08/09/15**	**jue 17/09/15**
Criterios de Inclusión	5 días	**vie 04/09/15**	jue 10/09/15
Justificación de Inclusión	5 días	**vie 04/09/15**	jue 10/09/15
Limpieza de los Datos	**5 días**	**mar 08/09/15**	**lun 14/09/15**
Informe de Limpieza de Datos	2 días	**vie 04/09/15**	lun 07/09/15
Construcción de los Datos	**4 días**	**jue 10/09/15**	**mar 15/09/15**
Atributos Derivados	2 días	**vie 04/09/15**	lun 07/09/15
Registros Generados	2 días	**vie 04/09/15**	lun 07/09/15
Integración de los Datos	**3 días**	**lun 14/09/15**	**mié 16/09/15**
Datos fusionados	1 día	**vie 04/09/15**	vie 04/09/15

Formateo de los Datos	**4 días**	**mar 15/09/15**	**vie 18/09/15**
Datos Formateados	1 día	**vie 04/09/15**	vie 04/09/15
Conjunto de Datos	1 día	**vie 04/09/15**	vie 04/09/15
Descripción de los Datos	2 días	**vie 04/09/15**	lun 07/09/15
Fase 4 : Modelado	**16 días**	**lun 21/09/15**	**lun 12/10/15**
Selección de la Técnica de Modelado	**3 días**	**lun 21/09/15**	**mié 23/09/15**
Técnica de Modelado	2 días	**lun 21/09/15**	mar 22/09/15
Supuestos del Modelo	1 día	**lun 21/09/15**	lun 21/09/15
Generación del diseño de Pruebas	**3 días**	**lun 21/09/15**	**mié 23/09/15**
Diseño de Pruebas			
Construcción del Modelo	**15 días**	**lun 21/09/15**	**vie 09/10/15**
Ajustes de los Parámetros	2 días	**lun 21/09/15**	mar 22/09/15
Modelos	7 días	**mié 23/09/15**	jue 01/10/15
Descripción del Modelo	2 días	**vie 25/09/15**	lun 28/09/15
Evaluar el Modelo	**5 días**	**lun 21/09/15**	**vie 25/09/15**
Evaluación del Modelo	2 días	**lun 21/09/15**	mar 22/09/15
Revisión de los ajustes de Parámetros	2 días	**jue 24/09/15**	vie 25/09/15
Fase 5 : Evaluación	**10 días**	**mar 06/10/15**	**lun 19/10/15**
Evaluación de Resultados	**6 días**	**mar 06/10/15**	**mar 13/10/15**
Evaluación de los resultados de minería de Datos con referencia a los criterios de éxitos de negocio	2 días	**vie 09/10/15**	lun 12/10/15
Aprobación del Modelo	3 días	**mar 06/10/15**	jue 08/10/15
Revisión del Proceso	**4 días**	**mar 06/10/15**	**vie 09/10/15**

Revisión del Proceso	3 días	**mar 06/10/15**	jue 08/10/15
Determinar los Pasos Siguientes	2 días	**mar 06/10/15**	mié 07/10/15
Lista Posible de Acciones de Decisión	3 días	**mar 06/10/15**	jue 08/10/15
Fase 6 : Despliegue	**23 días**	**mar 22/09/15**	**jue 22/10/15**
Plan de Despliegue	**3 días**	**mar 20/10/15**	**jue 22/10/15**
Elaboración del Plan de Despliegue	3 días	mar 20/10/15	jue 22/10/15
Plan de Monitoreo y Mantenimiento	**4 días**	**lun 19/10/15**	**jue 22/10/15**
Elaboración del Plan de Monitoreo y Mantenimiento	2 días	**lun 19/10/15**	mar 20/10/15
Producción de los Reportes Finales	**48 días**	**jue 01/10/15**	**jue 18/10/15**
Reportes Finales	32 días	mar 20/10/15	mié 21/10/15
Presentación Final	22 días	mar 20/10/15	mié 21/10/15
Revisión del Proyecto	**8 días**	**mar 22/09/15**	**jue 01/10/15**
Fase IV : Cierre	**69 días**	**vie 21/08/15**	mié 25/11/15
Metodología de Investigación	**4 días**	**sáb 21/11/15**	**mié 25/11/15**
Comprobación de las Hipótesis 1 y 2	2 días	lun 23/11/15	mar 24/11/15
Levantamiento de Observaciones Finales	**5 días**	**Lun 30/11/15**	**Vie 04/12/15**
Modificación del Cronograma del Proyecto	3 días	lun 30/11/15	Mié 02/12/15
Añadir riesgos de Cloud Computing en el Plan Gestión de Riesgos (Anexo 5)	2 días	Mie 02/12/15	jue 02/12/15
Descripción del Cambio en la Organización	1 día	Vie 04/12/15	Vie 04/12/15

ANEXO 2

ACTA DE CONSTITUCIÓN DE PROYECTO

Información del Proyecto

Datos

PROYECTO	Sistema de predicción de hechos delictivos para la mejora del proceso de prevención del delito en el distrito de la molina utilizando minería de datos
CLIENTE	Comisaria de la Molina : Santa Felicia
PATROCINADOR PRINCIPAL	USMP
GERENTE DE PROYECTO	Jorge Julio Jaulis Rua
COORDINADOR DE PROYECTO	Renato Vilcarromero Giraldo

Propósito y Justificación del Proyecto

- Desarrollar un Sistema de Predicción de Hechos delictivos que permita mejorar el proceso de prevención de delitos actuales que tienen las comisarias del distrito, la fase que contempla el proyecto considerará la implementación en la comisaria de Santa Felicia.
- Lograr un impacto favorable en la prevención de delitos que logre una mejora del servicio policial en beneficio de la sociedad.

Descripción del Proyecto y Entregables

Acta de Constitución del Proyecto

Requerimientos de alto nivel

Requerimientos del producto

En líneas Generales el producto debe presentar lo siguiente.

Identificación de Zonas de Riesgo

Muestreo de Zonas a través de Mapas Geográficos.

Visualización de Reportes en base a registros históricos

El sistema debe ser multiplataforma (web, móvil).

Requerimientos del proyecto

El sistema ayude y mejore a :

- Identificar zonas de riesgo
- Mejorar el servicio policial de prevención de delitos
- Reducir los tiempos de acción de los efectivos policiales

- Aumentar el número de patrullajes realizados por zona.

Objetivos

Principal
Mejorar el rendimiento del proceso de prevención del delito de las comisarías de La Molina utilizando Minería de Datos.
Específicos
Mejorar los tiempos de acción de los efectivos policiales en la identificación de zonas de riesgo en el distrito de La Molina
Mejorar el servicio policial frente a la prevención de hechos delictivos en el distrito de La Molina.
Unificar y Estructurar la información que manejan las comisarías del distrito de La Molina que permita predecir la ocurrencia de hechos delictivos en base a datos históricos.

Riesgos iniciales de alto nivel

- Acceso de información al registro de denuncias

- Falta de comunicación con los efectivos policiales debido al trajín de su trabajo
- Ausencia de efectivos para la toma de encuestas

Presupuesto estimado

El presupuesto que se estima contempla costos de recursos humanos así como de hardware y software. Los mismos que se mostraran en rangos estimados inicialmente. El detalle de los mismos se debe mostrar en una sección del proyecto.

- Costos de hardware
- Costos de software
- Costos de recursos humanos

Lista de Interesados (stakeholders)

Nombre	Cargo	Departamento / División
Adín Alarcón Quispe	Oficial PNP	DEINPOL
Juan Soto Santiago	Oficial PNP	Estadística

Requisitos de aprobación del proyecto

Cumplimiento de los Objetivos denotados a través de los entregables Visualización del Sistema de predicción de Hechos Delictivos Verificación de que el Sistema sea multiplataforma (Web , dispositivos móviles)

Aprobaciones

Se adjuntan dos documentos firmados por los Stakeholders de la comisaria de la molina los cuales son:

- Aprobación de acceso a la información de denuncias del distrito.
- Documento de conformidad de levantamiento de información del proceso de prevención del delito y de implementación del sistema.

ANEXO 3
PLAN DE PROYECTO

Introducción

Esta sección contiene una visión general del proyecto y el producto a desarrollar, una lista de los entregables del proyecto y la estrategia de evolución del Plan.

Alcance del Proyecto

El proyecto está determinado para desarrollar un modelo de predicción de delitos en el distrito de La Molina, el cual utilizará una técnica de aprendizaje automático que será determinada según una comparación entre 2 o más algoritmos, de la misma manera se evaluarán diversas herramientas disponibles en el mercado para determinar cuál es la que más se acomoda a nuestra necesidad.

Al final, este modelo será implementado como un sistema de información, el cual mostrará en el mapa del distrito las zonas de riesgo identificadas con un color (esto según una modalidad de delito seleccionada). Al entrar al detalle de cada zona se podrá visualizar la fecha, rango de hora, tipo de lugar y probabilidad de ocurrencia de un hecho delictivo.

Como funcionalidad adicional, el sistema mostrará una sección de reportes donde se muestran gráficos de barras con las cantidades de denuncias según zonas, modalidades, comisarias, fechas, etc. en base a las denuncias históricas que se han registrado en las comisarías del distrito.

La solución estará disponible en aplicativos móviles y las descargas se realizarán a los equipos móviles que posean los efectivos policiales de las comisarias del distrito.

Adicionalmente también se plantea la solución en versión web con las mismas funcionalidades descritas.

Entregables del Proyecto

A continuación se detalla la lista de todos los entregables para el usuario, las fechas de entrega, lugar de entrega y condiciones de satisfacción (de ser requeridas).

IDENTIFICACIÓN ENTREGABLE	DESCRIPCIÓN ENTREGABLE	FECHA DE ENTREGA	LUGAR DE ENTREGA	CONDICIONES SATISFACCIÓN
Sistema de predicción versión móvil	Es la versión en Android de la solución	15/10/2015	Comisaria de Santa Felicia	N/A
Manual de usuario	Manual con el detalle de las funcionalidades del sistema	15/10/2015	Comisaria de Santa Felicia	Fácil de entende por el usuario
Sistema de predicción versión web	Es la versión web que será utilizada por los miembros de la sección de estadística	20/10/2015	Comisaria de Santa Felicia	

Estrategia de evolución del Plan

El plan de proyecto debe ser respetado ante cualquier circunstancia a menos que:

- El usuario pida una funcionalidad adicional.
- Algún integrante del equipo del proyecto no pueda continuar en el mismo.
- Los entregables diversos entregables del producto no cumplan con los requerimientos acordados.

Consideraciones del plan de proyecto

- Responsable de monitorear el Plan de Proyecto: Jonathan Vilcarromero
- Modificaciones al Plan: todas deben ser evaluadas.

- Cambios al Plan: serán evaluados por los dos miembros del equipo del proyecto
- Comunicados los cambios al Plan: actualización de los cambios realizados y comunicarlo al usuario de manera presencial o via email.

Organización del proyecto

Interfaces e interacciones

En esta sección se describen los procedimientos administrativos y de gestión entre el proyecto y: el Cliente, Gestión de configuración, Gestión de calidad y Verificación.

ACTIVIDAD	PROCEDIMIENTO	RESPONSABLE	INVOLUCRADOS
Despliegue de la solución 50%	Crear un ambiente productivo donde se alojara la aplicación y pueda ser utilizada en las comisarías, este despliegue abarca la funcionalidad de mapa predictivo	Jorge Jaulis	Jonathan Vilcarromero
Despliegue reportes	Acomodar el ambiente productivo instalado para que soporte la funcionalidad de reportes	Jorge Jaulis	Jonathan Vilcarromero

Responsables

Se identifican las actividades más relevantes en el proyecto, los responsables de dichas actividades y los involucrados.

IDENTIFICACIÓN DE ACTIVIDAD	DESCRIPCIÓN DE ACTIVIDAD	RESPONSABLE	INVOLUCRADOS
Plan de proyecto, cronograma, estimación de costos	Actividades relacionadas con lo necesario para poner en marcha el proyecto	Jonathan Vilcarromero	N/A
Desarrollo del modelo de predicción	Es la actividad principal de proyecto donde se crea el modelo de predicción.	Jonathan Vilcarromero	N/A
Desarrollo del sistema.	Es la actividad relacionada a desplegar el modelo de predicción en un sistema que pueda ser entendido y utuilizado por los usuarios finales	Jorge Jaulis	Efectivos policiales.

Proceso de Gestión

Las restricciones relacionadas con los riesgos que existen en el proyecto serán mostrados en otro entregable, tal como lo indica la guía PMBOK.

Objetivos y Prioridades de Gestión

Los objetivos del proyecto son:

- Mejorar el rendimiento del proceso de prevención del delito de las comisarías de La Molina utilizando Minería de Datos.
- Mejorar los tiempos de acción de los efectivos policiales en la identificación de zonas de riesgo en el distrito de La Molina.

- Mejorar el servicio policial frente a la prevención de hechos delictivos en el distrito de La Molina.
- Unificar y Estructurar la información que manejan las comisarías del distrito de La Molina que permita predecir la ocurrencia de hechos delictivos en base a datos históricos.

Mecanismos de control y ajuste

Mecanismos para la Gestión de calidad

Para el tema de la gestión de la calidad y el monitoreo de las actividades relacionadas al proyecto se utilizaran los entregables de definidos cuando se desarrolle el proyecto. Los cuáles serán mostrados como anexos posteriores.

Se planteara un manual de usuario que contendrá todos los procedimientos necesarios para poder utilizar de manera correcta el sistema final.

Mecanismos para Verificación y gestión de la calidad

Para la verificación y validación del modelo de predicción se utilizaran técnicas estudiadas en diferentes investigaciones de minería de datos. Con respecto a las pruebas del sistema de predicción, se elaborará un plan de pruebas.

Mecanismos para la Gestión de proyecto

- Reuniones diarias de 1 hora para reportar avances y complicaciones.
- Planteamiento de fechas para realizar visitas a la comisaria y obtener la información del proceso de prevención del delito.
- Establecimiento de fechas no aplazables para la entrega de avances y/o cambios en las actividades del proyecto.

Recursos

Recursos Humanos:

- Jonathan Vilcarromero.
- Jorge Jaulis.

Recursos materiales y/o tecnológicos estimados.

DESCRIPCIÓN	CANTIDAD
Laptop Toshiba Satellite 845C Core i3	**1**
Laptop Toshiba Satellite I845 Core i5	**1**
SmartPhone Huawei Y3C con Android 4.4	**1**

DESCRIPCIÓN	CANTIDAD	LICENCIA
Base de Datos		
María DB 10.1.7	1	**Open Source**
ETL		
Pentaho Data Integration	1	**Open Source**
Análisis Estadístico y Data Mining		
Weka 6.3.12	2	**Open Source**
Gestión de Proyectos		
TeamGantt	2	**Open Source**
Programación		
Android Studio	1	**Open Source**
Diseño		
Balzamiq Mockups	1	**Trial**
Pruebas		
Apache Jmeter	1	**Open Source**
Sistemas Operativos		
Windows 8.1 ServicePack 1	1	**Propietario**
Windows 8 ServicePack 1	1	**Propietario**
Documentación		
Microsoft Office 2010	**2**	**Propietario**

Proceso técnico

La metodología base que engloba a las demás será la de gestión de proyectos PMI, la que a su vez, en la fase de ejecución contendrá una metodología de minería de datos. Finalmente también se plantea una metodología de investigación la cual servirá en la fase inicial y en la parte donde se miden las hipótesis.

Procedimientos técnicos, herramientas y tecnologías

Se optó por el software libre como herramienta de desarrollo del modelo y sistema de predicción.

Esto está sujeto a cambio según sea necesario, previa aprobación del equipo de desarrollo.

Funciones de soporte

Entregables generados al hacer el aseguramiento de la calidad del producto final.

Líneas de trabajo, distribución de recursos humanos y responsabilidades

ROL	RESPONSABLE	FUNCIONES GENERALES
Analista de BI	Jonathan Vilcarromero	• Análisis y modelado del esquema de base de datos (según información recolectada) • Unificación la información de las denuncias registradas en las distintas comisarías (La Molina, Las Praderas, Santa Felicia) de la Molina. • Modelado y carga de los procesos ETL.
Analista Estadístico	Jonathan Vilcarromero	• Selección de variables input para el Análisis Estadístico. • Establecer coeficientes y/o pesos de variables input. • Establecer ejecuciones del modelo de aprendizaje automático. • Evaluar la incertidumbre en los resultados.
Diseñador	Jorge Jaulis	• Realizar los Mockups de los flujos de navegación que tendrá la aplicación.
		• Análisis de la arquitectura de la aplicación. • Desarrollo de capa BackEnd (lógica de

Desarrollador	Jorge Jaulis	negocio) de la Aplicación móvil. • Desarrollo de FrontEnd (capa cliente) de la aplicación. • Desplegar la aplicación en Google Play.
Analista QA	Jorge Jaulis /Jonathan Vilcarromero	• Hacer las pruebas funcionalidad para que se ratifique el correcto funcionamiento del Sistema. • Pruebas de stress para simular la concurrencia a la aplicación. • Pruebas de Aceptación de Usuario

ANEXO 4

PLAN DE GESTION DE INTERESADOS

PROYECTO:	Sistema de predicción de hechos delictivos para la mejora del proceso de prevención del delito en el distrito de La Molina utilizando minería de datos				
GERENTE:					
PREPARADO POR:	Jonathan Vilcarromero	FECHA	20	11	15
REVISADO POR:	Jorge Jaulis				
APROBADO POR:	Oficial PNP Adin Alarcon				

REVISION	DESCRIPCION (REALIZADA POR)	FECHA		
01	Jorge Jaulis	22	11	15

CONTEXTO DEL PROYECTO

Registro de interesados

Los interesados principales para este proyecto son:

- Juan Soto Santiago
- Adín Alarcón Quispe

Nivel de compromiso de los interesados

Juan Soto Santiago: Responsable de la sección de estadística de la Comisaria de Santa Felicia, principal interesado de la implementación del sistema a fin de mejorar sus actividades que realiza diariamente. **NIVEL ALTO.**

Adin Alarcón Quispe: Nivel de compromiso relacionado a las rutinas diarias conocidas como patrullajes que lleva a cabo su equipo. Por eso se considera **NIVEL ALTO**

Identificar interrelaciones y las superposiciones entre interesados

El jefe de la división de estadística (Juan soto) es quien brinda la información sobre la peligrosidad de las zonas de la jurisdicción al equipo que se encarga de hacer las intervenciones y/o redadas (Adin Alarcón). La interacción entre estos dos agentes es constante pues uno entrega

información de entrada ara que el otro pueda realizar sus actividades.

Necesidades de información de interesados

- Sistema geográfico que permita identificar zonas de riesgo con más alta probabilidad de ocurrencia.
- Reporte de denuncias históricas en las comisarías de distrito.
- Plataforma moderna para señalizar zonas y poder filtrar por delitos más comunes.

Forma de entrega de información y frecuencia

Forma de entrega:

Sistema entregado de forma digital (para la aplicación móvil). Para la aplicación web se brinda la url. Dos entregables planteados el día 25 de septiembre y 1 de octubre.

Manual de usuario: Entregado de forma digital. Se llevara a cabo al final del despliegue del sistema. Fecha estimada: 5 noviembre.

Forma de actualizar el documento en el ciclo de vida del proyecto

Las actualizaciones del presente plan deben ser revisadas y aprobadas por los dos miembros del equipo.

Si existieran nuevos Stakeholders se debe analizar si requieren nuevas funcionalidades y si se puede contemplar en el cronograma.

Por ningún motivo se debe aceptar cambios referidos a la funcionalidad si ya se llegó a un acuerdo con los Stakeholders principales (Los dos mencionados al inicio del plan)

ANEXO 5

PLAN DE GESTION DE RIESGOS

Roles y responsabilidades

ROL	RESPONSABLE	RESPONSABILIDADES
Analista de BI	Jonathan Vilcarromero	• Modelado del esquema de base de datos • Modelado de datos. • Unificación la información de las denuncias de las comisarias. • Modelado y carga de los procesos ETL.
Analista Estadístico	Jonathan Vilcarromero	• Selección de variables de entrada para el Análisis Estadístico. • Establecer coeficientes y/o pesos de variables input. • Establecer ejecuciones del modelo de aprendizaje automático. • Evaluar la incertidumbre en los resultados.
Diseñador	Jorge Jaulis	• Diseñar flujos de navegación que tendrá la aplicación.
Desarrollador	Jorge Jaulis	• Análisis de la arquitectura de la aplicación. • Desarrollo de capa BackEnd (lógica de negocio) del sistema. • Desarrollo de FrontEnd (capa cliente) del sistema. • Despliegue de la solución en producción.

Analista QA	**Jorge Jaulis / Jonathan Vilcarromero**	• Pruebas de Funcionalidad que ratifiquen el correcto funcionamiento del Sistema. • Pruebas de stress para simular la concurrencia a la aplicación. • Pruebas de Aceptación de Usuario

Aspectos claves a considerar

Gestión de Riesgos

Al tener una sola iteración se consideraran ciertos aspectos, los cuales serán respetados a lo largo del desarrollo del proyecto.

Los riesgos deben ser identificados, analizados y seguidos según los lineamientos que se establecen en el presente documento. No se debería incurrir en cambios debido a que el tiempo de ejecución del proyecto es corto y una aplicación de cualquier magnitud supone también una ampliación de tiempos.

Seguimiento de Riesgos

El seguimiento de los riesgos a lo largo del desarrollo del proyecto se realizara de la siguiente manera:

- Por cada fase del proyecto, se analizan los riesgos identificados y se evalúa su posible ocurrencia en la etapa siguiente.
- Se examina la metodología establecida para el seguimiento de riesgos.
- Se pone en marcha la trata del mismo.

Resumen del circuito de trabajo

Para asegurar un proyecto exitoso debemos tener en cuenta los riesgos que pueden poner en peligro la continuidad del mismo. Estos deben ser

identificados, analizados y tratados de una manera eficaz a fin de conseguir el objetivo esperado.

Identificación y evaluación de riesgos

Etapa de trabajo

El proceso de Identificación y Evaluación de Riesgos deberá efectuarse al inicio de cada iteración del proyecto, siendo ésta una de las primeras tareas a planificar. Debido al tiempo reducido de la ejecución del proyecto, solo tendremos una iteración por lo que no habrá repetición en las tareas.

- Limitación en el acceso de la información
- Disponibilidad de data histórica corta
- Requerimientos adicionales de stakeholders que aumenten el alcance del proyecto
- Cambios en los tiempos de cronograma de presentación de entregables.
- Aumentos en los costos por falta de capacidad de las herramientas a utilizar

Riesgos Cloud Computing

- La seguridad de los datos, debido a que toda tecnología está expuesta a sufrir algún inconveniente técnico o de mantenimiento, los datos de la organización quedarían expuestos en la nube. Debemos conocer a detalle la reputación de la empresa que nos presta el servicio.
- Control de usuarios, es importante llevar el control de los usuarios que accederán a nuestros datos almacenados, con esto debemos evitar que nuestra información pueda llegar a personas no autorizadas.
- Recuperación de los datos y sus copias de seguridad en el caso de cese en el servicio.

Oportunidades Cloud Computing

- Garantía de la protección de la información frente a desastres naturales. Considerando que el Perú es un país sísmico, tener la información en la nube es una buena oportunidad de mantener la información segura.
- No es necesario estar en un lugar físico específico. Mediante cualquier computadora o dispositivo móvil con acceso a Internet, se puede acceder a las aplicaciones y datos.
- Outsourcing del área de Tecnología: Al implementar Cloud Computing, se delega el soporte tecnológico a otra empresa especializada en ello.

Impacto de riesgos

Se define un estándar para la medición de los riesgos y qué tanto afecta al desarrollo del proyecto.

La hoja está formada por diferentes columnas, las cuales se dividen de la siguiente manera:

- Una columna para identificar cada riesgo, mediante un ID único.
- Tres columnas para indicar los distintos riesgos, categorías e impactos; los cuales son completados automáticamente.
- Una columna indicando la probabilidad de ocurrencia del riesgo
- Una última columna, para el cálculo automático del factor de riesgo

Metodología de trabajo

A cada uno de los riesgos en la tabla mencionada anteriormente se le deberá asignar un identificador único de referencia y, de acuerdo al criterio del evaluador, una probabilidad de ocurrencia (de 0% a 100%) en la columna correspondiente.

Automáticamente se calcula un factor (Impacto x Probabilidad x 100).

Una vez finalizada la notación de las probabilidades, se ordenará la tabla de forma de observar los riesgos de mayor factor. Seguidamente se

seleccionarán los más destacados para realizar el seguimiento en la próxima etapa.

A continuación se muestra la tabla con los resultados de la metodología adoptada

ID	RIESGO	CATEGORIA	IMPACTO	PROB. DE OCURRENCIA	FACTOR DE RIESGO
100	Limitación en el acceso a la información	Acceso a datos	10	50%	5
101	Disponibilidad de la data histórica corta	Desarrollo del proyecto	6	30%	1.8
102	Requerimientos adicionales	Desarrollo del proyecto	5	10%	0.5
103	Cambios en los tiempos del cronograma	Desarrollo del proyecto	5	10%	0.5
104	Aumento en costos de herramientas	Desarrollo del proyecto	3	20%	0.6

Seguimiento de Riesgos

Etapa de trabajo

El seguimiento de riesgos se llevará a cabo durante todo el proyecto. Teniendo en cuenta la identificación realizada de los riesgos potenciales.
Descripción del documento

Cada riesgo tendrá en cuenta con cuatro incisos: Identificación, Análisis, Plan de Riesgos y Seguimiento, donde los primeros dos fueron completados durante la etapa previa.
Metodología de trabajo

Los riesgos establecidos en la parte inicial del Plan de Riesgos, serán demarcados por una estrategia para la trata de los mismos, colocando el tipo de estrategia propuesta (Eliminación, mitigación o Contingencia), el

responsable de la tarea, la respuesta al riesgo (Actividad a realizar), y la etapa del desarrollo durante la cual se deberá efectuar.

El siguiente cuadro detalla los riesgos con su respectiva estrategia, responsable y demás aspectos a contemplar:

Riesgo	Estrategia	Responsable	Respuesta al riesgo	Etapa donde se efectua
Limitación en el acceso a la información	Contingencia	Jonathan Vilcarromero	Pedir el acceso a la información con carta formal de la universidad con 20 días de anticipación	Planificación
Disponibilidad de la data histórica corta	Mitigación	Jonathan Vilcarromero	Realizar un modelo que pueda adaptarse a predecir con data mínimo de un año de antigüedad.	Desarrollo del proyecto
Requerimientos adicionales	Contingencia	Jorge Jaulis	Generar documento donde se detallen las funcionalidades del sistema.	Desarrollo del proyecto
Cambios en los tiempos del cronograma	Eliminación	Jorge Jaulis	Se debe respetar a cabalidad los tiempos en el cronograma debido a que el tiempo de realización del proyecto s fijo.	Planificación del proyecto

Aumento en costos de herramientas	Contingencia	Jonathan Vilcarromero	Cambio de herramientas que generen demasiados costos que sobrepasen los rangos establecidos en el plan de proyecto	Planificación del proyecto

ANEXO 6
PLAN DE DESPLIEGUE

Descripción

El plan de despliegue detalla todas las actividades necesarias para que el sistema sea desplegado dentro de un ambiente y al mismo tiempo sea accesible, hasta la especificación del muestreo de los reportes.

Lo que se busca aquí es comprobar que el sistema sea desplegado en ambiente seguro y disponible en todo momento ya que la misma es fundamental para que los interesados vean no solamente la implementación del modelo de predicción del delito sino también la visualización de la información recabada de cada una de las comisarías de La Molina.

Propósito

El propósito fundamental del plan de despliegue es dar un marco general sobre el proceso de despliegue que se va a realizar, detallando el stack utilizado para el desarrollo del SISTPRE, siendo a ser esta nuestra etapa de desarrollo para el muestreo de los datos recabados por el modelo de predicción.

Referencias

Este documento está basado y/o referenciado en base a las documentaciones descritas por la metodología CRISP-DM y son pertenecientes a la última fase de despliegue que comprende un conjunto de documentos y actividades a realizar

Alcance

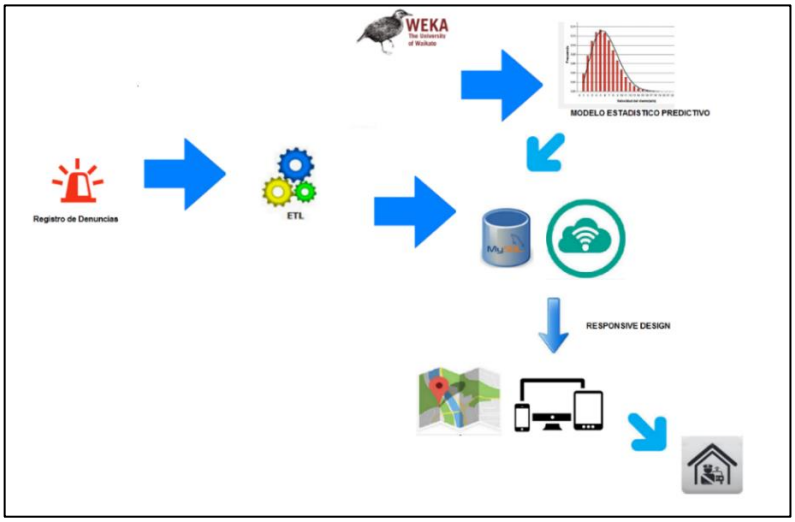
Desarrollo del sistema

El Plan de Despliegue inicia con el detalle de las actividades realizadas para la segunda etapa del proyecto; que es la implementación y desarrollo del SISTPRE.

ETAPA	SUBETAPA	COMPRENDE
DESARROLLO	Análisis	Recepción de los datos procesados por el Modelo de predicción, que ocurren en las anteriores fases de la metodología CRISP-DM - Comprendimiento del Negocio - Evaluación
	Diseño	El desarrollo del SISTPRE basado en dos componentes principales. - Maquetación - Programación
	Codificación	
IMPLEMENTACION	**Despliegue del SISTPRE**	**Compuesta por el despliegue de dos componentes principales** - Código Fuente (Frontend y Backend) - Base de Datos (Datos del Modelo de Predicción y de las comisarías)

Esquema del Despliegue

Para un mejor entendimiento de este plan se detalla el esquema general del SISTPRE siendo el siguiente.



Como podemos apreciar en el esquema ambos componentes de desarrollo se encuentran alojados en distintos repositorios, ya que esto nos permitía no solo trabajaren distintos entornos sino también hacia la aplicación mucha más modular, además de ello ya que están en la nube estas pueden ser accedidas por los usuarios desde cualquier punto.

CRONOGRAMA DEL PLAN

Responsables de las actividades de Despliegue

TAREA	RESPONSABLE
Elaboración y ajuste del Plan de Despliegue.	Jorge Jaulis
Configuración y despliegue del entorno de la base de datos	Jorge Jaulis
Configuración y despliegue del entorno del desarrollo del SISTPRE	Jorge Jaulis
Configuración de seguridad y automatización del SISTPRE	Jorge Jaulis

Especificaciones Técnicas

Base de Datos

PROPIEDADES	DESCRIPCIÓN
Versión de la Base de Datos	MySql 5.6
Región	Us – central
Replicación del Sistema de Archivos	Asíncrona
IP	173.194.251.77
Control de Acceso	SSL
Nombre de la BD	Sistpre
Copia de Seguridad	Diaria
Servidor	Linux Ubuntu 15.0
Registro Binario	Inhabilitado
Plan de Precio	Por uso
Hosting	Google Inc

SISTPRE

PROPIEDADES	DESCRIPCIÓN
Hosting	C9 io
Región	Us - central
RAM	42 MB de 512 MB
Disco	264 MB
Control de Acceso	SSH -RSA
Copia de Seguridad	Diaria
Servidor	Cent OS
Registro Binario	Inhabilitado
Plan de Precio	Free - Limitaciones

Como se ve ambos planes son free es decir son libres de pago pero poseen ciertas limitaciones, hasta el momento no se ha tenido ningún problema de accesibilidad sin embargo en este caso dependemos enteramente de estos Hosting.

Entre las principales requerimientos técnicos mostrados se ve que los servidores realizan ciertos guardados de réplica este es muy importante ya que se tiene un respaldo de la información guardada ante el surgimiento de cualquier eventualidad sea para nuestra BD o el repositorio del SISTPRE.

ANEXO 7

PLAN DE MONITOREO Y MANTENIMIENTO

Descripción

El plan de monitoreo y mantenimiento detalla todas las actividades necesarias para que el sistema sea monitoreado esto quiere decir que se mantengan métricas e indicadores que establezcan el acceso de los usuarios y su comportamiento dentro del SISTPRE, mientras la parte de mantenimiento es un proceso de retroalimentación para futuras mejoras que van a servir para que el sistema sea escalable.

Lo que se busca aquí es comprobar y tomar las previsiones respectivas para el mejoramiento de los servicios mostrados en el SISTPRE y mantener un control de los patrones de uso de los usuarios.

Propósito

El propósito fundamental del plan de monitoreo y mantenimiento es para brindar un marco referencial, de mejoramiento del SITPRE básicamente, para que este crezca en el tiempo y se adapte a nuevas funcionalidades o mejoras.

Alcance

Monitoreo del Sistema

El monitoreo del sistema comprende el análisis y seguimiento de la aplicación en base a patrones realizados por el usuario y este plan detalla cómo se hará tal seguimiento para ello.

Herramientas de Monitoreo

HERRAMIENTA	DESCRIPCIÓN
Google Analytics	Esta herramienta propia de Google permite realizar un monitoreo extensivo en base a los accesos del usuario, patrones de conducta y retorno de inversión, su versión free es bastante compacta y nos da muchas características , también existe una versión Pro
Applications Manager	Es un software propietario que posee herramientas

	iguales que las anteriores, ofrecen servicios específicos por tipo de tecnología y monitorea base de datos o aplicaciones. Cuesta 150 dólares mensuales el plan más básico.
Dynatrace	Dynatrace es un software de monitoreo, pero enfocado en móviles, a diferencia de los anteriores tiene una funcionalidad de geolocalización. Es también un software de pago con una versión de prueba de 30 días calendario y 30 dólares mensuales a partir de ello.

En base a un análisis donde el principal factor determinante fue el precio optamos por Google analytics que también tiene soporte en móviles y respaldada por una gran empresa como Google.

Escalabilidad del SISTPRE

Ya que el SISTPRE está sujeto a cambios a futuro es necesario especificar las tecnologías que actualmente usa tanto del Backend como en el Frontend. Tanto en su versión Web como Móvil

Tecnologías Frontend

Para el Lado del Cliente se ha utilizado Angular JS desarrollado por Google y que actualmente se encuentra en su versión 1.5

Tecnologías Backend

- Express Js se usa para el manejo de Middleware entre la capa de base de datos y la capa del cliente, está escrita en javascript y se encuentra en su versión 1.4.
- Node JS se ha utilizado como plataforma del Backend y el manejo de comunicación entre cliente servidor.
- Mysql como base datos de las denuncias registradas en el año 2015 y las predicciones probabilísticas del modelo.

Además de ello cabe mencionar que la versión Móvil está escrita en Java específicamente hecha para la plataforma Android que es la que posee mayor gama de celulares y es más extendida que IOS.

SISTPRE	VERSIÓN USADA
Angular JS	1.4
Mysql	5.6
Node	5.0
Express JS	4.0
Java	7.0

Cronograma del plan

En esta parte detallaremos básicamente las responsabilidades debido a que en el cronograma del proyecto ya figura un cronograma con la sección de tiempos de las tareas.

Responsables de las actividades de Despliegue

TAREA	RESPONSABLE
Elaboración y ajuste del Plan de Monitoreo y Mantenimiento.	Jorge Jaulis
Desarrollo de la interfaz de captura de comentarios	Jorge Jaulis
Configuración del entorno de BD para soporte de comentarios	Jorge Jaulis
Añadición y despliegue de la nueva característica	Jorge Jaulis

ANEXO 8
PLAN DE PRUEBAS

Descripción

El plan de pruebas detalla todas las actividades a seguir para asegurar que el sistema implementado cumple con las necesidades del negocio

Este plan comprende las pruebas de aceptación de la solución, que en nuestro caso contiene dos grandes bloques: el primero es el modelo de predicción y el segundo es nuestro sistema que implementa el modelo de predicción.
Adicionalmente presenta una serie de dashboards con información que maneja el proceso de prevención del delito.

Lo que se busca aquí es comprobar el cumplimiento de los requerimientos del negocio y detectar los defectos del producto para atacar la corrección de los mismos mediante en plan de acción.

Propósito

El propósito fundamental del plan de pruebas es dar un marco general sobre las pruebas que son necesarias realizar, según el tipo de solución desarrollada y los usuarios que utilizarán la misma y, además, designar los responsables para que cada prueba llegue a ser exitosa en los escenarios planteados.

Alcance

Ítems que van a ser probados

Se tienen dos grandes ítems:
El primero corresponde al ítem relacionado con la creación del modelo de minería de datos, donde se probara su eficacia hacia nuevos hechos delictivos no contemplados en la recolección de datos inicial.
EL segundo es un ítem de software que está relacionado a probar las funcionalidades del sistema, es decir los casos de prueba de las historias de usuario.

ITEM	CARACTERÍSTICA	TIPO DE PRUEBA
MODELO DE PREDICCIÓN	Precisión en predicciones	Datos de entrenamiento y prueba
	confiabilidad de predicciones	Validación cruzada (matriz de confusión).
	Utilidad de predicciones	
SISTEMA WEB	Ejecución de módulos separados	Pruebas Unitarias
	Concurrencia de usuarios al sistema	Pruebas de stress
	Funcionalidades de las HU	Pruebas funcionales

Debemos resaltar aquí que las pruebas del modelo de predicción serán ejecutadas en la fase 5 de la metodología de minería de datos (CRISP-DM) pero están planteadas aquí por ser un plan general del proyecto.

Documentación

EL documento fundamental de salida de la ejecución de las pruebas son los "Casos de prueba", donde se verá cada escenario con su resultado esperado. Adicionalmente se mostrara un resumen de las pruebas de stress (latencia del sistema, concurrencia de usuarios, etc.)

Características que van a ser probadas

Detallaremos la matriz que describe cada historia de usuario con su característica a ser probada.

Estrategia de regresión

Las pruebas de regresión se realizarán a aquellos módulos considerados como críticos, en los que se hayan detectado errores durante la ejecución de las pruebas.

En la matriz de funciones y casos de prueba de regresión se detallaran los casos de prueba que se ejecutarán de acuerdo al resultado y evolución del proceso de pruebas.

Criterio para decidir si un ítem supera la prueba

Para que un ítem supere la prueba es necesario que los errores de severidad 1, 2 y 3 que hayan sido encontrados sean removidos.

Los errores de severidad 4 se tratarán con el gerente del proyecto.

Definición de Niveles de Severidad

Cuando se reporta un defecto, los siguientes niveles de severidad seutilizaran:

NIVEL DE SEVERIDAD	DESCRIPCIÓN	EJEMPLO
1	Error inesperado en el sistema. No es posible continuar con el procesamiento.	Un error crítico ha sido encontrado y no permite que se continúe con la operación de la normal del sistema.
2	Bloquer: Conocido por ser una falla que al ocurrir, no es posible continuar con el proceso de la función que se había seleccionado	La funcionalidad que se intenta ejecutar no es de la forma esperada o simplemente no se carga (no se muestra).
3	Funciones restringidas, pero el procesamiento puede continuar	Componentes pocos relevantes no se muestran en la funcionalidad solicitada o se muestra de manera errada. Abarca información mostrada de manera incorrecta o algunos cálculos no son exactos.
4	Cambio de forma menor	Errores de usabilidad, pantallas o reportes de errores que no afectan la calidad, el uso ni la funcionalidad del sistema. Mensaje que no se entienden, falta de mensajes de confirmación ,etc.

Criterios de inicio y finalización de las pruebas

Criterios de Inicio

Se manejaran los siguientes criterios de inicio de las pruebas:

- Con respecto a las **pruebas funcionales**, se deben tener los casos de prueba específicos relacionados a su respectiva historia de usuario.
- Con respecto a las **pruebas de stress**, se debe tener definidos las funcionalidades a estresar, además de una herramienta que permita replicar “n” concurrencias al mismo tiempo
- Referido a las **pruebas de base de datos**, estas empezaran cuando se tenga un mino de 100 registros predichos. Debido a que si son menos podrían no verse fácilmente en el mapa temático.

Criterios de Finalización

- Las pruebas funcionales terminaran cuando el resultado de las actividades de los casos de pruebas sean las mismas a lo establecido en las historias de usuario. Esto involucra corregir los errores hallados al realizar los casos de pruebas.
- Las pruebas de stress terminaran cuando se tenga el número máximo de concurrencias a la aplicación manteniendo esta última de forma estable.

  Otro criterio de finalización es determinar la latencia (tiempo de respuesta en cada request) de las funcionalidades del sistema.

Criterios para la suspensión y reanudación de las pruebas

A continuación se enuncian los casos por los que las pruebas que se adelanten sobre el producto pueden ser suspendidas:

Una prueba funcional es suspendida si se encuentra un resultado diferente al esperado (el resultado esperado es el que figura en la Historia de Usuario).

Si se encuentra un bloquer que no permite continuar con las pruebas funcionales, entonces se detendrá la misma hasta solucionar el bloquer, luego se reanuda con la prueba funcional.

Responsabilidades del plan

Responsables de las actividades de Pruebas

TAREA	RESPONSABLE
Elaboración y ajuste del Plan de Pruebas de Aceptación.	Jonathan Vilcarromero
Elaboración y ajuste de los scripts de pruebas	Jonathan Vilcarromero
Elaboración y ajuste de los casos de pruebas	Jorge Jaulis
Elaboración de los datos de prueba	Jonathan Vilcarromero
Instalación del ambiente de pruebas	Jorge Jaulis
Ejecución de las pruebas de validación	Jorge Jaulis
Evaluación de las pruebas	Jorge Jaulis
Reporte de avance de las pruebas	Jonathan Vilcarromero
Reporte sumario de pruebas	Jonathan Vilcarromero

Requerimientos de ambiente de pruebas

Datos de Pruebas

Para la validación del modelo de predicción se utilizaron los siguientes datos:

Objetivos de cada prueba a realizar según su naturaleza

Esta lista representa qué será probado.

Pruebas a la Base de Datos

- Verificar que los datos predichos se inserten y se puedan consumir desde la aplicación.

Pruebas de Funcionalidad

- Verificar que los usuarios puedan ver la información que solicitan
  Verificar que el conjunto de datos capturados por el usuario se muestran correctamente
- Verificar que los usuarios que no tienen permisos no puedan ver la información
- Verificar que los contenidos de los reportes sean correctos
- Verificar que los reportes puedan ser exportados a Excel

Pruebas de interfaz de usuario (UI)

- Navegar a través de todos los casos del uso, verificando que cada panel de la interfaz de usuario pueda ser entendido fácilmente
- Verificar que todas las pantallas cumplan con los estándares de diseño de interfaz

Pruebas de Stress y desempeño

- Verificar que el tiempo de respuesta del sistema al recuperar la información en estación local
- Verificar el tiempo de respuesta cuando está conectado a la red de área local.
- Verificar el tiempo de respuesta de las conexiones en la red de acuerdo a los dispositivos de comunicación y sus velocidades (Modem, Internet, etc.).
- Verificar la respuesta del sistema con 1 usuario.
- Verificar el tiempo de respuesta con 2 usuarios concurrentes.
- Verificar la respuesta con 3 usuarios concurrentes.
- Determinar la cantidad de usuarios máximo que permite el sistema manteniendo el sistema estable.

Prueba de volumen

- Verificar el máximo número de clientes conectados todos desempeñándose de manera concurrente.
- Verificar que múltiples queries se ejecutan simultáneamente.

Pruebas de Fallas/Recuperación

- Verificar la recuperación del sistema ante la presencia de fallos de energía.
- Verificar la recuperación del sistema cuando exista corrupción de datos en la base de datos.

ANEXO 9

CUADRO DE OBJETIVOS VS VARIABLES

	PROBLEMA	OBJETIVO	VARIABLES A ANALIZAR	INDICADOR	MEDICIÓN
GENERAL	¿Se puede mejorar el rendimiento del proceso de prevención de delitos en las comisarías de la Molina utilizando minería de datos?	Mejorar el rendimiento del proceso de prevención del delito de las comisarías de La Molina utilizando Minería de Datos.	Percepción del personal sobre: • Identificación de zonas de riesgo. • Mejoran de controles preventivos	Preguntas analizadas según la escala de Likert	Encuesta de percepción. Encuestas sobre la realización el proceso
ESPECÍFICOS	¿Es posible mejorar los tiempos de acción de los efectivos policiales en la identificación de zonas de riesgos en el distrito de la molina?	- Mejorar los tiempos de acción de los efectivos policiales en la identificación de zonas de riesgo en el distrito de La Molina.	• Tiempo de identificación de zonas de riesgo.	T. antes T. después	Comparar dos situaciones: antes y después mediante encuesta sobre la identificación de zonas
	¿Es posible mejorar el servicio policial frente a la prevención de hechos delictivos en el distrito de La Molina?	- Mejorar el servicio policial frente a la prevención de hechos delictivos en el distrito de La Molina.	• zonas de riesgo detectadas	Cantidad de zonas antes del sistema. Cantidad de zonas después del sistema.	Comparar dos situaciones mediante encuestas sobre modo de asignación de zonas de riesgo
	¿Es posible estructurar y categorizar la información de las denuncias de las distintas comisarías del distrito de la Molina?	- Unificar y Estructurar la información que manejan las comisarías del distrito de La Molina que permita predecir la ocurrencia de hechos delictivos en base a datos históricos.	• Calidad de a información de BD. • Utilización de la BD.	Sistema, modelo de predicción, satisfacción del usuario.	Elaboración una BD unificada donde se registren todos los delitos, realizar encuestas a los usuarios.

ANEXO 10

ENCUESTA

Encuesta de percepción del sistema referida al proceso de prevención del delito

1. Con el sistema se pudo realizar mejores controles preventivos
   a) Totalmente en desacuerdo
   b) En desacuerdo
   c) Ni de acuerdo ni en desacuerdo
   d) De acuerdo
   e) Totalmente de acuerdo

2. El sistema me ayudó a identificar zonas de riesgo más rápido que con el método habitual
   a) Totalmente en desacuerdo
   b) En desacuerdo
   c) Ni de acuerdo ni en desacuerdo
   d) De acuerdo
   e) Totalmente de acuerdo

3. El sistema implantado me ayuda a prevenir delitos con dolo en la jurisdicción tales como robo, hurto, lesiones, daños, etc.
   a) Totalmente en desacuerdo
   b) En desacuerdo
   c) Ni de acuerdo ni en desacuerdo
   d) De acuerdo
   e) Totalmente de acuerdo

4. El sistema implementado sirve como herramienta de ayuda para tener una estadística adicional que colabora con la actual que posee la comisaria.
   a) Totalmente en desacuerdo
   b) En desacuerdo

c) Ni de acuerdo ni en desacuerdo
d) De acuerdo
e) Totalmente de acuerdo

Encuesta a los efectivos policiales

1. ¿Cuánto tiempo demoran en identificar una zona de riesgo en el distrito?
   a) Entre 1 – 4 minutos.
   b) Entre 5 – 8 minutos.
   c) Entre 9 – 12 minutos.
   d) Más de 12 minutos.
2. ¿Qué número de acciones preventivas realizan al mes?
   a) menos de 5
   b) Entre 6 y 10
   c) Entre 11 y 15
   d) Mas de 15
3. Según sus registros históricos y/o haciendo un estimado, ¿Cuántas acciones preventivas ha hecho es cada mes?
   a) Entre 10 y 20
   b) entre 20 y 30
   c) entre 30 y 40
   d) más de 40

ANEXO 11
MANUAL DE USUARIO

El manual de usuario del sistema de predicción de hechos delictivos, permite visualizar de manera gráfica los posibles delitos predichos en cada zona, con una probabilidad asignada de ocurrencia.

En este manual se explica detalladamente los pasos que deben seguir para el manejo general de la aplicación para la visualización de las zonas y también para poder visualizar los reportes. Además muestra el significado de los colores o marcas que se muestran en los mapas.

Los usuarios (los efectivos policiales) obtendrán información valiosa para mejorar el proceso de prevención del delito en la medida que el sistema acierte en sus predicciones.

Entre las bondades que ofrece el Sistema, se pueden citar las siguientes:

- ✓ Es amigable y de fácil manejo, ya que está disponible como solución web y móvil.
- ✓ Es configurable, permite filtrar por modalidad de delitos y ver la zona que se necesite.

Visión General del Sistema

Este manual de usuario será redactado tomando como referencia la parte móvil del sistema, debido las dos tecnologías (móvil y web) hacen exactamente lo mismo solo que por lo general las aplicaciones móviles se hacen más difíciles de comprender al ser una tecnología nueva.

Al ingresar al Sistema, el usuario tendrá una pantalla de bienvenida, donde podrá ver las noticias del distrito.

Para acceder a las opciones del sistema, el usuario encontrara dos opciones en la parte superior de la pantalla que son: ver mapa predictivo y ver reportes

Mapa Predictivo

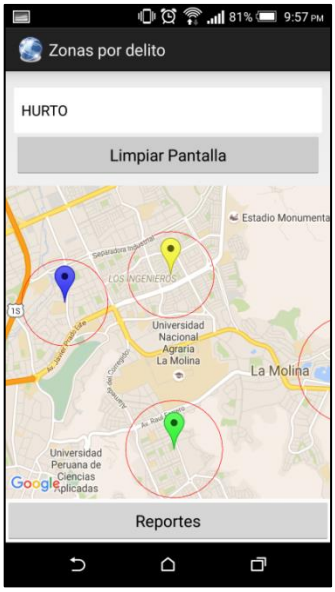
A continuacion se detallaran los paso para poder hacer búsquedas de las predicciones y poder ser mostradas en el mapa predictivo.

Al seleccionar la opcion “mapa predictivo”, el sistema carga el mapa, centrado en el distrito de la molina. En la parte inferior de muestra un combobox con las zonas disponibles para hacer busquedas tal y como se muestra en la figura:

Al elegir alguna opción del combo y seleccionar la opción mostrar zonas, inmediatamente se marcará en el mapa las zonas que tengan más coincidencias de búsqueda, esto es:

Si se eligió como modalidad de delito HURTO, entonces en el mapa se mostraran las zonas donde exista el delito predicho HURTO.

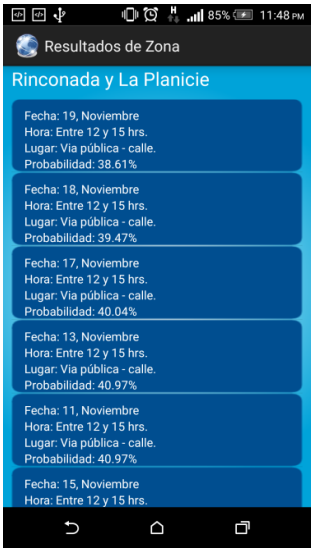
El resultado se muestra en una marcación en el mapa, la figura siguiente detalla una zona marcada en el mapa:



Esta zona demarcada corresponde a la zona demarcada. Aquí se muestra la siguiente opción:

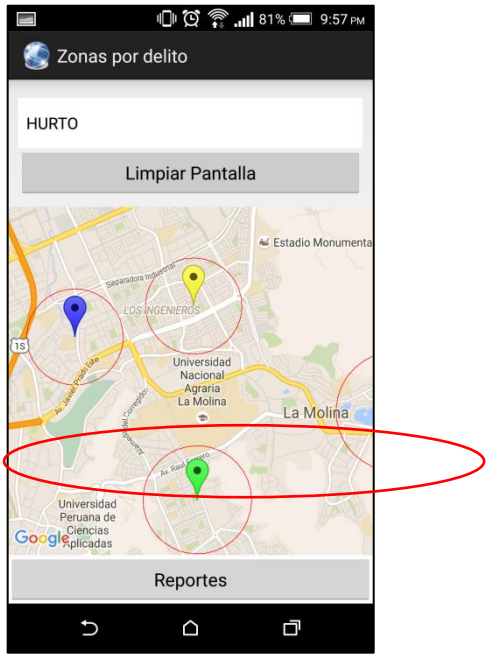
 La cual significa que la zona marcada contiene informacion predicha segun el delito ingresado.

Al seleccionar dicho marcador, el sistema muestra el detalle de las predicciones con los criterios de busqueda seleccionados (zona y modalidad de delito).



Reportes

Al entrar a la sección principal del sistema, en la parte inferior se ve la sección “Reportes”, tal como se muestra en la figura:



Al seleccionar esta sección, se podrán visualizar cuatro reportes que muestran información diversa con respecto a las denuncias:

Denuncias por zona

Este reporte muestra la cantidad de denuncias que existe en cada una de las zonas del distrito.

Modalidades de denuncia

El reporte muestra las denuncias agrupadas por modalidades, es decir, muestra las cantidades de denuncias que tiene cada modalidad.

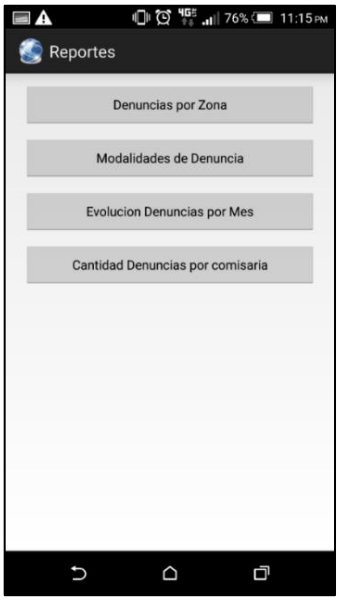
Evolución de denuncias por mes

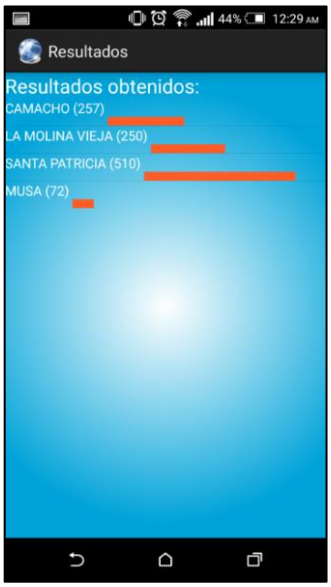
Este reporte muestra las denuncias registradas en cada mes del año para analizar el mes que tiene mayor frecuencia.

Cantidad de denuncias por comisaria

Muestras las denuncias en cada una de las comisarias.

La imagen que se muestra a continuación muestra la pantalla con el listado de reportes disponibles:





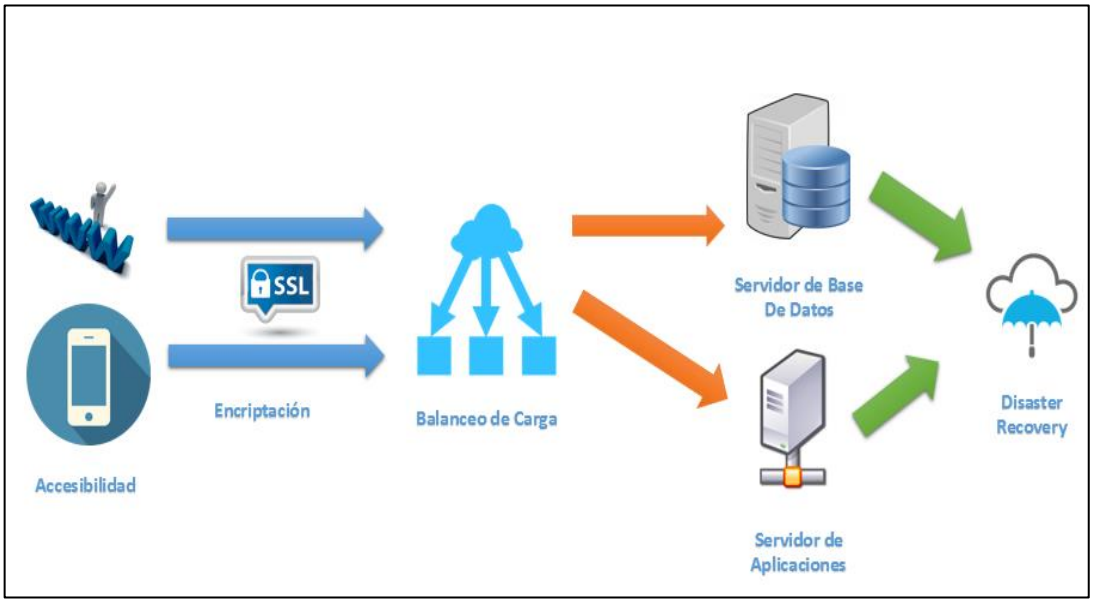
ANEXO 12
ARQUITECTURA DE SEGURIDAD

Introducción

Actualmente uno de los factores importantes en el desarrollo de software es el concepto de seguridad, teniendo en cuenta distintos especificaciones técnicas; entre las principales (credenciales, niveles de acceso, patrones de diseño, etc.)

En informática según (Woods, 2015) la **seguridad** se define como un conjunto de procesos y tecnologías que permiten monitorear y controlar la interacción entre un conjunto de "actores" y los recursos o componentes de un sistema.

Esquema de Seguridad



Como podemos observar en la figura de arriba el esquema de seguridad se subdivide básicamente en cuatro capas:

Accesibilidad: Hace referencia al número de conexiones existente que será determinado por la cantidad de usuarios por ello en SISTPRE se podrá acceder tanto desde dispositivos móviles como desktops.

Encriptación: Toda transmisión de información compartida a través del sistema es encriptado mediante SSL que garantiza que no se pueda hacer Sniffing (robo de información) cuando estén conectados a la red principal.

Balanceo de Carga: El balanceador de Carga básicamente se encarga de manejar la información que dispondrá el SISTPRE, en términos simples redirige las peticiones hechas por el usuario desde el lado del Backend que se conecta a los servidores.

Servidor: La cuarta y última capa por así llamarlo tiene que ver con la cantidad de servidores usados que en el caso del SISTPRE vendrían a ser dos los principales que son:

-Servidor de Base de Datos: Esta se encuentra alojado en el Cloud de Google

-Servidor de Aplicaciones: Esta se encuentra alojado en c9cloud

Concerniente a los servidores de disaster recovery (DR) o recuperación de desastres, se encargan las empresas donde están alojadas SISTPRE y nos facilitan la disponibilidad de las mismas.

Falencias

Como se ve actualmente en el esquema de seguridad no se han considerado los siguientes riesgos de software que también son muy comunes en el desarrollo de software.

- Gestión de Niveles Controles de Acceso
- Política de gestión de contraseñas
- Sistemas de detección de Intruso