UNIVERSIDAD NACIONAL DE TRUJILLO

Facultad de Ciencias Físicas y Matemáticas

Escuela Profesional de Informática

PREDICCIÓN DE LA RESPUESTA CORRECTA DE UNA PREGUNTA DE OPCIÓN MÚLTIPLE MEDIANTE TÉCNICAS DE APRENDIZAJE NO SUPERVISADO ASOCIADO CON LA EXPERIENCIA.

AUTOR: José Vicente Clavo Tafur

ASESOR: Jorge Luis Gutierrez Gutierrez

Trujillo - Perú

2019

PREDICCIÓN DE LA RESPUESTA CORRECTA DE UNA PREGUNTA DE OPCIÓN MÚLTIPLE MEDIANTE TÉCNICAS DE APRENDIZAJE NO SUPERVISADO ASOCIADO CON LA EXPERIENCIA.

José Vicente Clavo Tafur

PREDICCIÓN DE LA RESPUESTA CORRECTA DE UNA PREGUNTA DE OPCIÓN MÚLTIPLE MEDIANTE TÉCNICAS DE APRENDIZAJE NO SUPERVISADO ASOCIADO CON LA EXPERIENCIA.

Tesis presentada a la Escuela Profesional de Informática en la Facultad de Ciencias Físicas y Matemáticas de la Universidad Nacional de Trujillo, como requisito parcial para la obtención del grado de Bachiller en ciencia de la computación (Título profesional de Ing. Informático)

ASESOR: Jorge Luis Gutierrez Gutierrez

Trujillo - Perú

2019

HOJA DE APROBACIÓN

PREDICCIÓN DE LA RESPUESTA CORRECTA DE UNA PREGUNTA DE OPCIÓN MÚLTIPLE MEDIANTE TÉCNICAS DE APRENDIZAJE NO SUPERVISADO ASOCIADO CON LA EXPERIENCIA.

José Vicente Clavo Tafur

Tesis defendida y aprobada por el jurado examinador:

_____
Prof. Dr. XXXXXX - Asesor
Departamento de Informática - UNT

_____
Prof. Mg. XXXXXX
Departamento de Informática - UNT

_____
Prof. Mg. XXXXXXX
Departamento de Informática - UNT

Trujillo, xX de mayo del 2019

Dedico esta tesis a :

Mis padres quienes en todo momento me apoyaron. En especial a mi madre quien fue mi motor para seguir en mi camino y poder finalizar este proyecto.

Mi hermano con quien siempre compartimos conocimiento para así poder resolver algunos problemas que aparecieron en el transcurso del desarrollo de este proyecto.

A todas las personas que contribuyen con las ciencia en el Perú y cada día se esfuerzan para que esta siga creciendo

Agradecimientos

Agradezco a Dios por haberme bendecido en toda mi vida ....

A mis profesores del Departamento de Informática, de los cuales recibí una gran cantidad de conocimientos .

A mi asesor Prof. Dr. José Luis Gutierrez Gutierrez que siempre se mostró disponible e interesado en ayudarme.

. . .

Resumen

En la actualidad, hay algunas evaluaciones que son aplicadas en una determinada fecha y los resultados no son entregadas hasta algunos meses después, creando un ambiente de incertidumbre acerca de los resultados. Estas son evaluaciones que constan de preguntas tipo opción múltiple las cuales requieren de conocimiento académicos previos para poder responder correctamente. Por ejemplo una evaluación con esas características son las evaluaciones que miden el nivel en un idioma extranjero como puede ser el Inglés (ECCE,MET,etc). En este caso los resultados de estas evaluaciones son entregadas 2 meses después de haber sido aplicadas.

Esta investigación tiene como objetivo principal la predicción de la respuesta correcta de las preguntas de este tipo de evaluaciones. Para ello se usaran técnicas de aprendizaje no supervisado (Algoritmo K-means) asociado con la experiencia previa. Como experiencia previa se usará las calificaciones anteriores obtenidas de los 2 últimos exámenes de los estudiantes. Los resultados muestran que es posible estas predicciones mostrando un porcentaje de acierto de XX %.

Palabras claves: predicción, clusters, respuesta correcta, aprendizaje no supervisado.

Abstract

Nowadays, There are some tests which are taken in a specific date and their results are showed some months later. It creates a state of uncertainty about those results. These are tests which have multiple choice questions and previous knowledge is required to solve them. For example, a test with these characteristic is the test to measure the English level such as MET, ECCE and others. For these exams their results are showed two months later.

The main goal of this research is the prediction of the right answer of a multiple-choice question. To get this objective unsupervised learning algorithms (K-means)linked with previous experience will be used. In this case the previous experience will be the two last student's grades. The outcomes of this research show XX % right so these predictions are correct.

Keywords: prediction, cluster, rigth answer, unsupervised learning.

Lista de símbolos

Constantes:

(1) $r, \overline{r}$ Indice que denota regiones.

(2) n Indice de bienes finales deseados por los consumidores.

(3) ...

Variables:

(5) x^r Vector columna que denota la actividad de producción.

(6) u^r . . .

Índice de figuras

Índice de tablas

Índice general

Capítulo 1

Introducción

En esta investigación se realizará la predicción de la respuesta correcta de una pregunta de opción múltiple, este tipo de pregunta de una evaluación requiere de conocimiento previo para ser respondida, mediante técnicas de aprendizaje no supervisado asociado con la experiencia previa. Como experiencia previa se usará las calificaciones anteriores obtenidas de los 2 últimos exámenes de los estudiantes.

Para ello se plantea desarrollar un algoritmo de aprendizaje automático no supervisado llamado K-means, el cual sirve para crear clusters. Estos clusters se formarán con la data de las calificaciones anteriores de los estudiantes. Así se obtendría un valor promedio, Además, se utilizará técnicas de estadísticas para crear un histograma de las opciones de las preguntas respondidas. Combinando estas 2 técnicas se obtendría una predicción por cada pregunta.

Este proyecto fue pensado debido a la problemática que se tiene al aplicar una evaluación y tener que esperar demasiado tiempo para obtener las respuestas correctas. Pero con lo que se

desarrollara se obtendrá la respuesta correcta en corto tiempo.

Para cumplir los objetivos de este proyecto se tendrá primero que investigar acerca del aprendizaje automático no supervisado, así como las teorías de la probabilidad. Luego se desarrollará el algoritmo de aprendizaje automático no supervisado; ya hecho esto se procederá al desarrollo del prototipo de software web para poder ingresar los datos y hacer las pruebas. Finalmente se documentarán los resultados obtenidos.

1.1. Justificación de la investigación

La investigación es justificada por lo siguiente:

(a) En esta investigación pondremos en practica técnicas de aprendizaje de máquinas no supervisado, además se utilizará teorías de predicción para poder llevar acabo su objetivo.

(b) Esta investigación permitirá a los estudiantes tener un predicción del resultado de su examen en corto tiempo y así tener una idea de como sera su resultado. Además los educadores podrán usar esta información para la toma de decisiones.

(c) Los resultados de la investigación presentarán cómo resultado esperado, un prototipo de software que permita predecir la respuesta correcta de las pregunta de la evaluación. Este prototipo de software usará un algoritmo de aprendizaje no supervisado para trabajar con la información previa de los estudiantes (sus calificaciones anteriores).

1.2. Formulación del problema

En este trabajo, se propone responder a la siguiente pregunta:

¿ Cómo predecir la respuesta correcta de una pregunta de opción múltiple?

1.3. Hipótesis

Mediante el desarrollo de un algoritmo de aprendizaje no supervisado y experiencia previa, se puede predecir la respuesta correcta de una pregunta de opcion multiple.

1.4. Objetivos

1.4.1. Generales

a) Desarrollar un algoritmo de aprendizaje no supervisado para predecir la respuesta correcta de una pregunta de opción múltiple.

1.4.2. Específicos

a) Codificar el algoritmo de aprendizaje no supervisado para trabajar con clusters.

b) Desarrollar del prototipo de software y aplicarla a la predicción de la respuesta correcta.

c) Realizar las pruebas y documentar resultados.

1.5. Estructura de la tesis

El presente trabajo está dividido en cinco capítulos. El primer capítulo presenta los aspectos generales del tema tratado: justificación de la investigación, formulación del problema, hipótesis, objetivos y la estructura de la tesis.

En el segundo capítulo se contempla el marco teórico donde se hace referencia a los conceptos de predicción, pregunta de opción múltiple y de prototipo de software, además se hace referencia a la metodología de la investigación.

El tercer capítulo trata del tema central de la tesis, diseñándose el algoritmo K-means, se muestra el proceso para hacer la predicción y como se desarrolló el prototipo de aplicación web.

En el cuarto capítulo se presentan los resultados obtenidos en la investigación.

En el quinto capitulo se presentan las conclusiones, seguidas de las recomendaciones para futuras investigaciones relacionadas al tema en cuestión.

Finalmente las referencias bibliográficas usadas para la investigación. Además de los apéndices donde se presenta el algoritmo elaborado, los datos de los resultados y los instrumentos usados para recoger información.

Capítulo 2

Materiales y métodos

En este capítulo se explica cual fue la metodología empleada para la solución del problema formulado, además de una reseña del material bibliográfico investigado con relación a los temas considerados en esta investigación. Los conocimientos investigados son muy amplios, principalmente los que ayudaron a consolidar las bases del conocimiento científico para elaborar esta tesis, como lo son los temas de ********** FALTA COMPLETAR

2.1. Marco teórico

2.1.1. Aprendizaje no supervisado

2.1.1.1. Definición

Aprendizaje no supervisado es un método de adiestramiento automático donde un modelo es ajustado a las observaciones. Se distingue del aprendizaje supervisado por el hecho de que no hay un conocimiento a priori. En el aprendizaje no supervisado, un conjunto de datos de objetos de entrada es tratado. Así, el aprendizaje no supervisado típicamente trata los objetos de entrada como

un conjunto de variables aleatorias, siendo construido un modelo de densidad para el conjunto de datos.

En este método contamos con “objetos” o muestras, que tiene un conjunto de características, de las que no sabemos a que clase o categoría pertenece, entonces la finalidad es el descubrimiento de grupos de “objetos” cuyas características afines nos permitan separar las diferentes clases (Araujo, 2006).

2.1.1.2. Características

En este apartado se muestran las principales características del método de aprendizaje no supervisado. (Hinton and Sejnowski, 1999).

- No necesitan de un profesor externo.
- Muestran cierto grado de auto-organización.
- La red descubre en los datos de entrada y de forma autónoma: características, regularidades, correlaciones y categorías.
- Suelen requerir menores tiempos de entrenamiento que las supervisadas.
- Abordan los siguientes tipos de problemas: familiaridad, análisis de componentes principales, agrupamiento y relación de características.

2.1.1.3. Técnicas

Existen diferentes técnicas para el método de aprendizaje no supervisado.

Clustering: Agrupan objetos en regiones donde la similitud mutua es elevada.

Visualización: Permiten observar la amplitud de instancias en un espacio de menor dimensión.

Reducción de la dimensionalidad: Los datos de entrada son agrupados en subespacios de una dimensión más baja que la inicial.

Extracción de características: construyen nuevos atributos (pocos) a partir de los originales (muchos).

2.1.2. Clustering

2.1.2.1. Definición

Análisis de grupo o agrupamiento (en inglés, clustering) es un procedimiento de agrupación de una serie de vectores de acuerdo con un criterio. Esos criterios son por lo general distancia o similitud. La cercanía se define en términos de una determinada función de distancia, como la euclídea. La medida más utilizada para medir la similitud entre los casos es la matriz de correlación entre los NxN casos.

Generalmente, los vectores de un mismo grupo o clúster comparten propiedades comunes. El

conocimiento de los grupos puede permitir una descripción de un conjunto de datos multidimensional complejo. Esta descripción se consigue sustituyendo el detalle de todos los elementos de un grupo por la de un representante característico del mismo.

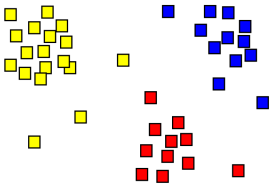

Figura 2.1: El resultado de un análisis de grupo, mostrado como la coloración de cuadrados en tres grupos. Fuente:

Clustering es considerado una técnica de aprendizaje no supervisado puesto que busca encontrar relaciones entre variables dependiendo del criterio de la distancia.

2.1.2.2. Técnicas de clustering

Existen dos grandes técnicas para el agrupamiento de casos:

- Agrupamiento jerárquico: En puede ser un agrupamiento aglomerativo o divisivo.
- Agrupamiento no jerárquico: En este agrupamiento el número de grupos se determina de antemano y las observaciones se van asignando a los grupos en función de su cercanía. Existen los métodos de K-means y K-medoid.

2.1.2.3. Algoritmos

Existen diversas implementaciones de algoritmos concretos. Por ejemplo, el de las K-medias o K-means. Es uno de los más antiguos pero su uso es muy extendido en la actualidad.

El algoritmo de K-means es un algoritmo particional y fue propuesto en los '50. Este algoritmo intenta encontrar una partición de nuestros ejemplos en K agrupaciones, de forma que cada ejemplo pertenezca a una de ellas, concretamente a aquella cuyo centro geométrico esté más cerca. El mejor valor de K para que la clasificación separe lo mejor posible los ejemplos no se conoce a priori, y depende completamente de los datos con los que trabajemos (Jain, 2010).

A pesar de que su primera aparición es desde hace más de 50 años sigue siendo de los algoritmos más utilizados para clustering por su facilidad de implementación, simpleza y buenos resultados empíricos.

Los principales pasos del algoritmo son los siguientes:

1 Generar las particiones iniciales determinada por los $c(1)....c(K)$ centros de los clusters ingresados.

2 Generar una nueva partición asignando cada dato$K$ al cluster$K$ cuyo centro está más cer-

cano.

3 Calcular los nuevos centros de los clusters $c(1)....c(K)$ (promediando los datos asignados a ese cluster en el paso anterior si la distancia es la euclídea).

4 Repetir los pasos 2 y 3 hasta que los clusters se lleguen a estabilizar.

El algoritmo K-means requiere del usuario los siguientes parámetros:

- Número de clusters.
- Inicialización de los clusters (centros).

2.1.2.4. Aplicaciones

- En marketing, para segmentar el mercado en pequeños grupos homogéneos donde realizar campañas publicitarias especificas.
- En biología, para dividir organismos en estructuras jerárquicas con el propósito de describir la diversidad biológica.
- En medicina, para diseñar tratamientos específicos para distintos grupos de riesgo.
- En psicología, para clasificar individuos en distintos tipos de personalidad, etc.

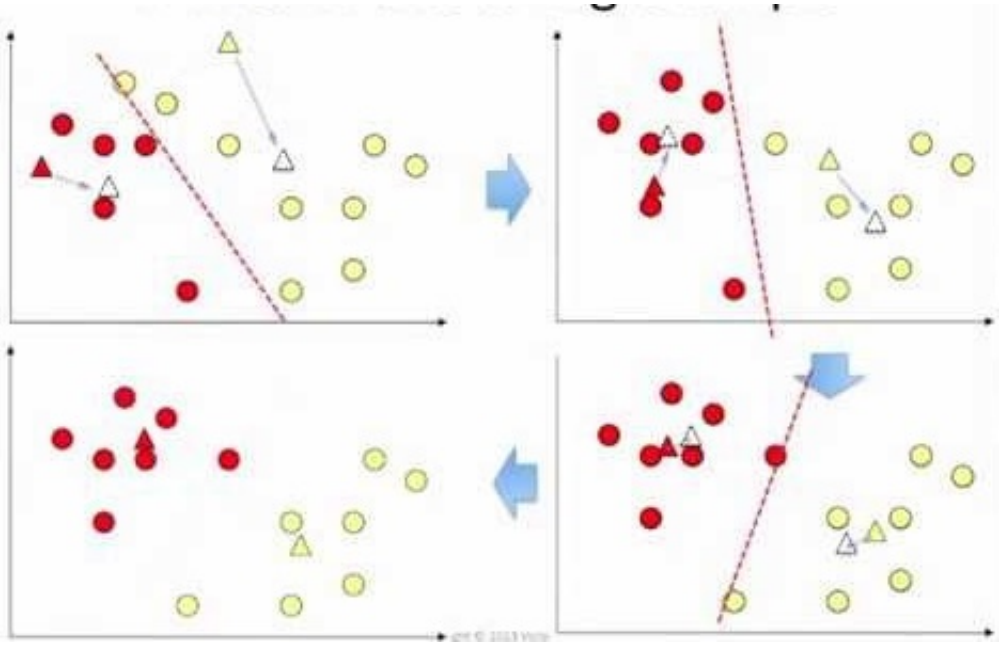

Figura 2.2: Proceso K-means.
Fuente:

2.1.3. Histograma

2.1.3.1. Definición

Según (UNAM, 2019) una histograma es una gráfica de la distribución de un conjunto de datos. Es un tipo especial de gráfica de barras y cada barra representa un subconjunto de los datos. Otra definición sostiene que un histograma es un gráfico de barras vertical que representa la distribución de frecuencias de un conjunto de datos (aiteco, 2019).

2.1.3.2. Tipos

Se agrupan los datos en clases, y se cuenta cuántas observaciones (frecuencia absoluta) hay en cada una de ellas. En algunas variables (variables cualitativas) las clases están definidas de modo natural, p.e sexo con dos clases: mujer, varón o grupo sanguíneo con cuatro: A, B, AB, O. En las variables cuantitativas, las clases hay que definirlas explícitamente (intervalos de clase) (HRC, 2019)

- **Simple:** Se representan los intervalos de clase en el eje de abscisas (eje horizontal) y las frecuencias, absolutas o relativas, en el de ordenadas (eje vertical).
- **Por grupos:** Se representan simultáneamente los histogramas de una variable en dos situaciones distintas.
- **Dirigido:** Se representan dos histogramas de la misma variable en dos situaciones distintas.
- **Estratificado:** Se representan el conjunto dividido en subconjuntos, a cada subconjunto se le denomina estrato.

2.1.3.3. Aplicaciones

HRC (2019) nos indica las siguientes utilidades.

- Proporciona, mediante el estudio de la distribución de los datos, un excelente punto de partida para formular hipótesis acerca de un funcionamiento insatisfactorio.

- El histograma es especialmente útil cuando se tiene un amplio número de datos que es preciso organizar, para analizar más detalladamente o tomar decisiones sobre la base de ellos.
- Es un medio eficaz para transmitir a otras personas información sobre un proceso de forma precisa.
- Permite la comparación de los resultados de un proceso con las especificaciones previamente establecidas para el mismo. Ayuda a determinar si el proceso satisface los requisitos del cliente.
- Hace posible determinar si ha habido cambios en un proceso.

2.1.4. Predicción

2.1.4.1. Definición

El término predicción puede referirse tanto a la acción y al efecto de predecir como a las palabras que manifiestan aquello que se predice; en este sentido, predecir algo es anunciar por revelación, ciencia o conjetura algo que ha de suceder (RAE, 2019).

Predicción es una expresión que anticipa aquello que "supuestamente"va a suceder. Se puede predecir algo a partir de conocimientos científicos, revelaciones de algún tipo, hipótesis o indicios.

2.1.4.2. Tipos

Científica: La predicción constituye una de las esencias claves de la ciencia, de una teoría científica o de un modelo científico. Así, el éxito se mide por el acierto que tengan sus predicciones (Mora, 1974).

La predicción en el contexto científico es una declaración precisa de lo que ocurrirá en determinadas condiciones especificadas. Se puede expresar a través del silogismo: "Si A es cierto, entonces B también será cierto".

El método científico concluye con la prueba de afirmaciones que son consecuencias lógicas de las teorías científicas. Generalmente esto se hace a través de experimentos que deben poder repetirse o mediante estudios observacionales rigurosos.

Una teoría científica cuyas aseveraciones no son corroboradas por las observaciones, por las pruebas o por experimentos probablemente será rechazada.

Predicción no científica: Los para-psicólogos y clarividentes, por su parte, apelan a pseudocientíficas para realizar predicciones. Estas personas dicen tener la posibilidad de conocer el futuro a partir de percepciones que reciben por sentidos que no son los cinco habituales (vista, olfato, gusto, oído y tacto). También hay sujetos que afirman recibir información por parte de entidades superiores(como dioses) para realizar sus predicciones (Julián and María, 2014)

En estos casos, las predicciones no tienen ningún sustento lógico, por lo que creer en ellas suele ser una cuestión de fe o depender de la susceptibilidad de la persona.

2.1.5. Probabilidad

2.1.5.1. Definición

La probabilidad es una medida de la certidumbre asociada a un suceso o evento futuro y suele expresarse como un número entre 0 y 1 (entre 0 % y 100 %).

Una forma tradicional de estimar algunas probabilidades sería obtener la frecuencia de un acontecimiento determinado mediante la realización de experimentos aleatorios, de los que se conocen todos los resultados posibles, bajo condiciones suficientemente estables. Un suceso puede ser improbable (con probabilidad cercana a cero), probable (probabilidad intermedia) o seguro (con probabilidad uno) (Loeve, 1978).

2.1.5.2. Teoria de la probabilidad

Según (Alvarez Franco and Rojas Rojas, 2010) la probabilidad p de que suceda un evento S de un total de n casos posibles igualmente probables es igual a la razón entre el número de ocurrencias h de dicho evento (casos favorables) y el número total de casos posibles n.

$$p = Prob(S) = \frac{h}{n}$$

La probabilidad es un número (valor) que varia entre 0 y 1. Cuando el evento es imposible se dice que su probabilidad es 0, si el evento es cierto y siempre tiene que ocurrir su probabilidad es 1. La probabilidad de no ocurrencia de un evento está dada por q, donde:

$$p = Prob(noS) = 1 - \frac{h}{n}$$

2.1.5.3. Teoría de bayes

El teorema de Bayes, en la teoría de la probabilidad, es una proposición planteada por el matemático inglés Thomas Bayes (1702-1761) (Bayes, 1763) y publicada póstumamente en 1763, que expresa la probabilidad condicional de un evento aleatorio A dado B en términos de la distribución de probabilidad condicional del evento B dado A y la distribución de probabilidad marginal de sólo A.

Las probabilidades Bayesianas se utilizan para encontrar la distribución de probabilidad condicional de un evento dados otros eventos , debido a su naturaleza se puede implementar al razonamiento bajo incertidumbre, los primeros estudios estaban basados en encontrar los errores sistemáticos en la estimación de probabilidades apoyados por la heurística, básicamente se da el análisis de la información que se utiliza para determinar la incertidumbre de la misma al realizar razonamientos probabilísticos, y encontrar el nivel de validez de los resultados obtenidos bajo esta perspectiva.

Formula de bayes:

Con base en la definición de Probabilidad condicionada (Parzen, XXXX) se obtiene la Fórmula de Bayes, también conocida como la Regla de Bayes.

Sea $A1, A2, ..., Ai, ..., An$ un conjunto de sucesos mutuamente excluyentes y exhaustivos, y tales que la probabilidad de cada uno de ellos es distinta de cero (0). Sea B un suceso cualquiera del que se conocen las probabilidades condicionales $P(B|Ai)$. Entonces, la probabilidad $P(Ai|B)$ viene dada por la expresión:

$$P(Ai|B) = \frac{P(B|Ai)P(Ai)}{P(B)}$$

Donde:

$P(Ai)$ son las probabilidades a priori.

$P(B|Ai)$ es la probabilidad de B en la hipótesis (Ai)

$P(Ai|B)$ son las probabilidades a posteriori.

2.1.6. Pregunta de opción múltiple

La pregunta de opción múltiple o de selección múltiple o multi-opción es una forma de evaluación por la cual se solicita a los encuestados o examinados seleccionar una o varias de las opciones de una lista de respuestas.

Este tipo de pregunta es usado en evaluaciones educativas (en lo que popularmente se llaman exámenes tipo test), en elecciones (para escoger entre múltiples candidatos o partidos políticos diferentes), en los cuestionarios para estudios de mercado, encuestas, estadística y muchas otras áreas.

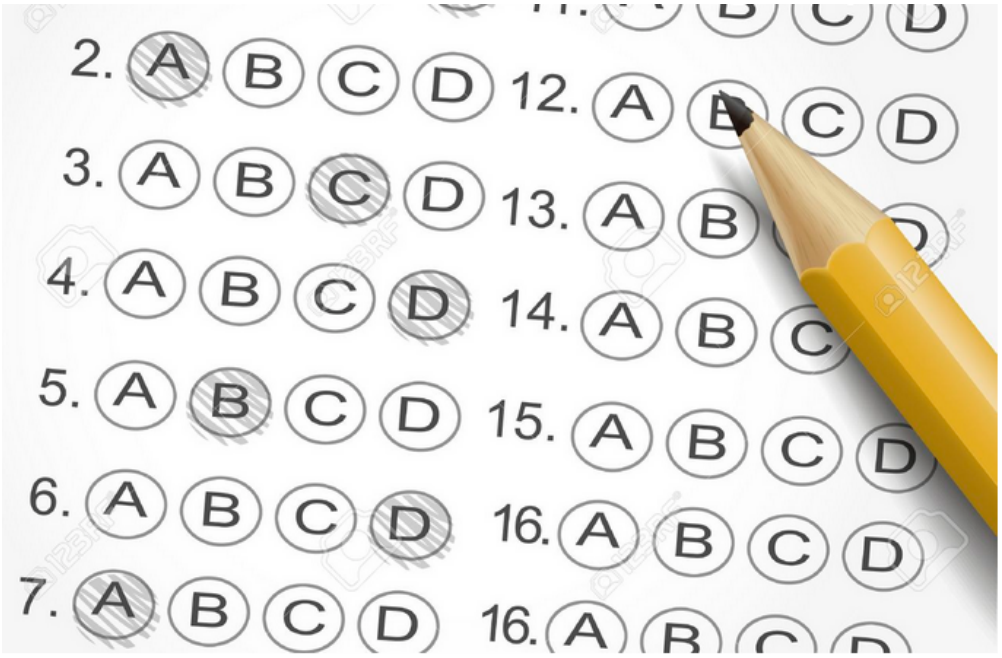

Figura 2.3: Hoja de respuestas de opción múltiple.
Fuente: (kchung, 2016)

Formato:

Un formato típico puede ser el de un enunciado seguido de una pregunta al respecto. El examinador ofrecerá normalmente de entre tres a cinco respuestas típicamente (a, b, c, d, e) de las cuales solamente una va a ser la respuesta correcta (la mejor respuesta) mientras las restantes respuestas

serán los distractores (Salkind, 2012).

2.1.7. Prototipo de software

2.1.7.1. Definición

En Ingeniería de software un prototipo es una representación limitada del diseño de un producto que permite a las partes responsables de su creación experimentar su uso (Pressman and Troya, 1988), el prototipo debe ser construido en poco tiempo, usando los programas adecuados y no se debe utilizar muchos recursos.

El diseño rápido se centra en una representación de aquellos aspectos del software que serán visibles para el cliente o usuario final. Este diseño conduce a la construcción de un prototipo, el cual es evaluado por el cliente para una retroalimentación; gracias a ésta se refinan los requisitos del software que se desarrollará. La interacción ocurre cuando el prototipo se ajusta para satisfacer las necesidades del cliente. Esto permite que al mismo tiempo el desarrollador entienda mejor lo que se debe hacer y el cliente vea resultados a corto plazo (Pressman and Troya, 1988).

2.1.7.2. Ventajas

- Este modelo es útil cuando el cliente conoce los objetivos generales para el software, pero no identifica los requisitos detallados de entrada, procesamiento o salida.
- También ofrece un mejor enfoque cuando el responsable del desarrollo del software está

inseguro de la eficacia de un algoritmo, de la adaptabilidad de un sistema operativo o de la forma que debería tomar la interacción humano-máquina.

- Se puede reutilizar el código.

2.1.7.3. Aplicación web

En la ingeniería de software se denomina aplicación web a aquellas herramientas que los usuarios pueden utilizar accediendo a un servidor web a través de Internet o una intranet mediante un navegador. En otras palabras, es una aplicación software que se codifica en un lenguaje soportado por los navegadores web (Mora, 2001).

2.1.7.4. LAMP

LAMP es el acrónimo usado para describir un sistema de infraestructura de internet que usa las siguientes herramientas:

- Linux, el sistema operativo.
- Apache, el servidor web.
- MySQL/MariaDB, el gestor de bases de datos.
- Perl, PHP, o Python, los lenguajes de programación.

La combinación de estas tecnologías es usada principalmente para definir la infraestructura de un servidor web, utilizando un paradigma de programación para el desarrollo (Apache, 2019).

A pesar de que el origen de estos programas de código abierto no han sido específicamente diseñado para trabajar entre sí, la combinación se popularizó debido a su bajo coste de adquisición y ubicuidad de sus componentes (ya que vienen pre-instalados en la mayoría de las distribuciones linux). Cuando son combinados, representan un conjunto de soluciones que soportan servidores de aplicaciones.

2.1.7.5. Software

- **Linux:** Linux es un núcleo de sistema operativo libre tipo Unix.
- **Apache HTTP Server:** El servidor HTTP Apache es un servidor web libre y de código abierto, el más popular en cuanto a uso, sirviendo como plataforma de referencia para el diseño y evaluación de otros servidores web.
- **MySQL:** MySQL es un Sistema de Gestión de Bases de Datos (SGBD) relacional, que por lo tanto utiliza SQL, multihilo y multiusuario del que se estiman más de un millón de instalaciones.
- **PHP:** PHP (acrónimo recursivo de "PHP: Hypertext Preprocessor") es un lenguaje de programación diseñado para producir sitios web dinámicos. PHP es utilizado mayormente en aplicaciones del lado del servidor.

Figura 2.4: Software usado en LAMP.
Fuente:

2.2. Método de la investigación

2.2.1. Metodología de la investigación

Se realizará una investigación aplicada, con una investigación cuantitativa, ya que se planea verificar con porcentajes y gráficos la cantidad de predicciones acertadas y al final verificar si las predicciones son correctas o no.

Para las pruebas se aplicara el estudio de caso con una sola medición en cual consiste en administrar un estimulo o tratamiento a un grupo y después aplicar una medición de una o mas variables para observar cual es el nivel del grupo en estas variables [AGREGAR CITA]. El diseño se diagrama de la siguiente manera:

$$G - - > X - - > O$$

Donde

$G =$ Grupo de sujetos.

$X =$ Tratamiento, estimulo o condición experimental.

$O =$ Medición de los sujetos de un grupo.

2.2.2. Población

La población para el presente estudio son las preguntas de la evaluación, para medir el nivel de inglés, de los 17 estudiantes de la sección 602 del centro de idiomas El Cultural. Esta evaluación consta de 18 preguntas y fue aplicada en Diciembre del 2016.

2.2.3. Muestra

Dado que cada evaluación consta de 18 preguntas y hay 17 estudiantes en la sección, en total tendremos 306 items de preguntas. Para esto se usara un muestreo aleatorio simple para poblaciones finitas con un nivel de confianza del 95 % y un error de muestreo del 5 %. [AGREGAR CITA]

$$n = \frac{N * p^2 z^2}{(N-1) * e^2 + p^2 z^2}$$

Donde

$n =$ Grupo de sujetos.

$N =$ Tamaño de la población. En este caso 170.57.

$z =$ Coeficiente de confiabilidad (1,96), lo que representa el 95 % de confianza.

$p =$ Desviación estándar de la población. 50 % = 0.50 (caso más desfavorable, desconocido).

$z =$ Margen de error 5 % = (0,05).

Aplicando la formula obtenemos que la muestra es de tamaño 170.57 = 171. La cual nos da un promedio de 10 preguntas por estudiante.

2.2.4. Variable de estudio

2.2.5. Instrumento de recolección de datos

- Formato para registrar las notas de los estudiantes. Apéndice B
- Formato para registrar las respuestas de los estudiantes. Apéndice C
- Prototipo de software web donde se ingresaran los datos.

2.2.6. Procedimiento

1. Formulación del problema principal de la investigación, justificando su importancia.
2. Búsqueda del material bibliográfico de los diferentes temas necesarios para la elaboración de la investigación, tales como aprendizaje automático no supervisado, probabilidades, estadísticas, desarrollo web, entre otros.
3. Estudio y análisis de los algoritmos de aprendizaje automático no supervisado.

4. Estudio de modelos probabilísticos y estadísticos que contribuyan con el tema de predicciones.

5. Desarrollo del algoritmo de aprendizaje automático no supervisado.

6. Desarrollo del prototipo de software web.

7. Elección de los datos de prueba.

8. Testear, validar y documentar los resultados de la investigación.

Capítulo 3

Tema central de la tesis

Basados en los conceptos discutidos en los capítulos 1 y 2. Se procederá a realizar el modelamiento de todo el proceso de predicción de la respuesta correcta, el desarrollo del algoritmo y de cada paso del proceso hasta llegar a resultado final.

3.1. Desarrollo del algoritmo de aprendizaje no supervisado K-means

3.1.1. Implementación

3.1.1.1. Diagrama de flujo

El algoritmo tiene diferentes pasos en su proceso los cuales vemos en la **figura 3.1**

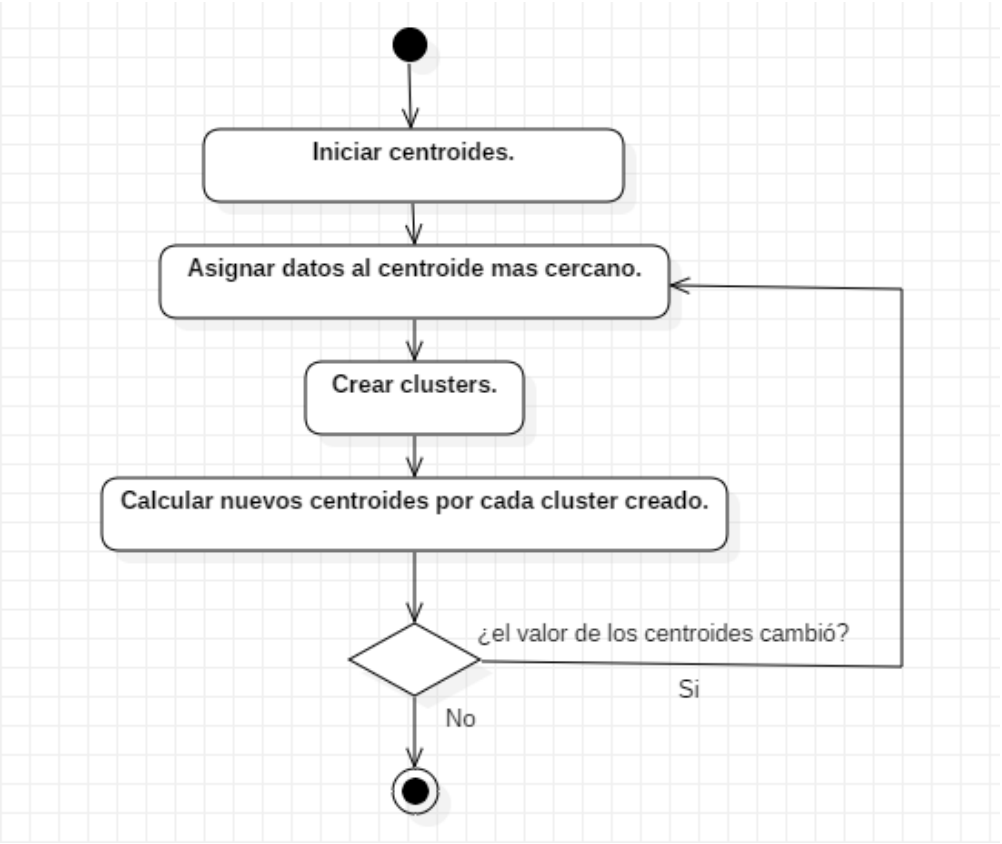

Figura 3.1: Diagrama de actividades/flujo algoritmo Kmeans.
Fuente: Propia

3.1.1.2. Pseudocodigo

Se muestra los pasos del **algoritmo 1** representado en pseudocodigo.

Algoritmo 1 Pseudocodigo Kmeans.

Entrada: Grupo de datos, Número de clusters.
Salida: Centroides, Datos con sus respectivos clusters.

1: Iniciar centroides.
2: Asignar datos al centroide.
3: Crear clusters.
4: Calcular nuevos centroides por cada cluster creado.
5: **SI** El valor de los centroides cambió **ENTONCES**
6: **Goto** [Paso 3]
7: **FIN SI**
8: **RETORNAR** Centroides, Datos con sus respectivos clusters.

3.1.2. Proceso

En esta sección se muestra el proceso que sigue K-means en las **figuras 3.7, 3.3, 3.4,3.5** a través de todas las iteraciones que realiza hasta lograr que los grupos se puedan estabilizar en ese momento las iteraciones finalizan.

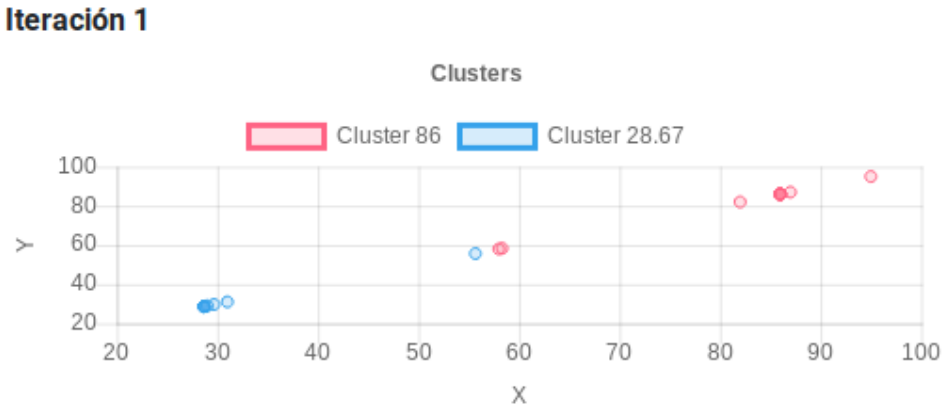

Figura 3.2: Grupos obtenidos en la iteración 1 .
Fuente: Propia.

Iteración 2

Figura 3.3: Grupos obtenidos en la iteración 2 .
Fuente: Propia.

Iteración 3

Figura 3.4: Grupos obtenidos en la iteración 3 (Final) .
Fuente: Propia.

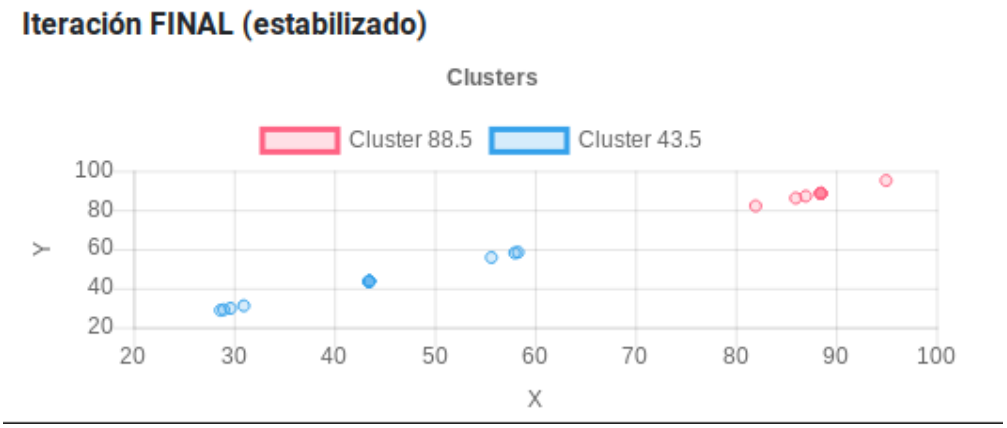

Figura 3.5: Grupos obtenidos en la iteración FINAL (estabilizado) .
Fuente: Propia.

3.2. Histograma

3.2.1. Implementación

3.2.1.1. Diagrama de flujo

El histograma tiene diferentes pasos en su proceso los cuales vemos en la **figura 3.6**

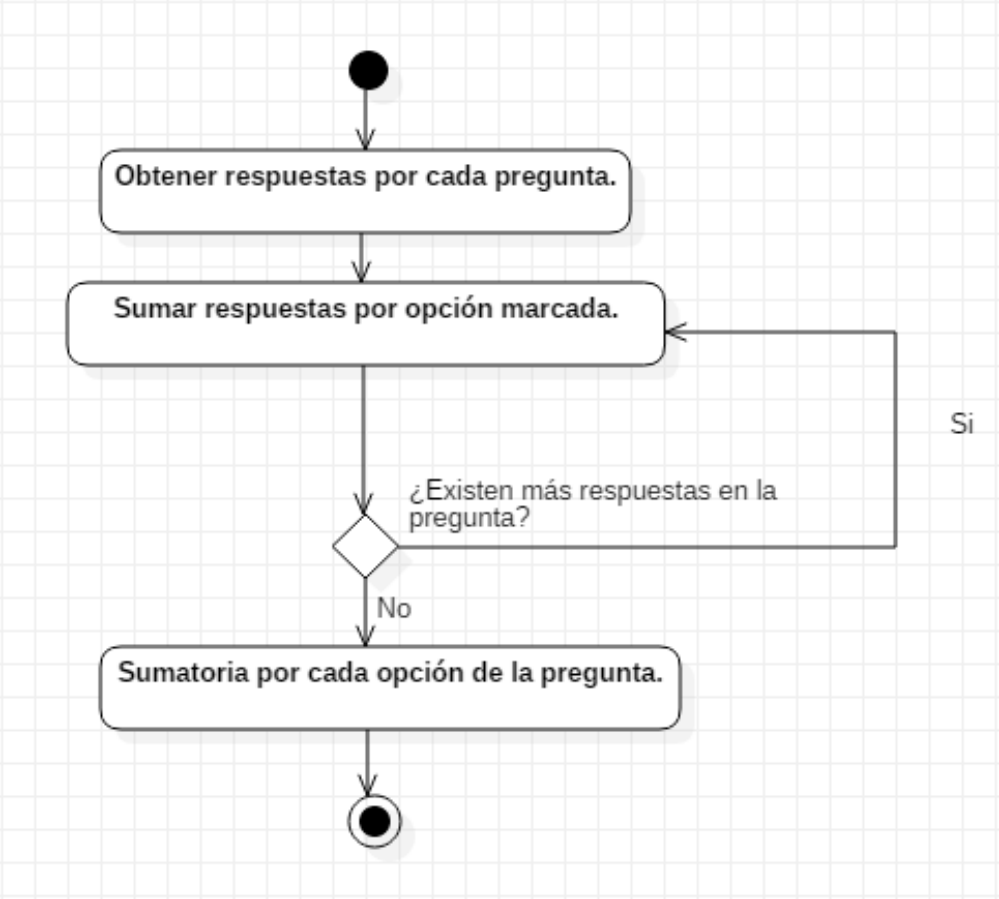

Figura 3.6: Diagrama de actividades/flujo Histograma.
Fuente: Propia.

3.2.1.2. Pseudocodigo

Los pasos representados en **algoritmo 2**.

Algoritmo 2 Pseudocodigo del actividades/flujo histograma.

Entrada: Respuestas de una pregunta.
Salida: Sumatoria por cada opción de la pregunta .

Obtener respuestas por pregunta.
Sumar respuestas según la opción marcada.
SI Existen más respuestas en la pregunta **ENTONCES**
    Goto [Paso 2]
FIN SI
RETORNAR Sumatoria por cada opción de la pregunta.

3.2.2. Proceso

En esta sección se muestra el proceso donde se genera el histograma, en las **figuras 3.7, 3.3, 3.4,3.5** a través de todas las iteraciones que realiza hasta lograr que los grupos se puedan estabilizar en ese momento las iteraciones finalizan.

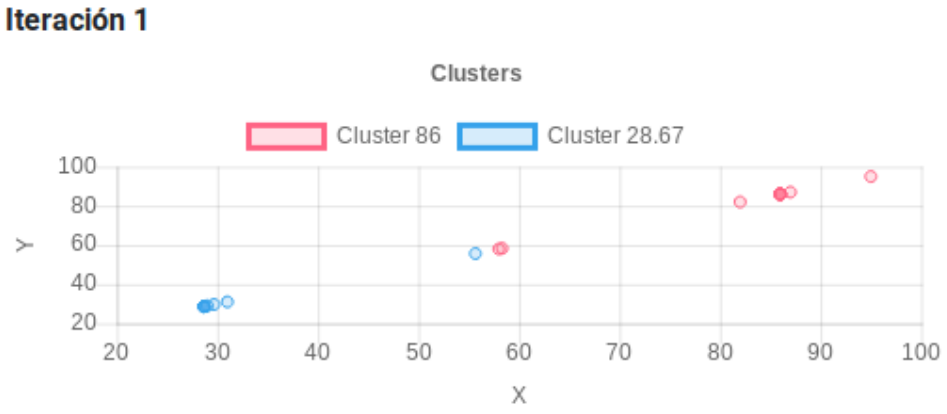

Figura 3.7: Grupos obtenidos en la iteración 1 .
Fuente: Propia.

3.3. Prototipo de aplicación web

3.3.1. Modelamiento del proceso

En esta sección vemos los pasos para poder realizar todo el proceso de la predicción de la respuesta correcta usando el prototipo desarrollado. En la figura **??** se muestra el diagrama de actividades de los procesos, la interaccion entre el usuario y el prototipo de aplicación web.

3.3.2. Desarrollo

El desarrollo del prototipo se realizo bajo un entorno Linux, usando Apache como servidor web, PHP como el lenguaje de servidor y HTML para la vistas. En la Figura 3.7 podemos ver la pantalla principal del prototipo donde muestra las diferentes opciones habilitadas. (MOSTRAR PANTALLA PRINCIPAL)

3.3.3. Base de datos

Algunos de los datos que se necesitan debían de ser permanentes, por lo cual se creo una pequeña base de datos para almacenarlos En la figura 3.8 vemos el diagrama relacional de la base de datos. (MOSTRAR DIAGRAMA DE LA BDD)

3.4. Proceso para realizar la predicción

3.4.1. Ejecutar K-means

En la figura 3.9 se muestra los resultados de ejecutar K-means. Esta ejecución nos da como resultado los clusters ordenados. Mostrando los datos de los estudiantes, el promedio de sus notas, el cluster al que pertenecen y el valor de estos clusters. (MOSTRAR RESULTADO DE LA EJECUCION KMEANS)

3.4.2. Histograma por pregunta

Por cada pregunta se obtiene cuantos estudiantes marcaron cada uno de las opciones, creando un histograma usando las respuestas por cada pregunta.

3.4.2.1. Modelado

El proceso del crear el histograma tiene diferentes pasos los cuales vemos en la Figura 3.10 **??**

3.4.2.2. Modelado

Se realizo la prueba de ejecución utilizando una pregunta de 4 opciones y con la respuesta de 7 estudiantes. En la figura 3.11 se muestra los resultados el histograma de una pregunta. En donde se muestra las opciones de la pregunta, el nombre los estudiantes que marcaron cada pregunta y el total de estudiantes por pregunta. (MOSTRAR RESULTADO DEL HISTOGRAMA)

3.4.3. Asignación valores de los clusters al histograma

Después de haber obtenido el histograma de una pregunta, Se asigna el valor del cluster que pertenece a cada estudiante en el histograma, obteniendo la suma de ellos por cada opción. Luego se suman los resultados de todas las opciones para sacar un porcentaje por cada una y así la opción con mayor porcentaje sera elegida como la correcta. (EXPLICAR EL PROCESO)

3.4.3.1. Modelado

El proceso de asignar valores de los cluster al histograma tiene diferentes pasos los cuales vemos en la Figura **??**.

3.4.3.2. Resultado

Se realizo la prueba de ejecución utilizando una pregunta de 4 opciones y con la respuesta de 7 estudiantes. En la figura 3.13 se muestra los resultados de la asignación valores de los 33cluster al histograma, mostrando también cual es la opción elegida como la correcta. En donde se muestra las opciones de la pregunta, el nombre los estudiantes que marcaron cada pregunta, el total de estudiantes por pregunta, la suma de los valores de los clusters de cada estudiante por cada opción, el porcentaje de cada opción de la suma de los clusters y cual es la opción correcta según la que tiene mas porcentaje de la suma de los valores de los clusters. (MOSTRAR CAPTURA Y ARREGLAR TEXTO)

3.5. Comparación de resultados

Las respuestas predecidas de cada preguntas se comparan con las respuestas correctas para poder obtener cuantas han sido predecidas correctamente.

3.5.1. Resultado

Se realizo la prueba de ejecución con 3 preguntas. La figura 3.14 muestra primero un resumen de cuantas predicciones de las respuestas de las preguntas fueron correctas y cuantas no Después muestra la lista de preguntas con su respuesta predecida, la respuesta que es la correcta y una opción que muestra si las 2 respuestas coinciden. (MOSTRAR DONDE SE COMPRAN LAS PREGUNTAS)

3.6. Proceso de modelamiento

Ejemplo:

La planificación y modelamiento del sistema de logística reversa de una área urbana es una fase importante y estratégica, para obtener en el futuro óptimos resultados en el proceso de gerenciamiento y operación del sistema reverso de RSU. El modelamiento permite determinar la localización de las estaciones de colecta y de unidades especiales necesarias, asi como el flujo que será movido a los largo de la red permitiendo dimensionar todo el sistema y sus componentes (Figura 3.1).

3.6.1. Proceso de ruteo

3.7. Implementación

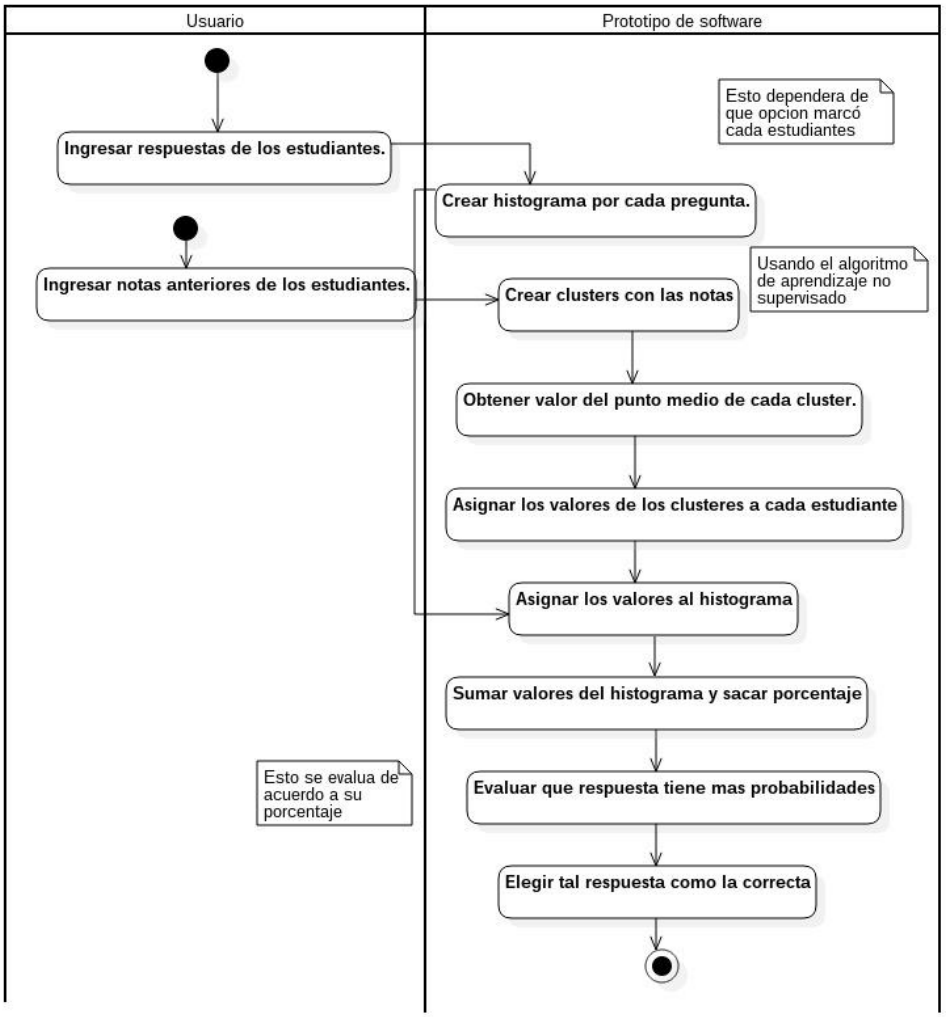

Figura 3.8: Diagrama de actividades de todo el proceso.
Fuente:

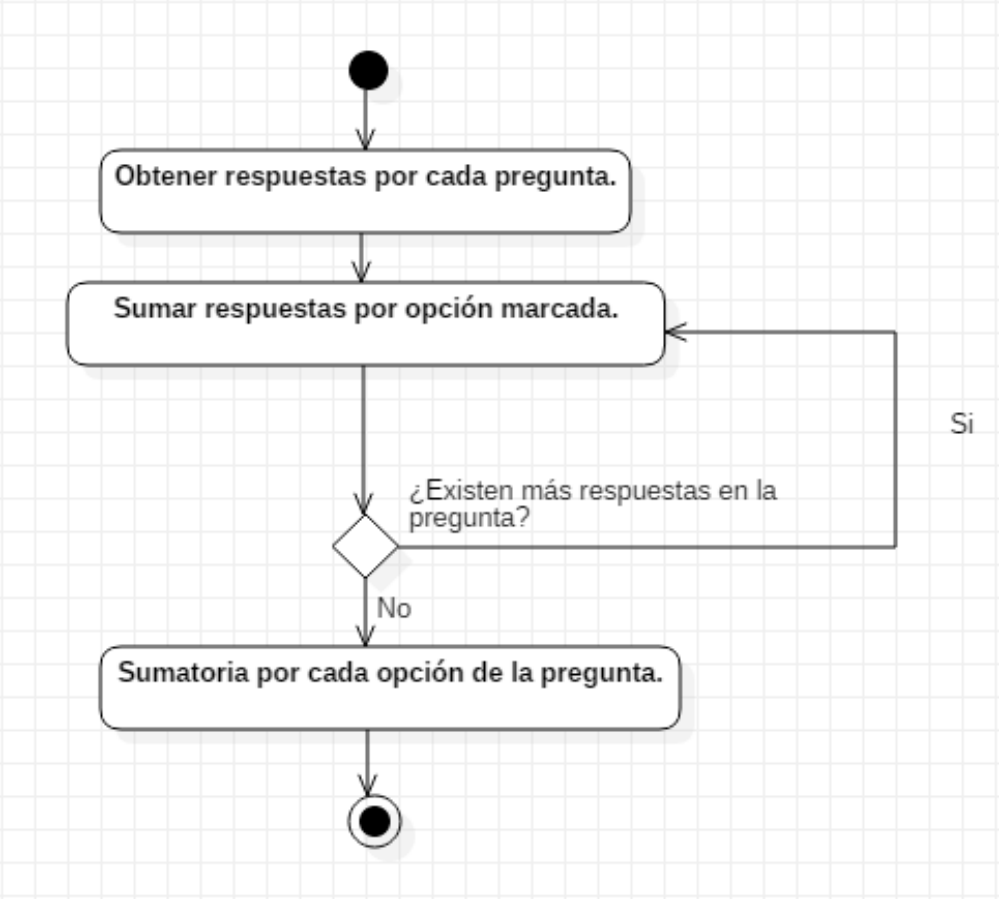

Figura 3.9: Pasos para el proceso del histograma.
Fuente: Propia

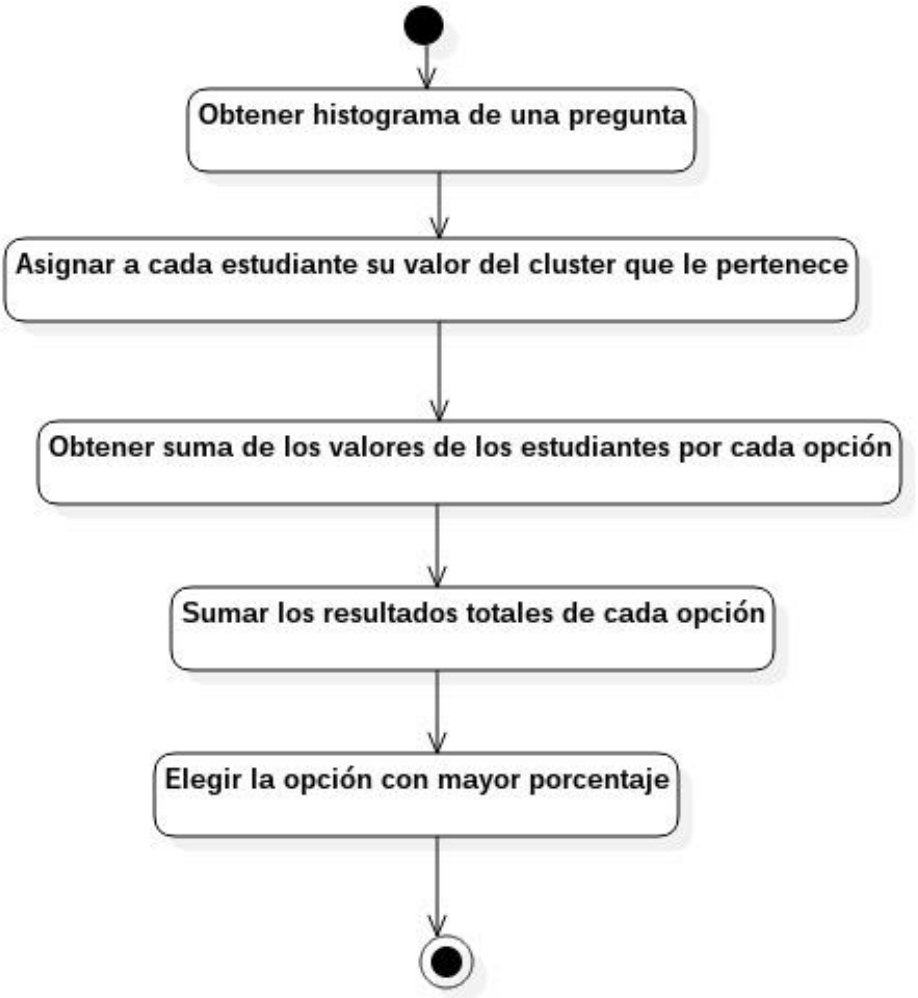

Figura 3.10: Pasos para la asignación valores de los cluster al histograma.
Fuente: Propia

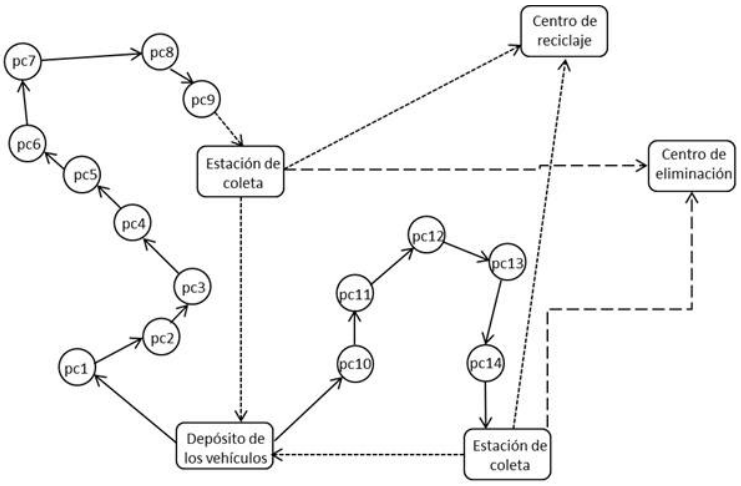

Figura 3.11: Esquema del proceso de colecta y transporte de RSU.
Fuente: Elaboración propia

Capítulo 4

Resultados y discusión de la tesis

Al culminar con la investigación se llegaron a resultados interesantes del punto de vista tanto teórico como computacional. Estos resultados muestran que se contrasta la hipótesis planteada durante el proceso de elaboración del plan de investigación, es decir, que se logró demostrar la relación entre las variables de estudio formuladas en la investigación.

4.1. Teóricos

4.2. Computacionales

Capítulo 5

Consideraciones finales

5.1. Conclusiones

Ejemplo

La investigación bibliográfica revela que realmente existe una preocupación de los gobiernos con el destino final de los residuos sólidos, con el objetivo de preservar la salud de la población y el medio ambiente urbano y rural. Por ejemplo, se observa la creación de la Ley 12305. Sin embargo existe una laguna entre las metas propuestas en la ley con las metas reales de los gobiernos locales. Eso se debe a la falta de una buena estructura organizacional, gerencial y operacional de los gobiernos locales capaz de atender las demandas locales y las necesidades de la población.

La falta de cuadros especializados, tanto en los gobiernos centrales como locales, para realizar la planificación y modelamiento de una red logística reversa puede ser compensada con la contribución de los investigadores que actúan en ese campo del conocimiento. Es muy difícil la formación de un equipo que tenga todo el conocimiento en las áreas de ciencia de la computación,

de geo procesamiento, de modelamiento matemático y de logística reversa, entre otras. Esa es una de las principales justificativas que los gobiernos, argumentan a la falta de planificación de una red logística reversa que funciones eficaz y eficientemente.

Por lo tanto, como quedó demostrado a lo largo de este trabajo, es posible realizar el modelamiento matemático para este tipo de problema con baja inversión, así como aplicarlo en varias regiones sin necesidad de grandes cambios en el modelamiento propuesto. El modelo propuesto calcula los flujos en la red logística reversa, permitiendo dimensionar la cantidad y capacidad de las unidades productivas y de los vehículos.

...

5.2. Trabajos futuros

Bibliografía

aiteco (2019). Histograma – herramientas de la calidad. https://www.aiteco.com/histograma/.

Alvarez Franco, L. C. and Rojas Rojas, J. B. (2010). *Teoría de la probabilidad.* Sello Editorial de la Universidad de Medellín.

Apache (2019). Xamp. https://www.apachefriends.org/es/index.html.

Araujo, B. (2006). *Aprendizaje automático: conceptos básicos y avanzados*. Pearson Education; 1st. edition (2006).

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, 1(2):370–418.

Hinton, G. E. and Sejnowski, T. J. (1999). *Unsupervised learning: foundations of neural computation*. MIT press.

HRC (2019). Ejemplos de tipos de representaciones gráficas. http://www.hrc.es/bioest/Ejemplos

Jain, A. (2010). Data clustering: 50 years beyond k-means. *Pattern Recognition Letters*, 31(8):651–666.

Julián, P. and María, M. (2014). Definición de predicción. http://definicion.de/prediccion/.

kchung (2016). Hoja de respuestas de opción múltiple.

Loeve, M. (1978). *Probability theory*. Technical report.

Mora, J. F., B. E. G. (1974). *Diccionario de filosofía abreviado.* Editorial Sudamericana.

Mora, S. L. (2001). *Programación en Internet: clientes web.* Editorial Club Universitario.

Parzen, E. (XXXX). *Teoría moderna de probabilidades y sus aplicaciones*. Limusa Grupo Noriega Editores.

Pressman, R. S. and Troya, J. M. (1988). *Ingeniería del software.*

RAE (2019). Predicción. http://dle.rae.es/srv/search=prediccion.

Salkind, N. (2012). *Métodos de investigación.* Pearson Education.

UNAM (2019). Histograma. http://asesorias.cuautitlan2.unam.mx/Laboratoriovirtualdeestadistica/DOCUMEN

Apéndice A

Primer apendice

hola como estas

Apéndice B

Segundo apendice

si te escucho

Apéndice C

Tercer apendice

sdsdsd

Declaración jurada y autorización

Los autores suscritos de la tesis:

escribir nombre de la tesis

declaramos su originalidad y autorizamos su publicación en el repositorio digital institucional y repositorio RENATI-SUNEDU, con el siguiente tipo de acceso: NOTA: Elegir un tipo de acceso

a) Acceso abierto: SI/NO

b) Acceso restringido (datos del autor y resumen del trabajo): SI/NO

c) No autorizo su publicación:

Si eligió la opción restringido o NO autoriza su publicación sírvase justificar:

El equipo investigador integrado por:

Apellidos y nombres	Condición de docente/estudiante	Código docente/Num. mat.	Autor/asesor
Peche Perlado Edgar	Estudiante	10101010	Autor
Pérez Yon Manuel	Estudiante	10101010	Autor
Rodríguez M. José	Docente	4010	Asesor

Edgar M. Peche Perlado

Manuel E. Pérez Yon

DNI: DNI:

José A. Rodríguez Melquiades

DNI: