

Universidad de Oriente

Sede “Julio Antonio Mella”

Facultad de Ingeniería en Telecomunicaciones Informática y Biomédica

Ingeniería Informática

3er Año

**Informe de la Práctica Laboral**

Autor: Fidel Alejandro Cepero Slazar

Tutor: MSc. Dionis López Ramos

Tema: Simulación del transito de una ciudad usando técnicas de Big Data

Proyecto de Investigación: Soluciones Informáticas para la gestión de los procesos universitarios a partir de plataformas de gobierno electrónico.

Comercio electrónico con tecnologías emergentes CELTEM.

2022

“Año 64 de la Revolución”

**Resumen:**

El presente trabajo consiste en la realización de un estudio y análisis de los datos obtenidos por un dispositivo.

El sistema de estudio se basa en la creación de un proceso de tratamiento de datos mediante el uso del lenguaje informático Python. El objetivo de este análisis es la obtención de la ocupación de un vehículo en cada punto de su recorrido y la verificación de su fiabilidad y eficiencia mediante el estudio de los datos obtenidos. También se valora cómo se podría realizar una predicción del comportamiento de esta ocupación de duchos vehículos, a partir de estos datos.

**Introducción:**

El proyecto de investigación para lograr el uso de las tecnologías de Información y Comunicación (gobierno electrónico), perteneciente a la Facultad de Ingeniería en Telecomunicaciones Informática y Biomédica de la Universidad de Oriente, se ha basado en el diseño de un programa para tratar los datos y en la evaluación del actual uso de sistemas Big Data en el transporte, para establecer la información relevante para el estudio de tráfico de usuarios de una red.

Con el presente estudio se pretende ofrecer varias contribuciones en este campo de investigación. En primer lugar, se recogen los principales conceptos teóricos sobre esta tecnología. En segundo lugar, se analizará el aprendizaje de la programación empleando técnicas de Big Data en el lenguaje Python. Y en tercer lugar, diseñar la solución de las mismas con elementos de sostenibilidad.

1. ****Estado del Arte****

**Se introducirán diversos aspectos relacionados con la pesquisa realizada sobre trabajos anteriores, y de información relativa al uso del Big Data, con la intención de mejorar la eficiencia en procesos de movilidad y transporte. Todo ello, con el objetivo de abordar el proyecto de la forma más eficaz posible. También se presentaran diversos tipos de algoritmos utilizados para realizar predicciones.**

* 1. ****Concepto de Big Data****

**El Big Data consiste en la captura del mayor número de datos posible con la idea de poder utilizar estos en un futuro para estudiar y mejorar la materia de la cual se han obtenido estos. En los últimos años, el valor de datos almacenados ha**

**crecido exponencialmente.**

**Por un lado, la tecnología ha permitido un aumento de la capacidad de almacenamiento. La aparición de “La Nube” permite a los usuarios guardar sus datos en servidores externos a través de Internet, en lugar de en sus dispositivos físicos. Esto es posible gracias a la aparición de equipos muy potentes y económicos.**

**Por otro lado, aunque ciertamente esta relacionado con lo mencionado anteriormente, se encuentra el crecimiento de dispositivos generadores de datos. Además, el surgimiento del “Internet de las Cosas” permite que cada vez más dispositivos estén conectados a Internet, por lo que también son**

**productores de datos en potencia.**

**Para explicar mejor que era este concepto, la empresa IBM publicó en 2012 “La Regla de las 4 V’s”, en las que indicaba las 4 dimensiones del Big Data [1]:**

**• Volumen: los datos se encuentran en diferentes escalas, de terabytes a petabytes.**

**• Variedad: también se presentan en diferentes formas, desde texto o multimedia a estructurados o no estructurados.**

**• Velocidad: el análisis de los datos debe hacerse de la forma más rápida posible para poder tomar decisiones con mayor presteza.**

**• Veracidad: la gestión de la fiabilidad y la previsión del tipo de datos obtenidos es muy importante para un buen resultado posterior.**

**Posteriormente, se añadieron nuevas V’s para definir este concepto, entre ellas:**

**• Visualización de los datos: una presentación fácilmente comprensible para los usuarios es vital para un buen uso de estos.**

**• Valor de los datos: este se obtiene de aquellos datos que se transforman en información, la cual es la que después se puede utilizar para mostrar a los usuarios.**

**Como se puede ver, actualmente ya existe una tecnología capacitada para esta captura y almacenamiento de capacidades ingentes de datos.**

* 1. ****Big Data en procesos de movilidad y transporte****

**Viendo el potencial del Big Data en aspectos como la movilidad, muchas ciudades y empresas han decidido invertir en este sentido, buscando mejorar diferentes aspectos.**

**Aprovechando los datos creados por los ciudadanos de las mismas, o utilizando sensores específicos para obtener los datos buscados, se esta buscando cada vez más la** **integración de esta tecnología en los procesos de mejora de eficiencia, dejando de lado** **los métodos “tradicionales”. Algunos ejemplos de esto:**

**• El proyecto Transforming Transport, organizado por la Unión Europea y con un presupuesto de 18,7 millones de euros. Compuesto por 48 socios a nivel europeo, entre los que se encuentran empresas como Indra o Ferrovial, con 13 desarrollos piloto programados. Entre éstos se encuentra la implementación en la ciudad de Valladolid de la tecnología utilizada en Finlandia con la idea de regular el tráfico por el centro de la ciudad y la entrega de mercancías. [2]**

**• En Santander, una de las ciudades que primero invirtió en esta tecnología, se trabaja con más de 10.000 sensores con la intención de captar el mayor número de datos posible. Actualmente existe un proyecto piloto que informa sobre los lugares de aparcamiento disponibles para reducir el tiempo de búsqueda de los conductores [3].**

**• En la ciudad de Lyon (Francia), utilizando la captura de datos repartidos por toda la ciudad, se pueden analizar las congestiones de tráfico y predecir posibles atascos, de esta manera se puede variar el cambio de las señales de tráfico para mejorar la circulación.**

**• En Santiago de Chile, y gracias a un acuerdo con dos compañías de telefonía móvil, analizan el flujo de desplazamiento de los usuarios de autobuses y metros de la ciudad mediante el uso de tecnología LBS, entre las que se encuentra el Wi-Fi o el Bluetooth [4].**

**Muchas otras ciudades han iniciado proyectos de características similares. La**

**cooperación entre empresas privadas, que disponen de la tecnología y los medios para llevar a cabo estos trabajos, y los organismos públicos son fundamentales para una mejora de la eficiencia del transporte en las ciudades utilizando todos los datos disponibles.**

**Uno de los primeros artículos referentes en este sector es The real-time city? Big data and smart urbanism” [5], en el que se citan diversos estudios sobre cómo el uso de Big Data puede afectar al transporte y urbanismo de un ciudad.**

**Como se ha podido comprobar, este es un tema novedoso y en auge dentro del mundo del Big Data. Por lo tanto, se cree que puede ser un campo con una aportación importante en el futuro inmediato.**

****1.3. Minería de Datos****

**Como se ha indicado anteriormente, el volumen actual de datos almacenados por gobiernos, empresas y organizaciones es enorme. La razón de este hecho es el valor que tiene la información que se puede obtener de estos datos. Los datos son la materia prima y cuando el usuario pasa a darle cierto significado, se convierten en información.**

**Siguiendo esta lógica de pensamiento, para obtener información, será necesario analizar cantidades masivas de los anteriores mediante procesos de filtrado y selección. Una vez obtenida la información, esta puede ser estudiada con la intención de obtener un modelo el cual de sentido a esta y proporcione conocimiento.**

**Es en este proceso donde la minería de datos entra en acción. Esta trabaja buscando patrones o tendencias que sigan los datos tratados en busca de modelos que nos permitan comprender mejor el área estudiada. Estos modelos podrán ser utilizados para ayudar a tomar decisiones.**

**Desde el punto de vista académico, la minería de datos es un término que representa una etapa de un proceso mayor, la extracción de conocimiento en bases de datos .**

**Actualmente, en el entorno comercial, se utilizan ambos términos como sinónimos. La minería de datos no es un concepto moderno. Desde mediados del siglo pasado ya se nombraba la acción de encontrar correlaciones sin una idea previa en bases de datos sin filtros previos con términos similares. Por otro lado, a causa del aumento de datos almacenados gracias a las mejoras tecnológicas en este campo y al menor coste y mayor capacidad de los procesos de tratamiento de datos, el crecimiento de mineria de datos ha sido exponencial en los últimos 20 años [6].**

****1.3.1. Etapas de un proyecto de Minería de Datos****

**La tecnología usada en un proceso de minería de datos es muy variada. Son procesos que se desarrollan en diversas etapas y requieren de diferentes softwares. En términos generales, las etapas de un proyecto son las siguientes [7]:**

1. **Determinación de los objetivos: previamente al tratamiento de los datos, es necesario marcar los propósitos que se quieren obtener en el proyecto.**
2. **Selección del conjunto de datos y análisis inicial de sus propiedades: una vez se han indicado los objetivos, se escoge la base de datos existente más óptima y se analiza. Esto significa que se muestran los datos buscando posibles valores atípicos o ausencia de datos.**
3. **Preprocesamiento de los datos: en esta etapa se realiza el proceso de filtrado, selección, reducción y transformación de los datos iniciales con la idea de poder analizarlos. Esta parte del proceso consume la mayor parte del tiempo total del proyecto y es la etapa más laboriosa y costosa.**
4. **Determinación del modelo: trata de analizar estadísticamente los datos y mostrarlos gráficamente con la idea de poder visualizarlos y tener una imagen inicial de estos.**
5. **Análisis de los resultados: verificación de la coherencia de los resultados obtenidos en los análisis anteriores y mostrados gráficamente. En esta etapa se valorará si la información es valiosa para el usuario y si el modelo puede ayudar a la futura toma de decisiones.**

**Estas etapas son muy generales y normalmente se dividen en numerosas sub-etapas más específicas. Cada proyecto es diferente y puede utilizar diferentes formas de llevar a** **cabo el proceso.**

****1.3.2. Principales herramientas de la Minería de Datos****

**Las técnicas utilizadas para la realización de minería de datos provienen, principalmente, de la estadística y la inteligencia artificial. Estas técnicas consisten en algoritmos con un** **nivel menor o mayor de sofisticación ya existentes y conocidos por su uso en otros ámbitos. Las principales herramientas utilizadas son las siguientes [8]:**

**• Redes neuronales artificiales: se trata de modelos predictivos no lineales y automáticos que se asemejan, estructuralmente hablando, a la forma en que funciona el sistema nervioso de los animales. Se trata de un sistema de interconexión de neuronas en una red que colabora para producir un estímulo de salida.**

**• Reglas de inducción: en base a cálculos estadísticos es posible generar reglas útiles que siguen un patrón y pueden aplicarse a los datos a analizar para extraer conclusiones y hacer pronósticos.**

**• Árboles de decisión: modelo de predicción en el que se construyen diagramas en forma de árbol a partir de una base de datos. Sirven para representar y categorizar una serie de condiciones que suceden de forma sucesiva.**

**• Regresión lineal: Rápida y eficaz pero insuficiente en espacios tridimensionales.**

**• Agrupamiento: Agrupa vectores según criterios, habitualmente de distancia, disponiendo vectores de entrada de forma que estén más cercanos aquellos con características comunes.**

****1.3.3. Principales algoritmos de la Minería de Datos****

**Según el objetivo del análisis de datos, los algoritmos se dividen, principalmente, en los siguientes tres tipos [9].**

****1.3.3.1Algoritmos aprendizaje supervisado****

**El aprendizaje supervisado consiste en realizar predicciones futuras basadas en comportamientos ya vistos en el histórico de datos almacenado. Se busca patrones en los datos históricos relacionándolos todos los campos (campos descriptivos) con un campo objetivo. En una primera instancia se determinan los patrones de comportamiento (esta seria la fase de aprendizaje), para después poder clasificar los nuevos datos. Dentro del aprendizaje supervisado encontramos:**

**• Clasificación y regresión: los sistemas de clasificación predicen categorías mientras los de regresión predicen números. Dentro del lenguaje python se utilizan los algoritmos de Regresión Logística (LR), que es un sistema de clasificación lineal; y los árboles de clasificación y regresión (CART), correspondientes a un sistema no lineal.**

**• K Nearest Neighbours (K-NN): el algoritmo clasifica cada dato nuevo en el grupo que corresponda, según tenga k vecinos más cerca de un grupo o de otro. Calcula la distancia del elemento nuevo a cada uno de los existentes, y ordena dichas distancias de menor a mayor para ir seleccionando el grupo al que pertenecer. Este grupo será, el de mayor frecuencia con menores distancias.**

****1.3.3.2. Algoritmos aprendizaje no supervisado****

**Los algoritmos de clustering o clasificación no supervisada son métodos donde se separan las diferentes observaciones en clases de forma que las observaciones del mismo grupo se parezcan mucho entre ellas y sean muy diferentes a las de las otras clases. Estos métodos se diferencian de la clasificación supervisada en que no hay una variable predictora, de esta forma los algoritmos no supervisados tratan los datos de entrada como variables aleatorias y construyen un modelo para este conjunto de datos.**

****1.3.3.3. Algoritmos aprendizaje por refuerzo****

**Es una técnica basada en la prueba y error. Mediante un sistema de valoración de las situaciones, “premios” y “castigos”, el sistema aprende a partir de la experiencia. Al tomar una decisión, se evalúa si es errónea, o por contra, correcta. De esta manera el modelo aprende de sus situaciones pasadas cómo debe actuar en un futuro. Es un modelo muy utilizado en la robótica y la inteligencia artificial ya que no requiere de la incorporación de una gran cantidad de datos inicialmente.**

****1.4. Machine Learning****

****1.4.1. Conepto de Machine Learning****

**El concepto de Machine Learning se ha definido de diversas maneras entre profesionales y expertos en la materia. Aquí se mostrarán dos de ellas [10]:**

**• "El aprendizaje automático en su forma más básica es la práctica de usar algoritmos para analizar datos, aprenderlos y luego hacer una determinación o predicción sobre algo en el mundo”. - Nvidia**

**• "El aprendizaje automático se basa en algoritmos que pueden aprender de los datos sin depender de la programación basada en reglas”. - McKinsey & Co.**

**Como se puede observar en estas definiciones, se habla del Machine Learning como el uso de algoritmos para permitir a las computadoras analizar datos, para posteriormente aprender de ellos de manera autónoma o semi-autónoma. Los algoritmos utilizados deben realizar parte de las acciones de aprendizaje por su cuenta, sin necesidad de ser programadas por una persona. A partir de los datos obtenidos, realizan sus propios** **cálculos. Por consiguiente, cuanto mayor sea el número de datos, más preciso será el resultado.**

**Es en este punto donde se relaciona esta ciencia, con otra mencionada anteriormente, el Big Data. La recopilación e ingesta de una gran cantidad de datos es uno de los** **principales valores actualmente en las empresas. Estos datos pueden ser utilizados para realizar predicciones que pueden colocarlas en posiciones de ventaja respecto a** **productos, mercados o clientes o competidores. Además de poder mejorar su** **organización interna y eficacia. Es por eso que el uso de esta tecnología marcará la competitividad y el desarrollo de las empresas en los próximos años.**

**La construcción de un modelo de Aprendizaje Automático incluye diferentes pasos, entre los cuáles se encuentra el uso del algoritmo de aprendizaje deseado.**

**Un modelo podría dividirse en los siguientes pasos [11]:**

**1. Recolección de datos: Los datos pueden extraerse de diferentes fuentes o, por otro lado, utilizar dispositivos que recolecten los datos.**

**2. Preprocesar los datos: Con los datos ya obtenidos, hay que asegurarse que el formato y el diseño sea el correcto para poder nutrir a nuestro el de aprendizaje.**

**3. Exploración de datos: Una vez se tienen los datos en el formato correcto, se puede realizar un pre-análisis con la idea de encontrar patrones o posibles errores en ellos. En este punto, los gráficos 2D y 3D son herramientas muy útiles.**

**4. Entrenamiento del algoritmo: En este apartado entra realmente el concepto de machine learning.**

**5. Evaluación del algoritmo: Se pone a prueba la información obtenida por el algoritmo evaluando la precisión de este en sus predicciones.**

**6. Uso final del modelo: En esta etapa, y con el rendimiento del modelo ya estudiado, se utiliza este para tratar todos los datos obtenidos. El objetivo es tratar el problema real y buscar soluciones a este.**

****1.4.2. Librerías de Python para Aprendizaje Automático****

**Las principales librerías disponibles en Python son [11]:**

**• Scikit-Learn: es la principal librería para trabajar con machine learning, incluyendo la implementación de un gran número de algoritmos de aprendizaje. Se puede utilizar para todas las fases de un modelo y se integra muy bien con el resto de paquetes científicos**

**de Python. Creada a partir de las extensiones NumPy, SciPy y matplotlib.**

**• Statsmodels: Esta librería hace foco en los modelos estadísticos y se utiliza principalmente ara análisis predictivos y exploratorios.**

**• PyMC: implementa principalmente modelos estadísticos bayesianos.**

**• NTLK: Posee numerosas herramientas para el procesamiento de texto y es la librería líder en lenguaje natural.**

**2. Diseño del programa**

**2.1. Datos necesarios para el caso de estudio**

Los datos obtenidos de la fuente, la cual seria un teléfono que disponga de conexión de datos móviles, del caso de estudio serán limitados con la idea de mostrar los parámetros más útiles para el análisis deseado en este trabajo. Los consisten en:

• El RSSI de la medición.

• La MAC de un dispositivo móvil.

• La latitud y la longitud del lugar donde se realizo esta medición.

• La fecha y la hora de la misma, además de un valor el cual no se conoce con exactitud su significado.

Para el desarrollo de este caso de estudio se asume que la información se encuentra almacenada en un archivo de tipo texto (txt) y cada parámetro está separado de los contiguos mediante comas. En la primera columna encontramos el RSSI, seguido de la MAC, la latitud, la longitud y la fecha y hora.

**7.3. Diseño y organización del modelo**

Una vez se tienen claros los datos necesarios al final del proceso de tratamiento, es necesario organizar su desarrollo y las diferentes partes que lo componen. Para ello, se enumeran los procedimientos necesarios para tratar la información mediante Python y su librería Pandas, los procesos de filtrado a realizar y la posterior obtención de la información de la forma necesaria para la posible visualización de los usuarios.

• Conversión del archivo inicial creado por la fuente de datos en txt a CSV. De esta forma, se podrán tratar los datos con el lenguaje Python.

• Eliminar las detecciones consecutivas realizadas sobre una misma MAC en un mismo instante. La idea es obtener únicamente una medida que representa su localización en ese instante determinado.

• Eliminar las medidas detectadas una única vez durante todo el recorrido.

• Suministrar la información de la posición del vehículo con la configuración necesaria para su visualización.

• Creación de un aplicación para la visualización de estos datos.

**3. Implementacion**

**3.1. Conversion de archivos**

El primer paso a realizar es la conversión del archivo para facilitar su posterior manipulación. Partiendo de un archivo txt se pasan todos los datos a un archivo separado por comas (CSV), ya que se gana en sencillez a la hora de trabajar con este tipo de archivos en Python. En este mismo paso se aprovecha para eliminar donde se encuentran la fecha y hora, ya que no es utilizable.

**3.2. Filtro inicial**

Una vez se dispone del archivo CSV se facilita su posible manipulación con Pandas. Usando este último, se procede a filtrar los datos en dos puntos diferentes.

Primeramente, se realiza una criba en las mediciones en función de su RSSI. Como se ha puntualizado en el capítulo Fuente Obtención de Datos, este valor limita el rango espacial de detección de dispositivos representando la calidad relativa de una señal recibida en un dispositivo. RSSI indica el nivel de potencia que se recibe teniendo en cuenta todos los factores que están afectando su señal en el punto de recepción. Cuanto mayor sea el valor RSSI, más fuerte será la señal. Al representar números negativos, el número más cercano al cero normalmente significa mejor señal. En este sentido, -75 sería una señal razonablemente buena y -100 sería la señal más débil que se puede percibir.

Inicialmente, se decidió eliminar registros con un RSSI menos que -90. Este se consideró

un criterio suficientemente estricto ya que la intención es eliminar el mayor número posible

de datos incorrectos, como por ejemplo las personas ubicadas en los coches cercanos al

autobús. Para verificar que este valor era adecuado en este proyecto, se realizó una

prueba con el dispositivo para comprobar la distancia aproximada en la que la señal

obtenida mostraba un RSSI similar a -90.

Los datos fueron tomados con el aparato colocado en la parte central del autobús. Las

medidas del vehículo en el que se realizaron las mediciones en Barcelona son 18 metros

de largo por 2,6 metros de ancho. En el caso de los datos tomados en Viena, la largura

del vehículo se veía reducida a 12 metros. De esta manera, es necesario que las

detecciones tengan un RSSI mayor de -90 a una distancia de unos 9 metros en caso de

ser tomadas sin obstáculos intermedios. En cambio, es necesario que sean menores de

-90 en distancias superiores a 1,5 - 2 metros si las medida es obtenida a través de una

pared de cristal o ventana. Para realizar la prueba de verificación se utilizó un smartphone

del cual se conocía su MAC con la intención de poder identificarlo fácilmente.

Primeramente se llevo a cabo una comprobación en estático, en la que el usuario con el

teléfono inteligente se iba alejando una distancia determinada en un espacio sin

obstáculos. De esta forma se obtuvo la siguiente información.

**Referencias Bibliográficas:**

[1] Haydée A., L. 2017. BIG DATA Y SU APLICACIÓN EN EL TRANSPORTE. <http://www.aacarreteras.org.ar/pdf/Lordi-BIG-DATA-AAC-VF.pdf> . Citado el [ 30/10/2022]

[2] Transforming Transport Project. 2018. <https://transformingtransport.eu/> .Citado el [ 30/10/2022]

[3] Santander, modelo español de “Smart City” - CIC Consulting Informático. 2013. <https://www.cic.es/santander-smart-city/> . Citado el [ 30/10/2022]

[4] Transporte público de Santiago utilizará sistema de Big Data a través de la red deEntel. 2014. <https://www.telesemana.com/blog/2014/07/29/transporte-publico-de-santiago-utilizara-sistema-de-big-data-a-traves-de-la-red-de-entel/> . Citado el [ 30/10/2022]

[5] Kitchin, R. 2013. The Real-Time City? Big Data and Smart Urbanism. SSRN Electronic Journal, 1-14. doi: 10.2139/ssrn.2289141 [Última consulta: 18/02/2019]

[6] Molina Félix, L. 2002. Data mining: torturando a los datos hasta que confiesen. UOC. <https://www.uoc.edu/web/esp/art/uoc/molina1102/molina1102.html> .Citado el [ 30/10/2022]

[7] Datamining. 2019. <https://www.sinnexus.com/business_intelligence/datamining.aspx> . Citado el [ 30/10/2022]

[8] Modelos de data mining y las herramientas más usadas. 2015. <https://blog.es.logicalis.com/analytics/modelos-de-data-mining-y-las-herramientas-mas-usadas> .Citado el [ 30/10/2022]

[9] González, A. 2015. De los datos a las predicciones. Un caso real de Machine Learning -. <https://cleverdata.io/datos-predicciones-machine-learning-real/> . Citado el [ 30/10/2022]

[10] Faggella, D. 2019). What is Machine Learning? | Emerj.

<https://emerj.com/ai-glossary-terms/what-is-machine-learning/> . Citado el [ 30/10/2022]

[11] González, A. 2015. De los datos a las predicciones. Un caso real de Machine Learning -. <https://cleverdata.io/datos-predicciones-machine-learning-real/> [ 30/10/2022]