PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

FACULTAD DE CIENCIAS E INGENIERÍA

FUNDAMENTOS DE DATA SCIENCE Y SUS APLICACIONES EN DISTINTAS INDUSTRIAS

Trabajo de investigación para la obtención del grado de BACHILLER EN CIENCIAS CON MENCIÓN EN INGENIERÍA INDUSTRIAL

AUTOR

Jean Franco Ramos Torres

ASESOR:

Wilmer Atoche Díaz

Lima, diciembre, 2020

Resumen

Este trabajo de investigación tiene la finalidad de brindar una guía de aprendizaje de los conocimientos, a nivel general, que un profesional debe adquirir con la finalidad de desempeñarse como Data Scientist. A través de este trabajo, se inicia enunciando lo que es Data Science y lo que hace un Data Scientist, y en base a esto discernir cinco categorías de actividades principales.

Partiendo de estas cinco actividades se desarrollan los siguientes apartados del primer capítulo, en los que se presentan los conocimientos estadísticos, matemáticos e informáticos que se deben poseer vinculados a cada una de las actividades. Aunque es de mencionar que los conocimientos asociados a estas actividades principales son transversales entre sí para una correcta aplicación del Data Science. También, se debe tener en cuenta que este trabajo solo pretende brindar una pauta para los conocimientos base necesarios para desempeñarse en el área de Data Science, esto implica que no se profundiza en temas relacionados a algoritmos de modelos, de los cuales solo se harán mención por ser relevantes por sus aplicaciones.

En el segundo capítulo se mencionan distintas aplicaciones del Data Science en cuatro industrias: servicios de salud, transporte, finanzas y e-commerce. En cada una de estos se muestran distintos casos de aplicación de Data Science entre los que están las predicciones, análisis de decisiones, detecciones de escenarios, optimizaciones, control de sistemas y sistemas de recomendaciones. En cada una de estos casos se refieren de forma concisa los procedimientos seguidos, pasando desde la recolección de los datos hasta el modelo de los mismos, y mencionando los resultados logrados.

Finalmente, se presentan conclusiones recabadas de lo que implica una formación como Data Science en la actualidad, así de como su importancia en los campos de aplicación, más ahora, en tiempos donde hay más información disponible y mejores capacidades de cómputo.

Tabla de Contenidos

Índice de Figuras

Capítulo 1. Marco conceptual

En este capítulo se detallan los conceptos y fundamentos del Data Science para comprender las implicaciones que conllevan un adecuado aprendizaje del mismo. Para esto se inicia por conocer las actividades principales que realiza un Data Scientist para posteriormente desglosar los conocimientos que son necesarios poseer para desarrollar cada una de estas actividades.

1.1. Data Science

En este apartado se busca conocer las principales actividades que realiza un Data Scientist para con ello abordar los conceptos y conocimientos necesarios para su aplicación.

1.1.1. ¿Qué es Data Science?

Skiena (2017) remarca que aún no se ha definido por completo el concepto de Data Science, pero aporta el concepto que tiene respecto al mismo y lo define como la intersección de la informática, la estadística y los dominios de aplicación real. Menciona que de la informática proviene el Machine Learning y las tecnologías informáticas para hacer frente a la escala de datos; de igual forma, la estadística aporta el análisis exploratorio de datos, pruebas de significación y técnicas de visualización; y de los dominios de aplicación real en los negocios y las ciencias surgen desafíos y estándares de evaluación para constatar cuando se ha aprendido. De forma similar, Grus (2015), concuerda que no existe aún una definición de lo que es Data Science, pero acota que surge de la intersección entre habilidades de hackeo, conocimientos de matemática y estadística, y experiencia sustancial. Entonces, se puede apreciar que, en general, Data Science involucra 3 pilares fundamentales: conocimientos informáticos, conocimientos estadísticos y matemáticos, y conocimientos del negocio al que se busca aplicar el Data Science.

1.1.2. ¿Qué hace un Data Scientist?

Aclarado los conocimientos fundamentales vinculados al Data Science, con el fin de trazar una guía de aprendizaje del mismo vinculado a las actividades que realiza un Data Scientist es importante conocer sus principales funciones. Así, Donoho (2017) clasifica las actividades en que está involucrado un Data Science en 5 divisiones, las cuales marcarán las pautas para el desarrollo de los siguientes apartados en los cuales se explicarán en más detalle cada una de estas 5 actividades principales:

- Exploración y preparación de datos: El mayor esfuerzo dedicado al Data Science se dedica sumergiéndose en los datos desordenados y adecuarlos de modo que los datos puedan estar listos para una mayor explotación.
- Representación y transformación de datos: Un Data Scientist trabaja con muchas fuentes de datos diferentes durante su carrera y estos asumen una variedad de diferentes formatos, y el Data Scientist tiene que adaptarse fácilmente a todos ellos.
- Computación con datos: Todo Data Scientist debe conocer y utilizar varios lenguajes para análisis de datos y procesamiento de datos. Estos pueden incluir lenguajes muy difundidos como R y Python, pero también se debe tener conocimiento de lenguajes específicos para transformar y manipular texto, así como datos no estructurados.
- Visualización y presentación de datos: La visualización de datos involucra el desarrollo de gráficos de análisis exploratorio de datos (EDA - Exploratory Data Analysis) tales como son los histogramas, gráficos de dispersión, gráficos de series de tiempo, diagramas de caja, entre otros. Así también, un Data Scientist crea dashboards para monitorear procesos e indicadores mediante un flujo de datos en tiempo real. Un Data Scientist también desarrolla visualizaciones para presentar conclusiones ante las gerencias.
- Modelado de datos: Un Data Scientist construye modelos predictivos a partir de los datos con

la finalidad de inferir propiedades de los datos provistos y realizar predicciones con nuevos datos. Los modelos predictivos coinciden con lo que es el Machine Learning moderno, el cual cuenta con una variedad de aplicaciones industriales.

1.1.3. ¿Qué pregunta se busca responder?

De acuerdo a Peng y Matsui (2016), para un Data Science es muy relevante y se debe enfatizar en definir la pregunta que se busca responder incluso antes de observar una base de datos para evitar gastar energías con el fin de responder la pregunta correctamente. Por su parte, Leek y Peng (2015) definen 6 tipos básicos de preguntas y concuerdan que comprender el tipo de pregunta que se está haciendo puede ser el paso más fundamental para asegurar que la interpretación de los resultados sea correcta. A continuación, se hace mención de estos 6 tipos de preguntas.

- Pregunta descriptiva: El objetivo es describir o resumir un conjunto de datos. Siempre que se tiene un nuevo conjunto de datos para examinar, este usualmente es el primer tipo de análisis que realizará. Generalmente se busca obtener un resumen simple sobre una muestra y sus medidas como pueden ser medidas de tendencia central, posición y dispersión; estos serán mencionados posteriormente.

Por ejemplo, esto se puede apreciar en el trabajo de Sabater, Molero y Pla (2010) en el que se recopila las frecuencias de visitas de menores de edad con sus familias biológicas en centros de acogida. El objetivo de esto es solo describir la distribución. No hay inferencias sobre lo que esto significa ni predicciones sobre cómo los datos podrían tener una tendencia en el futuro. Es solo para mostrarle un resumen de los datos recopilados.

- Pregunta exploratoria: El objetivo es examinar o explorar los datos y encontrar relaciones que no se conocían anteriormente; explorar cómo diferentes medidas pueden estar relacionadas entre sí, pero no confirman esa relación como causal. El hecho de que observe una relación

entre dos variables durante un análisis exploratorio no significa que una necesariamente cause la otra. Esto permite formular hipótesis e impulsar el diseño de estudios futuros y la recopilación de datos, pero un análisis exploratorio por sí solo nunca debe usarse como la última palabra sobre por qué o cómo los datos podrían estar relacionados entre sí.

Por ejemplo, esto se puede apreciar en el estudio de Esteban, Bernardo, Tuero, Cervero y Casanova (2017), en el que se analiza la influencia de las variables que influyen en el progreso académico de estudiantes universitarios y su permanencia en la universidad, y se concluye que la variable de mayor influencia es el rendimiento académico en la universidad.

- Pregunta inferencial: El objetivo es utilizar una muestra de datos para inferir algo sobre la población de la que proviene la muestra. Un análisis inferencial es comúnmente el objetivo del modelado estadístico, donde se tiene una pequeña cantidad de información para extrapolar y generalizar esa información a un grupo más grande.

Un ejemplo de esto se puede apreciar en el trabajo de Caro-Hernández y Tobar (2020), en el que se realiza un análisis microbiológico de superficies en contacto con alimentos con el fin de analizar el recuento de bacterias en locaciones de preparación y venta de alimentos para posteriormente inferir el comportamiento de la contaminación bacteriana a nivel de todo el país donde se realizó dicho estudio.

- Pregunta predictiva: El objetivo es utilizar datos actuales para hacer predicciones sobre datos futuros. Consiste en usar datos actuales e históricos para encontrar patrones y predecir la probabilidad de resultados futuros. Al igual que en el análisis exploratorio, el hecho de que una variable pueda predecir a otra no significa que una cause la otra; simplemente se hace uso de la relación observada para predecir la segunda variable.

Un ejemplo de esto se aprecia en el trabajo de Pérez, Escobar y Toledo (2017), quienes realizaron un modelo de predicción de la deserción de estudiantes de primer año en la Universidad Bernardo O'Higgins en Chile. Para ello se usaron datos históricos como son:

colegio de procedencia, resultado en el examen de admisión, ingreso familiar, lugar de residencia, edad, estado civil, entre otros. En base a este análisis se tomaron medidas preventivas en apoyo al estudiante que, según su condición de entrada a la universidad, presentan mayor riesgo de abandonar sus estudios.

- Pregunta causal: El objetivo es ver qué le sucede a una variable cuando manipulamos otra variable, observando la causa y el efecto de una relación. Generalmente, este tipo de análisis es bastante complicado de realizar solo con datos observados, pues siempre habrá incertidumbre sobre si es la correlación la que impulsa sus conclusiones o si las suposiciones inherentes al análisis son válidas.

Los ensayos controlados aleatorios de fármacos son un ejemplo de este tipo. Así, por ejemplo, un ensayo de control aleatorio examinó los efectos de un nuevo fármaco en el tratamiento de bebés con atrofia muscular espinal (Finkel, Mercuri, Darras et al. 2017). Al comparar una muestra de bebés que reciben el medicamento con una muestra que recibe un control simulado se miden varios resultados clínicos en los bebés y observan cómo el medicamento afecta los resultados.

- Pregunta mecanista: El objetivo del análisis es comprender los cambios exactos en las variables que conducen a cambios exactos en otras variables. Estos análisis son más difíciles de usar para inferir mucho, excepto en situaciones simples o en aquellas que están bien modeladas. A menudo, cuando se aplica este tipo de análisis, el único error en los datos es el error de medición, que se puede tener en cuenta.

Se puede mencionar un estudio sobre biocomposites que examina cómo el tamaño de las partículas de biocarbono, el tipo de polímero funcional y la concentración afectan las propiedades mecánicas del plástico (Behazin, Misra y Mohanty, 2017). Son capaces de realizar análisis mecanicistas mediante un cuidadoso equilibrio entre el control y la manipulación de variables con medidas muy precisas tanto de esas variables como del resultado deseado.

1.2. Exploración y Preparación de Datos

En este apartado se detallan los conceptos y procesos relacionados a la exploración y preparación de datos que se deben realizar a fin de tener un buen entendimiento de los datos y posteriormente realizar un buen modelado de datos.

1.2.1. Preparación de datos

Castanedo (2015) dice que la preparación y limpieza de datos para cualquier tipo de análisis es notoriamente costoso, lento y propenso a errores; y se convencionalmente se estima que el 80% del tiempo total dedicado al análisis se dedica a la preparación de datos. En este punto es importante considerar que las técnicas de análisis de datos y el posterior modelado de datos asumen que los datos están en un estado apropiado para realizar el análisis. Sin embargo, los datos que se recaban de las distintas fuentes muy difícilmente cumplen este supuesto pues los usualmente estos presentan errores como es la presencia de etiquetas incorrectas, datos en formato incoherente, presencia de datos faltantes, entre otros, lo cual hace necesario preparar los datos.

Por su parte, Dietrich, Heller y Yang (2015) menciona que la fase de preparación de datos es generalmente el más iterativo y el que se tienden a subestimar con más frecuencia debido a que se es presuroso por comenzar a analizar los datos, probar hipótesis y obtener respuestas a algunas de las preguntas planteadas. Sin embargo, se afirma que esta fase es crucial pues sino posteriormente se percatará que los datos con los que se está trabajando no les permite ejecutar los modelos y tendrán que realizar la preparación de datos de todas formas, pero con un desperdicio de tiempo en el proceso.

1.2.2. Identificación del tipo de datos y preparación

Entonces, Castanedo (2015) menciona que, dependiendo del tipo de los datos de entrada, puede usar diferentes métodos para prepararlos para el análisis:

• Datos de fecha y hora: Para este tipo de datos se puede usar formatos POSIX, que es un sistema para la descripción de instantes de tiempos, y almacenar el valor como el número de segundos que han pasado desde el 1 de enero de 1970 00:00:00. La ventaja de este formato es que facilita los cálculos al restar o sumar directamente los valores. Así también, se recomienda convertir las fechas a un formato estándar porque los datos de fechas se pueden describir de muchas formas diferentes.

• Datos categóricos: El trabajo de clasificar texto no procesado en variables categóricas se conoce como codificación. Esta codificación puede ser de tipo nominal, en el cual sus valores representan categorías que no obedecen a una clasificación intrínseca, o puede ser de tipo ordinal, en el cual sus valores sí representan categorías con una clasificación intrínseca. Ejemplos del primer tipo pueden ser el código postal, religión y sexo; y ejemplos del segundo tipos son nivel de instrucción o educación, grado de satisfacción y calificación en un examen en el sistema estadounidense.

• Datos de cadena o texto: Son uno de los tipos de datos más difíciles para detectar errores o inconsistencias en los valores. La mayoría de las veces, estos datos provienen de aportaciones humanas, lo que contribuye a incurrir en inconsistencias. Entre las técnicas más básicas para tratar las inconsistencias en cadenas de textos se encuentra la normalización de cadenas. En la normalización de cadenas consiste en la transformación de una cadena de texto en un conjunto común y más pequeño, en el cual se involucran dos fases: la primera consiste en encontrar un patrón en la cadena, y la segunda consiste en reemplazar un patrón por otro. Como ejemplo de esto se puede mencionar el eliminar espacios en blanco o de tabulaciones en cadenas de texto.

• Datos numéricos: Cuando se tienen datos numéricos se debe constatar que los datos sean coherentes con la variable que representa. Como primer paso se debe verificar que el subtipo de dato numérico sea correcto, identificándolo como un dato discreto o continuo, ya que esto

impactará en gran medida cuando se realice el modelado de datos. Por otro lado, se debe verificar la consistencia de datos y la presencia de errores que se detallarán en mayor extensión a continuación.

1.2.3. Verificación de posibles problemas en los datos

Por otro lado, Osborne (2013) remarca que algo que se debe tener en cuenta en todo análisis, esto es que la calidad de los resultados a los que se llegue con el análisis depende de la calidad de los datos de entrada. Por ende, la importancia de filtrar los datos y realizar una buena limpieza con el fin de lograr buenos resultados. A continuación, se presentan algunos de los problemas más comunes presentes en los datos y las alternativas para abordar sus correcciones:

- Datos faltantes o datos incompletos: McCallum (2012) menciona sobre falta de datos que esto puede ocurrir por muchas razones como pueden ser: los participantes de un estudio pueden no responder a las preguntas, mal funcionamiento del equipo y los mecanismos de recolección y/o registro de datos, los partícipes de un estudio pueden retirarse del mismo antes de que se complete en su totalidad, errores de ingreso de datos en bases de datos, entre otros. Existen distintas maneras de abordar esta problemática; aunque, por supuesto, se debe determinar cuál es la mejor solución en base al escenario en que se encuentre. Así pues, se pueden realizar las siguientes acciones:

  - La primera opción es verificar si la persona o grupo que recolectó la data puede decir cuál es el valor faltante. Esta opción no siempre es viable, dado que la persona en cuestión puede no recordar o ayudar a encontrar el valor faltante, o debido a que las fuentes de los datos no lo posibilitan.
  - Otra posibilidad es remover el valor faltante. Cuando se opta por esta posibilidad se puede eliminar la totalidad de datos correspondiente a esa variable o solamente eliminar

el registro u observación que presenta el dato faltante. Cuando no se cuenta con grandes volúmenes de datos es preferible solo eliminar registros u observaciones.

- La alternativa de reemplazar los datos es bastante más usual ante ese problema dado que no se produce perdida de registros ni variables. Sin embargo, al realizar esto se es menos preciso con la data resultante ya que implica reemplazar los datos faltantes con una suposición de cuáles deberían ser los datos. En el caso de que los datos sean numéricos, un estándar para este procedimiento es reemplazar los datos faltantes por el promedio de dicha variable. Pero en el caso en que no se puede utilizar el valor promedio de la variable, como sucede con los datos categóricos, una posibilidad es usar la moda de la variable.
- Así también, a veces es posible encontrar otra forma de adivinar los datos que faltan a través de algunas relaciones o patrones de aparición del dato faltante alguna otra variable. Así pues, se puede determinar el valor exacto o aproximado a través de correlaciones con otras variables.
- Y finalmente, en algunos casos, es posible que simplemente sea permisible dejar los datos faltantes como tales ya que por alguna u otra razón puede ser útil mantener esa observación o registro incluso si faltan datos. Esto por supuesto dependerá del tipo de análisis a realizar y se deberá tener en cuenta si se desea modelar los datos posteriormente.

• Outliers o valores atípicos: De acuerdo a Irizarry (2019), los outliers son muy comunes en el Data Science. Esto asociado con que el registro de datos puede ser complejo y con ello generar datos atípicos por error. Así, por ejemplo, esto se puede dar cuando un dispositivo de monitoreo se descompone y con ello realiza malas mediciones; así también, se puede dar como resultado del error humano cuando el registro de datos se realiza manualmente, como puede ser ingresar un dato en unidades de medición erróneas por equivocación.

Se debe poder identificar un outlier como una medición que es muy grande o muy pequeña en comparación a los otros datos. La cuestión ahora es identificar cuán grande o cuán pequeño debe ser dicho valor para ser considerado como tal. Para ello, según la estadística descriptiva se acepta como outlier a todo valor fuera del rango [Q1 – 1.5*(Q3 - Q1), Q3 + 1.5*(Q3 - Q1)] (Córdova, 2003). De acuerdo a esto, el 99.3% de los datos estarían dentro de este rango. Esto puede parecer una proporción bastante aceptable; sin embargo, considerando que en Data Science se puede trabajar con grandes volúmenes de datos, esta proporción no es muy alta. Por ello, Tukey (1977) propuso modificar el rango intercuantil tradicional bajo la siguiente formula: [Q1 – 3*(Q3 - Q1), Q3 + 3*(Q3 - Q1)]. Con esto se espera considerar como outliers solo a los valores que resulten severamente atípicos y extremos.

Osborne (2013) hace mención de cómo lidiar ante la presencia de outliers:

- Si el outlier proviene de un error humano o de un error del sistema de medición, entonces se puede tratar de corregir dicho valor de ser posible, o en su defecto se elimina la observación o registro.
- Se puede dar el caso que se presenten muchos valores muy atípicos, en esta situación se debe evaluar si es un error de muestro o de un error inherente a la recolección o fuente de los datos. En este escenario se recomienda recolectar nuevamente los datos.
- Por último, hay escenarios en que los outliers corresponden a valores que sí se pueden observar en la variable de estudio, con lo cual sería un valor extremo, mas no un valor atípico, y, por ende, no sería recomendable la eliminación del registro. Por ejemplo, esto se puede presentar si se realiza un muestreo aleatorio de salarios en una ciudad, en la cual habrá muy pocas personas que presenten salarios bastante distantes del promedio de las personas, y también muy pocas personas que ganen mucho menos salario mínimo. Entonces, los salarios de estas personas serán valores extremos, pero no son atípicos, pues sí corresponden a la población de la que fueron extraídos.

• Escalamiento de datos: La normalización de datos es una técnica importante en el pre procesamiento de datos. Singh y Singh (2019) hace énfasis en puede ser necesario realizar un escalamiento de los datos en una variable para que el rango sea consistente, y con esto se facilitan algunos análisis estadísticos posteriormente. Así pues, al hacer que los rangos sean consistentes entre las variables, el escalamiento permite una comparación justa entre las diferentes variables, asegurando que tengan el mismo impacto. También es importante para realizar un correcto modelado de datos, ya que hay algunos modelos en los cuales el escalamiento de datos desempeña un gran impacto y puede llevar a desarrollar inadecuados modelos si no se realizan. Por ejemplo, al realizar una regresión lineal, si no se realiza un escalamiento, entonces las variables que tengan valores más grandes influirán intrínsicamente más en el resultado de la regresión debido a su mayor valor, pero eso no significa necesariamente que sea más importante como predictor. Por ello, la naturaleza de los datos sesga el modelo de regresión lineal y para evitar ello es recomendable realizar un escalamiento de los datos para que las variables se encuentren en un rango de cero a uno. A continuación, se hace mención de algunos métodos de escalamiento de datos:

- El primer método, llamado “escalamiento de características simples”, divide cada valor por el valor máximo de la variable en cuestión.
- El segundo método, llamado “min-max”, toma cada valor y le resta el valor mínimo de la variable en cuestión para luego lo dividirlo por el rango de la variable.
- El tercer método se llama puntuación z o puntuación estándar. En este método, a cada valor se le resta el promedio de la variable y luego se le divide por la desviación estándar de la misma. Los valores normalmente oscilan entre tres negativos y tres positivos, pero pueden ser mayores o menores.

1.2.4. Análisis exploratorio de datos

Komorowski, Marshall, Salciccioli y Crutain (2016) dicen que el análisis exploratorio de datos (EDA) es un paso esencial en cualquier análisis de investigación y que el objetivo principal del mismo es examinar los datos en busca de distribución, valores atípicos y anomalías para posteriormente realizar pruebas específicas de su hipótesis. Así también, proporciona herramientas para la generación de hipótesis al visualizar y comprender los datos, generalmente a través de una representación gráfica. Komorowski et al. (2004) mencionan que los objetivos del EDA se pueden resumir en 5 puntos:

- Maximizar el conocimiento de los datos y comprender su estructura.
- Visualizar las posibles relaciones las variables.
- Detectar outliers.
- Desarrollar modelos parsimoniosos (modelo predictivo que usa la menor cantidad posible de variables) o una selección preliminar de modelos apropiados.
- Extraer y crear variables relevantes para un posterior modelado de datos.

Por su parte, Cleff (2013) divide el EDA respecto a la cantidad de variables que se analizan, así pues, si se analizan las variables individualmente, es llamado análisis de datos univariados; si se analiza la relación entre dos variables, se conoce como análisis bivariados; y si se analizan las relaciones entre más de dos variables se denomina análisis multivariado.

1.2.4.1. Análisis de datos univariados

Dentro del análisis de datos univariados existen diferentes mediadas y técnicas que buscan distintos objetivos, los cuales son descritos por Cleff (2013) y serán abordados a continuación.

- **Medidas de tendencia central y posición**

Córdova (2003) dice que las medidas de tendencia central ubican el centro de los datos,

como son la media aritmética, media geométrica, media armónica y la mediana. Las medidas de posición indican el lugar de los datos más frecuentes (moda) o de los menos frecuentes a partir de los cuartiles. A continuación, se realizan definiciones breves de estas medidas, las cuales fueron tomadas de Córdova (2003).

- Media aritmética: Resulta de la suma de los valores observados dividido por el número de observaciones.
- Media geométrica: Es la enésima raíz del producto de n observaciones.
- Media armónica: Es el recíproco de la media aritmética de los recíprocos del total de observaciones.
- Mediana: Es el valor observado que separa en dos partes de igual número de observaciones al total de observaciones ordenados de menor a mayor.
- Moda: Es la observación que más se repite del total de observaciones.

De las medias, la media aritmética es la que se usa con mayor frecuencia, pero no siempre es una buena medida, ya que esta afecta a la presencia de valores extremos, es decir, cuando la distribución de datos es asimétrica. Por ello, en escenarios en que se presenten estos valores es recomendable usar la mediana pues esta solo depende del número de observaciones y no del valor de estos, por lo cual no es sesgada por los valores extremos. En la Figura 1 se muestra la relación entre la media, mediana y moda, relacionado a la presencia de outliers, en el que se observa como la media aritmética es afectada por la presencia de los valores extremos.

- **Medidas de dispersión, asimetría y curtósis**

Córdova (2003) menciona que las medidas de tendencia central no son suficientes para describir un conjunto de valores de una variable pues estos solo determinan el centro, mas no brindan información acerca de cómo están distribuidos los datos respecto al centro. Por ello, se necesita una medida del grado de dispersión o variabilidad con respecto al centro con la

finalidad de ampliar la descripción de los datos o de comparar variables. Así también, se requiere de un grado de asimetría o deformación en ambos lados del centro con el cual describir la forma de la distribución de los datos. Y, por último, se requiere de una medida que permita comparar el apuntamiento o curtosis de la distribución de los datos con una distribución simétrica normal. A continuación, se realizan definiciones breves de estas medidas, las cuales fueron tomadas de Córdova (2003).

- Rango: Es una medida de dispersión que resulta de la diferencia entre el valor máximo y el valor mínimo de un conjunto de observaciones. Es un valor fácilmente calculable, pero es muy inestable ya que depende únicamente de los valores extremos y está afecta a estos valores.
- Rango intercuantil: Es una medida de dispersión que resulta de la diferencia entre el cuantil tercero y el cuartil primero (Q_3 - Q_1). Es una medida que excluye el 25% más alto y más bajo de datos, resultando en un rango del 50% de los datos centrales. Por ello, a diferencia del rango, no se encuentra afecta a la presencia de outliers.
- Varianza y desviación estándar: La varianza es una medida que cuantifica el grado de dispersión de los valores de una variable respecto a su media aritmética. Por ende, si los valores están concentrados cerca de la media, la variación será pequeña; caso contrario, si los valores están distribuidos lejos de la media, se espera una variación grande. Respecto a la desviación estándar, esta resulta de la raíz cuadrada de la varianza. Es la medida de dispersión usada más frecuentemente conjuntamente con la media aritmética, pero como sucede con esta última, no es recomendable usarla cuando se presentan valores extremos ya que estará sesgada por estos. En cuyo caso es recomendable usar el rango intercuantil.
- Índice de asimetría de Pearson: Es una medida de asimetría que resulta de la resta de la

media aritmética con la moda, y dividido por la desviación estándar. Si el resultado es cero, entonces la distribución es simétrica. Si el resultado es positivo, se considera una asimetría positiva; y lo contario si el resultado es negativo. En la Figura 1 se muestra la relación de la simetría con los valores de la media, mediana y moda.

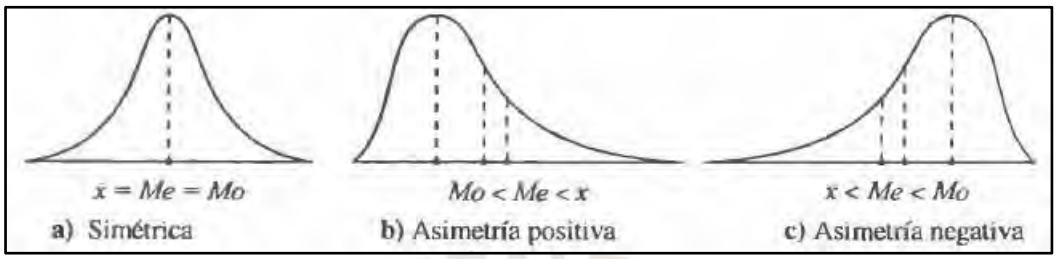

Figura 1 Relación de la media, mediana y moda

Tomado de *"Estadística: Descriptiva e Inferencial"*, por Córdova, 2003.

- **Curtosis**: Es una propiedad de la distribución de frecuencias por la cual se compara la dispersión de las observaciones cercanas al centro con la dispersión de las observaciones cercanas a ambos extremos de la distribución. La distribución se suele comparar respecto a la curva simétrica normal o mesocúrtica. Si la curtosis es mayor a la de la normal se denomina curva leptocúrtica, y si es menor se denomina platicúrtica. En la Figura 2 se muestran las características de cada una de estas curvas.

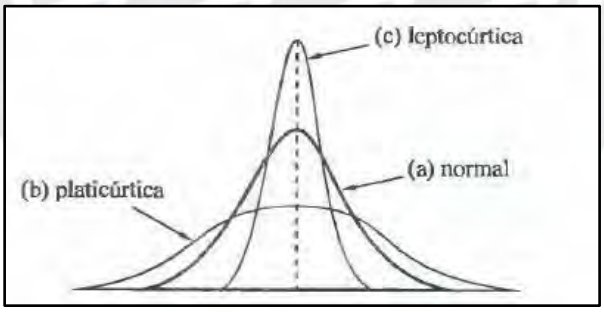

Figura 2 Tipos de curtosis

Tomado de "*Estadística: Descriptiva e Inferencial*", por Córdova, 2003.

1.2.4.2. Análisis de datos bivariados y multivariados

Dentro del análisis de datos bivariados y multivariados existen diferentes técnicas y métodos descritos por Cleff (2013) y serán abordados seguidamente.

• **Covarianza:** Levine, Krehbiel y Berenson (2012) definen la covarianza como la fortaleza de la relación lineal entre dos variables numéricas. Una distinción de la varianza con la covarianza es que esta última puede ser negativa. Así, si la covarianza es positiva, esto indica una dependencia positiva, es decir, a grandes valores de una variable corresponden grandes valores de la otra. Si la covarianza es cero, esto significa que no existe relación lineal entre las dos variables. Y si la covarianza es negativa, esto indica una dependencia negativa, es decir, a grandes valores de una variable corresponden pequeños valores de la otra. Sin embargo, la covarianza tiene un defecto importante como medida de la relación entre dos variables, y esto es debido que puede tener cualquier valor, con lo cual resulta imposible determinar la fortaleza relativa de la relación. Por ello, es necesario calcular el coeficiente de correlación.

• **Coeficiente de correlación**: Levine et al. (2012) mencionan que el coeficiente de correlación mide la fortaleza relativa de una relación lineal entre dos variables. Los valores del coeficiente de correlación van desde el +1, que implica una correlación positiva perfecta, hasta el -1, que implica una correlación negativa perfecta. En la figura 3 se muestran las relaciones de dos variables y sus coeficientes de correlación.

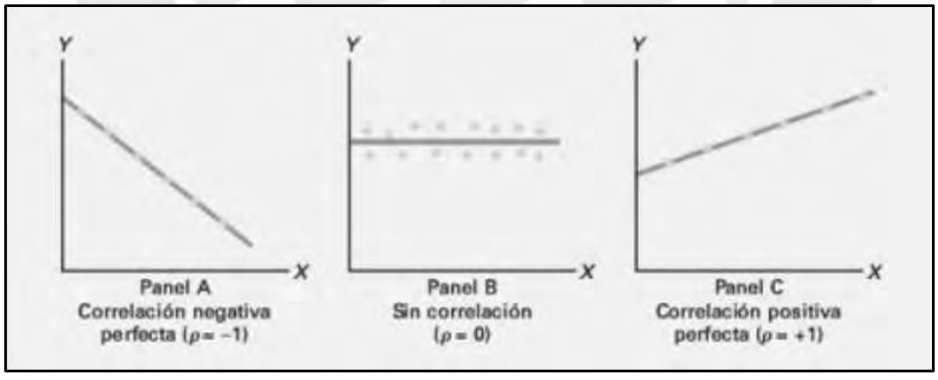

Figura 3 Tipos de correlación entre dos variables
Tomado de "*Estadística Descriptiva*", por Levine et al., 2012.

• **Pruebas de hipótesis**: Martínez (2012) menciona que las pruebas de hipótesis tienen como objetivo evaluar suposiciones acerca de los valores estadísticos de poblaciones, denominados parámetros. También se puede definir como la afirmación acerca de una característica ideal de

una población sobre la que hay incertidumbre al momento de formularla, y que es expresada de tal forma que pueda ser rechazada. La hipótesis nula (H_0) representa el supuesto que es aceptado temporalmente como verdadera y cuya veracidad será comprobada. La hipótesis alternativa (H_1) representa a la hipótesis contraria a la hipótesis nula que se acepta en caso que la hipótesis nula sea rechazada. En la decisión de aceptar o rechazar la hipótesis nula se pueden cometer dos tipos de errores:

- Error tipo I: Consiste en rechazar la hipótesis nula cuando es en realidad verdadera. También es conocido como un falso positivo. La probabilidad de cometer este error se denota por la letra α, la cual también representa el nivel de significancia de la prueba de hipótesis.
- Error tipo II: Consiste en aceptar la hipótesis nula cuando es en realidad falsa. También es conocido como un falso negativo. La probabilidad de cometer este error se denota por la letra β, y a diferencia del error tipo I que es controlado, esta probabilidad de error depende de la diferencia que existe entre los valores hipotéticos y el valor real del parámetro poblacional.

Es importante mencionar que las pruebas de hipótesis también pueden ser del tipo análisis univariado, pero dado que está presente en ambos escenarios se optó por presentarlo en este apartado.

	Situación real	
Decisión estadística	**H_0 verdadera**	**H_0 falsa**
No se rechaza H_0	Decisión correcta Confianza = $1 - \alpha$	Error tipo II P(error tipo II) = β
Se rechaza H_0	Error tipo I P(error tipo I) = α	Decisión correcta Potencia = $1 - \beta$

Figura 4 Tipos de errores en pruebas de hipótesis

Tomado de "*Estadística Descriptiva*", por Levine et al., 2012.

• **ANOVA**: Selvamuthu y Das (2018) mencionan que la técnica ANOVA o análisis de varianza es una herramienta que sirve para el estudio del efecto de uno o más factores o variables (cada uno con dos o más niveles) sobre la media aritmética de una variable en estudio. El análisis de varianza puede relacionarse con las pruebas hipótesis en el sentido que prueba la hipótesis que las medias de dos o más poblaciones son iguales, pero con la diferencia que los ANOVA evalúan la importancia de uno o más factores al comparar las medias de la variable en estudio en los diferentes niveles de los factores. Esta técnica puede generalizarse también para estudiar los posibles efectos de los factores sobre la varianza en una variable.

Llegado la conclusión de este apartado, es importante mencionar que dentro del análisis exploratorio de datos también se utilizan gráficas estadísticas, tanto en el análisis univariado como en el bivariado y multivariado, pero se ha considerado abordar las gráficas y visualizaciones en el apartado cinco del presente capítulo, del cual también forman parte.

1.3. Representación y transformación de datos

Un Data Scientist encontrará que un paso central en su trabajo es implementar una transformación apropiada para reestructurar los datos dados originalmente en una forma nueva que conlleven una mejor representación. En ese sentido, en el presente apartado se recapitulan algunas transformaciones matemáticas útiles cuando se trabajan con datos de tipos especiales.

1.3.1. Transformada Wavelet

Stark (2005) menciona que una Wavelet es una onda corta de duración limitada, es decir, su energía se encuentra concentrada alrededor de un punto en el tiempo, lo cual proporciona una herramienta útil para el análisis de fenómenos transitorios, no estacionarios y variables en el tiempo. Por su parte, Kouro y Musalem (2002) mencionan que la Transformada Wavalet es una técnica mediante ventanas con regiones de tamaño variable que permite el uso de intervalos grandes de tiempo en segmentos en los que se requiere mayor precisión en baja frecuencia, y regiones más pequeñas donde se requiere información en alta frecuencia. Así pues, dentro de las aplicaciones de la Transformada Wavelet se tienen:

- Detección de discontinuidad en señales.
- Eliminación de ruido.
- Compresión de imágenes.
- Multiplicación rápida de matrices.
- Identificación de frecuencias puras

Respecto a su uso en Data Science se puede mencionar el uso de la Transfromada Wavelet cuando se trabajan con datos vinculados a imágenes y se busca representar el ruido presente las imágenes. Un ejemplo de esta aplicación la realizó Resendiz (2017), en su estudio de filtros Wavelet óptimos para la detección de ruidos en imágenes, en el cual hace posible la representación gráfica del ruido presente en las imágenes y con ello desarrollar filtros que

mitiguen la presencia del ruido y la distorsión de la imagen. En la Figura 5 se puede apreciar la representación del ruido en una imagen mediante Transformada Wavelet.

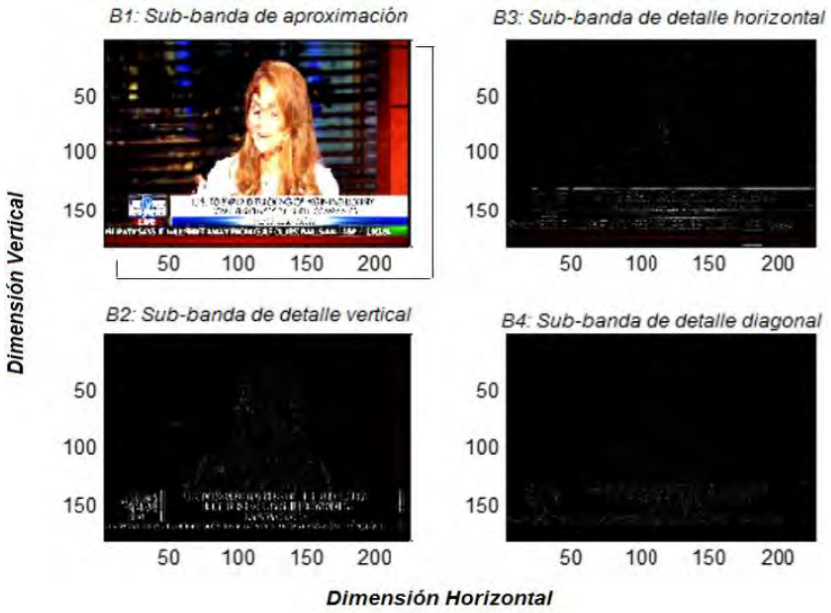

Figura 5 Aplicación de Transformada Wavelet en representación de datos de imágenes
Tomado de "*Filtros Wavelet óptimos para detección de ruido en imágenes SD y HD*", por Resendiz, 2017.

1.3.2. Transformada de Fourier

De acuerdo a Serov (2017), la Transformada de Fourier es una transformación matemática empleada para transformar señales entre el dominio continuo del tiempo y de la frecuencia, es decir, que toma infinitos valores en un intervalo finito de tiempo y de frecuencia. Resulta ser una herramienta sumamente potente, ya que proporciona métodos para la resolución de ecuaciones difíciles de manejar, como son las respuestas dinámicas de sistemas eléctricos, lumínicos y térmicos. Entonces, en términos generales, la Transformada de Fourier consigue realizar un cambio de dominio; es decir, el paso de información contenida en una señal de dominio temporal o espacial, al de la frecuencia y viceversa, de tal modo que permite mejorar el análisis de dicha señal. Así pues, dentro de las aplicaciones de la Transformada de Fourier se tienen:

- Reforzar señales.
- Generar formas de onda de corriente por medio de superposición de senoides.
- Analizar el contenido de frecuencia de las señales.
- Diseñar sistemas de transmisión de señales para transmitir información.
- Balanceo de motores y eliminación de vibración producto del desbalanceo.

Respecto a su uso en Data Science se puede mencionar el uso de la Transformada de Fourier para realizar un monitoreo visual de procesos que vinculan señales acústicas. Un ejemplo de esta aplicación la realizó Blasina (2019), en la que evaluó la atenuación de frecuencias cuando las señales acústicas incidían en distintos materiales mediante experimentos de atenuación de ultrasonido. En la Figura 6 se observa la representación de la evolución de las señales acústicas en distintos materiales mediante Transformadas de Fourier.

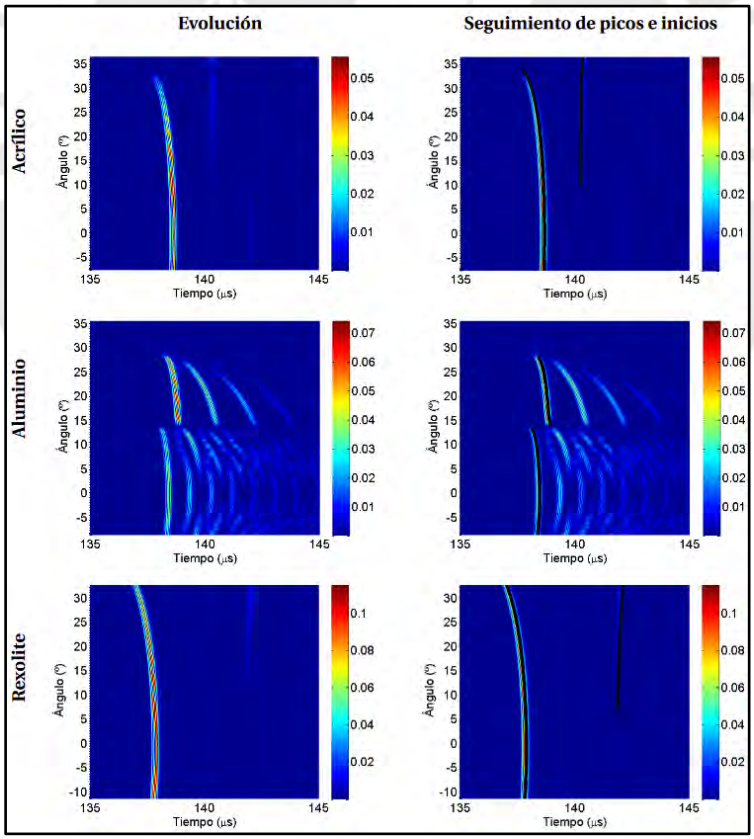

Figura 6 Aplicación de Transformada de Fourier en representación de datos acústicos

Tomado de "*Procesamiento de señales acústicas aplicado al monitoreo de procesos*", por Blasina, 2019.

1.3.3. Representación Piramidal

De acuerdo a Brito y Diday (1990), la Representación Piramidal es una representación multiescala diseñada para el procesamiento de señales y procesamiento de imágenes, en la cual la señal o imagen es suavizada y sometida a submuestreos repetidos. El modelo piramidal viene dado por el nivel de la multiescala; ya que, por ejemplo, se considera una imagen a diferentes niveles de resolución representada como una imagen. Dentro de sus aplicaciones se consideran las siguientes:

- Eliminar ruidos.
- Analizar texturas.
- Reconocimiento de objetos.
- Etiquetado de características de imágenes

A continuación, se muestra la Figura 7, que representa un ejemplo de detección de objetos mediante Representación Piramidal a través de segmentación basada en color (Bustamante, 2014).

Figura 7 Aplicación de Representación Piramidal en representación de datos de imágenes
Tomado de "*Algoritmos de procesamiento de imagen aplicados a la detección de figuras geométricas y sus propiedades espaciales*", por Bustamante, 2014.

1.4. Computación con datos

En teoría, todo lenguaje de programación suficientemente potente es capaz de expresar cualquier algoritmo. Pero en la práctica, ciertos lenguajes de programación resultan mucho mejores que otros en tareas específicas. Sucede lo mismo también con los programas usados en Data Science, de los cuales hay una gran variedad, pero cada uno de ellos tiene ventajas sobre los otros dependiendo de la aplicación. Es por eso que en el presente apartado se detallarán los lenguajes de programación y programas de mayor uso es Data Science.

1.4.1. Data Wrangling

Data Wrangling es un término usado en Data Science que hace referencia al proceso general de manipulación de datos desestructurados o desordenados con la finalidad de transformarlos a una forma estructurada o limpia para usarlos en un posterior análisis (McKinney, 2017). Dentro de los principales lenguajes de programación usados en Data Science se tienen:

- Python: De acuerdo a VanderPlas (2016), es uno de los lenguajes de programación más usados para Data Science. Python contiene una variedad de características para facilitar la manipulación de datos, como por ejemplo las expresiones regulares. Es un lenguaje de programación interpretado, lo que hace que el proceso de desarrollo sea más rápido y agradable. Así también, Python cuenta con una gran comunidad que se ha encargado de desarrollar bibliotecas para distintos propósitos. Una biblioteca se debe entender como un conjunto de herramientas (llamadas funciones) que hacen tareas en nuestros datos. Dentro de estas bibliotecas se tienen las siguientes recabadas de Albon (2018).

  - Numpy: Su característica principal es que permite trabajar con matrices o arrays de n dimensiones. También ofrece funciones básicas de algebra lineal, transformada de Fourier, generación de números aleatorios, entre otros.

- Pandas: Es una de las bibliotecas más útiles en Data Science ya que permite la estructuración de datos en una dimensión (llamada Series) o en dos dimensiones (llamada DataFrame). Es bastante usado para la preparación y manipulación de datos, ya que proporciona herramientas para el análisis de datos.
- SciPy: Desarrollada a partir de NumPy. Es una de las más útiles por la gran variedad de funciones de alto nivel que posee sobre ciencia e ingeniería, como transformada discreta de Fourier, álgebra lineal, interpolación, integración numérica y optimización.
- Matplotlib: Es la biblioteca gráfica estándar de Python y la más conocida. Con matplotlib se puede crear diferentes tipos de gráficos: series temporales, histogramas, diagramas de barras, diagramas de errores, gráficas de dispersión, entre otros.
- Seaborn: Es una biblioteca gráfica basada en matplotlib, especializada en la visualización de datos estadísticos. Se caracteriza por ofrecer un interfaz de alto nivel para crear gráficos estadísticos visualmente atractivos e informativos.
- Statsmodels: Es una biblioteca bastante útil para el modelado estadístico. Permite explorar datos, hacer estimaciones de modelos estadísticos y realizar test estadísticos.
- Scikit Learn: Es una de las bibliotecas más usadas para Machine Learning en Python. Contiene una gran variedad de herramientas y modelos, tanto de aprendizaje supervisado, como de no supervisado, entre los que tenemos: regresiones, árboles de decisión, k-nearest neighbors, redes neuronales, support vector machines, entre otros.

La mayor desventaja de Python es la eficiencia, esto referido a su velocidad, debido a que es un lenguaje interpretado y por ello es inferior en velocidad frente a los lenguajes compilados. Esta es una de las razones por las que Python no suele ser usado cuando la velocidad es un aspecto importante en un proyecto (VanderPlas, 2016). En la Figura 8 se observa la tendencia creciente de usuarios que utilizan Python, ello debido a su versatilidad y facilidad de aprendizaje.

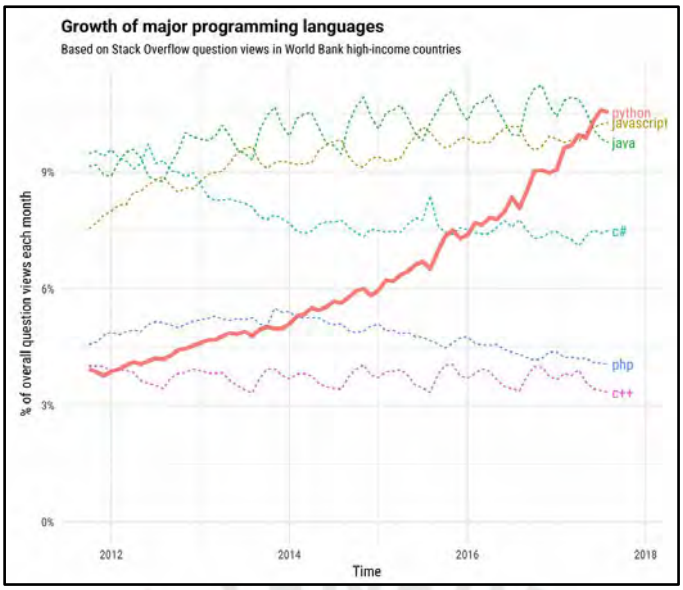

Figura 8 Crecimiento del uso de Python frente a otros lenguajes de programación
Tomado de Stack Overflow (2017) The Incredible Growth of Python.

• R: Es lenguaje de programación con un enfoque estadístico que, como Python, ha sido ampliamente adoptado en el sector de Data Science en los últimos años. Los profesionales que prefieren R generalmente lo hacen debido a su programación estadística avanzada y sus capacidades de visualización de datos, capacidades que simplemente no se pueden replicar en Python. Por ello, cuando se trata de Data Scientist específicamente, la base de usuarios de R es más amplia que la de Python (Skiena, 2017).

R no es tan fácil de aprender como Python, pero R puede ser más poderoso para ciertos tipos de análisis estadísticos avanzados. Aunque el Python es considerado el lenguaje de más rápido aprendizaje, R también es relativamente sencillo de aprender. R también cuenta con una gama de bibliotecas, algunas de las cuales se mencionan a continuación recabadas de Pimpler (2018) y Mailund (2017).

- Dplyr: Se utiliza principalmente para la manipulación de datos en R y se basa en cinco funciones: seleccionar ciertas columnas de datos, filtrar datos, organizar filas de datos, realizar trasformaciones a los datos, y resumir los datos.
- Tidyr: Se utiliza para transformar los datos sucios en datos limpios y ordenados. Para

esto se basa en tres reglas: cada variable debe tener su propia columna, cada observación debe tener su propia fila y cada valor debe tener su propia entrada.

- Ggplot2: Es una de las mejores bibliotecas para visualización de datos en R. Permite producir visualizaciones gráficas de alta calidad expresando relaciones entre los atributos de los datos y su representación gráfica.
- Esquisse: Desarrollada sobre ggplot2. Es una herramienta de visualización más simple y directa que lleva las mejores características de Tableau a R, y permite la exploración sus datos de forma interactiva visualizándolos con ggplot2.
- MLR: Es la biblioteca de R para Machine Learning. Ofrece modelos de aprendizaje supervisados como clasificación, regresión y análisis de supervivencia, así como métodos no supervisados de clustering.
- Shiny: Es una biblioteca usada para el desarrollo de aplicaciones web interactivas directamente desde R. También permite crear dashboards.
- Lubridate: Usado para agilizar y facilitar el manejo de fechas, horas y períodos de tiempo. Esta biblioteca permite expandir el tipo de operaciones matemáticas que se puede realizar con objetos de fecha y hora.

De acuerdo a Sahinaslan (2019), los lenguajes de programación más populares para Data Science en la actualidad son Python, R y SQL (este último es mencionado en el siguiente apartado). Aparte de los mencionados anteriormente, para el data wrangling se pueden incluir otros lenguajes de programación que también son usados para este fin, aunque con menor cantidad de usuarios a comparación de los que usan Python y R. Dentro de estos otros lenguajes se incluyen; Julia, Scala, MatLab, Java y C++. Estos dos últimos convencionales usados para el desarrollo de grandes sistemas en aplicaciones de Big Data, dado que aportan mayor velocidad de procesamiento en grandes volúmenes de datos por ser lenguajes de programación compilados.

1.4.2. Database Management

Los datos estructurados se pueden almacenar, procesar y manipular en un sistema de gestión de bases de datos relacionales tradicional (RDBMS - Relational Database Management System). Estos datos pueden ser generados por personas o máquinas, como pueden ser: encuestas, reportes de ventas, mediciones de sensores, entre otros. Por su parte, los datos no estructurados vienen completamente desestructurados; comúnmente generados a partir de actividades humanas y no encajan en un formato de base de datos estructurado. Dichos datos pueden generarse de publicaciones de blogs, correos electrónicos, redes sociales, documentos de texto, entre otros (Pierson, 2017).

Entonces, Database Management consiste esencialmente en un grupo de programas que pueden editar, indexar y manipular la base de datos. A continuación, se hace mención de los cuatro RDBMS más usados en la actualidad, los cuales tienen en común que usan el lenguaje de consulta SQL. Estos presentan cambios de sintaxis menores entre sí, pero la sintaxis SQL básica sigue siendo en gran medida la misma (Taylor, 2013).

- MySQL: Es un RDBMS multiusuario y de código abierto utilizado con bastante frecuencia en el desarrollo de páginas web actuales. Además, es el más usado en aplicaciones creadas como software libre. Tiene como ventaja su soporte multiplataforma, pero el lado negativo es su escalabilidad, es decir, que es poco eficiente con bases de datos grandes.
- PostgreSQL: Es un RDBMS orientado a objetos y de código abierto. Dentro de sus características están su robustez, eficiencia y estabilidad para trabajar con bases de datos grandes, pero como desventaja presenta lentitud para gestionar bases de datos pequeñas ya que se encuentra optimizado para gestionar grandes volúmenes de datos.
- Microsoft SQL Server: Es el RDBMS desarrollado por Microsoft, y tiene como atributos su capacidad de poner a disposición de muchos usuarios grandes cantidades de datos de manera simultánea, así como integrar la opción de cancelar consultas. Tiene como desventaja su precio,

ya que no es de código abierto.

• Oracle: Considerado el RDBMS más utilizado en el rubro empresarial por ser el más completo y robusto de los cuatro. Destaca por su escalabilidad, estabilidad y soporte multiplataforma. Al igual que Microsoft SQL Server, tiene como desventaja de su precio pues no es de código abierto.

ClientName	*Address1*	*Address2*	*City*	*State*
Butternut Animal Clinic	5 Butternut Lane		Hudson	NH
Amber Veterinary, Inc.	470 Kolvir Circle		Amber	MI
Vets R Us	2300 Geoffrey Road	Suite 230	Anaheim	CA
Doggie Doctor	32 Terry Terrace		Nutley	NJ
The Equestrian Center	Veterinary	7890 Paddock Parkway	Gallup	NM
Dolphin Institute	1002 Marine Drive		Key West	FL
J. C. Campbell, Credit Vet	2500 Main Street		Los Angeles	CA
Wenger's Worm Farm	15 Bait Boulevard		Sedona	AZ

Figura 9 Ejemplo de base de datos estructurada
Tomado de "*SQL for Dummies*", por Taylor, 2013.

Seguidamente, ahora se detallarán sobre los sistemas de gestión de bases de datos no estructurados. Pare ello, se detallarán los cuatro tipos de bases de datos NoSQL (denominados así por no ser relacionales) existentes y los DBMS NoSQL asociados a estos (Herranz, 2014).

• Bases de datos clave-valor: Este tipo de base de datos posee la estructura de datos más sencilla, lo cual proporciona ventajas a la hora de particionar y escalar el sistema a lo largo de clústers de decenas e incluso cientos de nodos. Por lo cual, este tipo de base es altamente divisible y permite escalado horizontal a escalas que otros tipos de bases de datos no pueden alcanzar. Los DBMS NoSQL que soportan este tipo de base de datos son: Amazon DynamoDB, Redis, Cassandra, Riak y Oracle NoSQL Database.

Tabla 6. Ejemplo de tabla en DynamoDB[45]

ID (clave primaria)	Atributos
101	{ Title = "Book 101 Title" ISBN = "111-1111111111" Authors = "Author 1" Price = -2 Dimensions = "8.5 x 11.0 x 0.5" PageCount = 500 InPublication = 1 ProductCategory = "Book" }
201	{ Title = "18-Bicycle 201" Description = "201 description" BicycleType = "Road" Brand = "Brand-Company A" Price = 100 Gender = "M" Color = ["Red", "Black"] ProductCategory = "Bike" }

Figura 10 Ejemplo de base de datos no estructurada clave-valor

Tomado de "*Bases de Datos NoSQL: Arquitectura y Ejemplos de Aplicación*", por Herranz, 2014.

• Base de datos en columna: Este tipo de base de datos almacena los datos en familia de columnas en vez de registros. Destaca por ser más amigable al programador, ser orientados a la web, y favorables para consultas y agregaciones de grandes de datos. Los DBMS NoSQL que soportan este tipo de base de datos son: Cassandra, Apache Hbase y Google Cloud Bigtable.

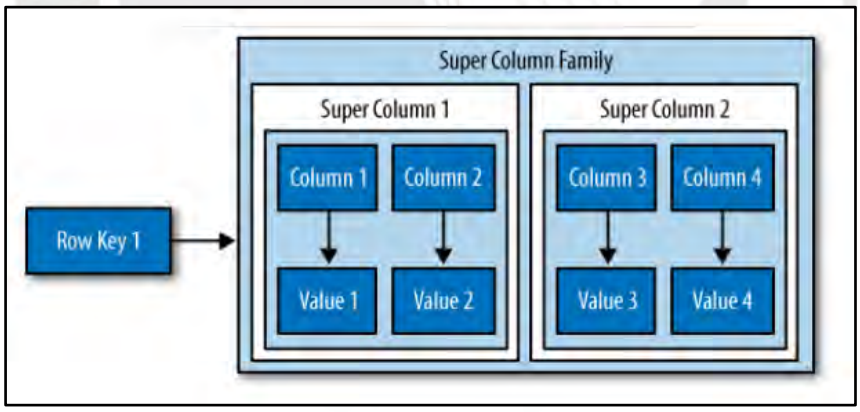

Figura 11 Ejemplo de base de datos no estructurada en columna

Tomado de "*Bases de Datos NoSQL: Arquitectura y Ejemplos de Aplicación*", por Herranz, 2014.

• Base de datos por documentos: Este tipo de base de datos almacena los datos en documentos en forma de datos semi estructurados. Destaca por ser amigable para el programador, ser orientados a la web, y permitir el modelado natural de datos. Los DBMS NoSQL que soportan este tipo de base de datos son: MongoDB y CouchDB.

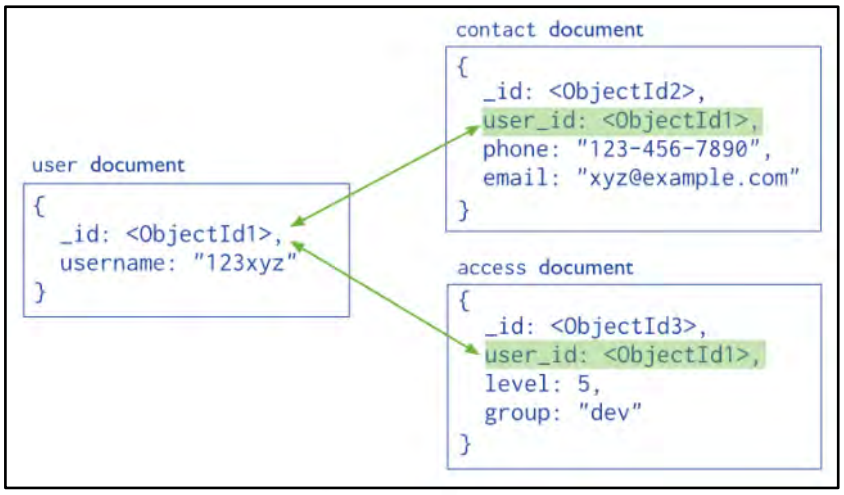

Figura 12 Ejemplo de base de datos no estructurada por documentos

Tomado de "*Bases de Datos NoSQL: Arquitectura y Ejemplos de Aplicación*", por Herranz, 2014.

• Base de datos orientado a grafos: Este tipo de base de datos almacena los datos mediante una representación de relaciones a través de nodos y aristas. Por eso, permite representar y entender conjuntos de datos altamente complejos y conectados. Así también, presenta un alto rendimiento cuando los datos están interconectados y no son tabulares. Los DBMS NoSQL que soportan este tipo de base de datos son: Neo4j, OrientDB, DataStax Enterprise Graph y HyperGraphDB.

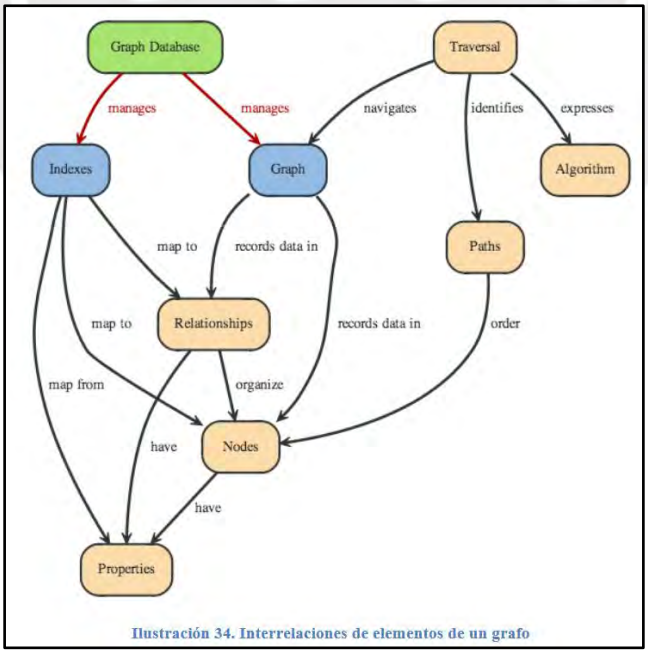

Figura 13 Ejemplo de base de datos no estructurada orientada a grafos

Tomado de "*Bases de Datos NoSQL: Arquitectura y Ejemplos de Aplicación*", por Herranz, 2014.

Finalmente, para concluir este apartado, a continuación, se presenta la Figura 14 en el que se muestra un ranking de popularidad de DBMSs en el que se observa que los DBMS NoSQL presentan una tendencia de uso creciente con los años y se espera que en el futuro supere el uso de las RDBMS, esto debido a la relevancia que cada vez más tienen los datos no estructurados en el Data Science.

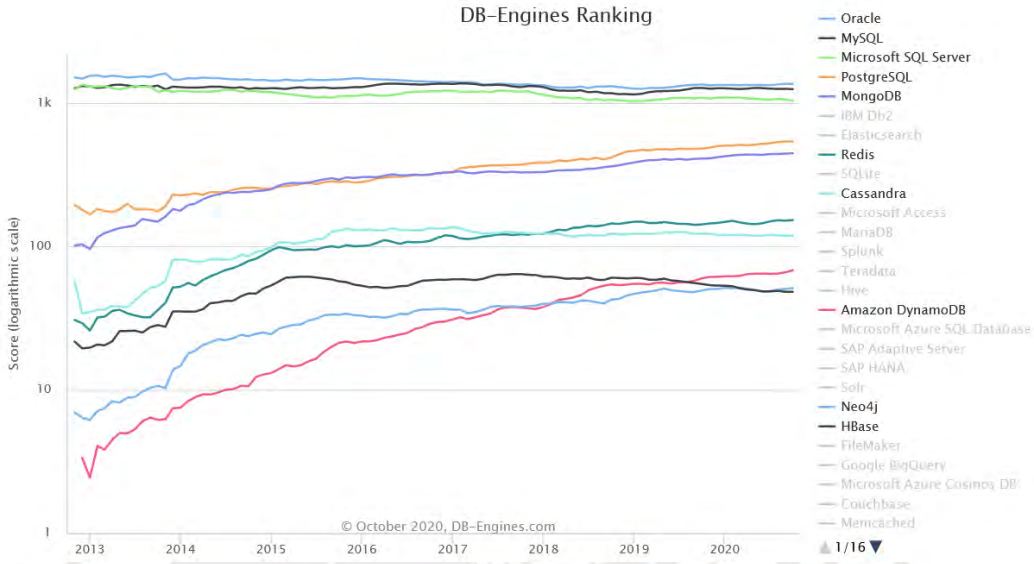

Figura 14 Ranking de popularidad de DBMSs

Tomado de DB-Engines (2020) DB-Engines Ranking - Trend Popularity.

1.4.3. Data Visualization

Gartner (2020), en un informe de investigación de mercado, presenta la tendencia del mercado en el uso de herramientas de visualización, hasta el presente año, en el rubro de inteligencia de negocios y analítica de negocios. En la Figura 15 se observa que las plataformas líderes y dominantes en visualización son Power BI (plataforma de Microsoft) y Tableau. Estos se caracterizan por proveer interfaces para la generación de reportes, dashboards y análisis de grandes volúmenes datos.

Así también, hay otras plataformas que se están posicionando cada vez más en el mercado como es el caso de QlikView, que también lidera el mercado. Esta es predominantemente una herramienta de descubrimiento de datos y, por lo tanto, tiene algunas características distintas

de visualización de datos como es la búsqueda impulsada de patrones y tendencias en conjuntos de datos.

Figura 15 Plataformas líderes en visualización de datos
Tomado de Gartner (2020) Gartner Magic Quadrant for Analytics and Business Intelligence Platforms.

Finalmente, no se debe olvidar que los lenguajes de programación mencionados anteriormente como Python y R también incluyen bibliotecas destinadas a visualización que son usados para realizar análisis exploratorio de datos (EDA). Estas tienen la ventaja que se pueden controlar los elementos específicos de los gráficos que se crean y hacer que esas especificaciones sean repetibles a través del código. Como desventaja de los lenguajes de programación frente a las plataformas presentes en este apartado se puede mencionar que estas últimas están más orientadas a la generación de reportes y dashboards, y por ello son bastante usados en la presentación de resultados ante gerencias o en el control de procesos con flujo de datos en tiempo real.

1.5. Visualización y presentación de datos

Como se ha mencionado en los anteriores apartados, dentro de las actividades de un Data Scientist se involucra la visualización de datos. Esta actividad se puede dividir en base a su propósito: el análisis de datos y la presentación de resultados o conclusiones (Donoho, 2017). Entonces, en el presente apartado se discuten los aspectos que involucran el desarrollo de estas dos categorías de visualizaciones.

1.5.1. Análisis exploratorio de datos

En este apartado se hará mención de los distintos de gráficos que se pueden utilizar para el Análisis exploratorio de datos (EDA). Esto servirá para complementar lo mencionado en el apartado dos de este capítulo, en el que se mencionó que el EDA es frecuentemente acompañado de visualizaciones, ya que brindan un mejor entendimiento de los datos de forma gráfica.

A continuación, se hará mención de unas breves definiciones y características de los tipos de gráficos más usados en el EDA, los cuales fueron tomadas de VanderPlas (2016).

- Gráficos de puntos y gráficos de líneas: Son los gráficos más simples que existen. El gráfico de puntos solo muestra una serie de puntos de datos, y en el caso del gráfico de líneas se conectan los puntos con líneas rectas. Estos gráficos son útiles, por ejemplo, si se quiere ver el comportamiento de una variable en el tiempo.

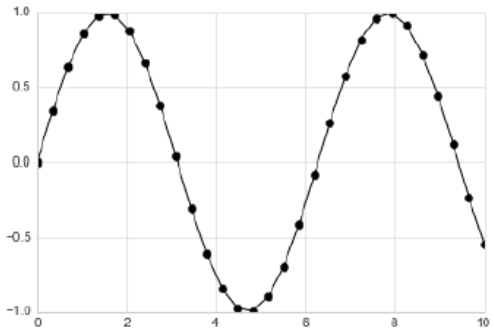

Figura 16 Ejemplo de gráfico de puntos y de gráfico de líneas
Tomado de "*Python Data Science Handbook: Essential Tools for Working with Data*", por VanderPlas, 2016.

• Gráfico de Barras: Es comúnmente usado para comparar los valores de una variable cuantitativa en un punto fijo en el tiempo, o para comparar variables categóricas.

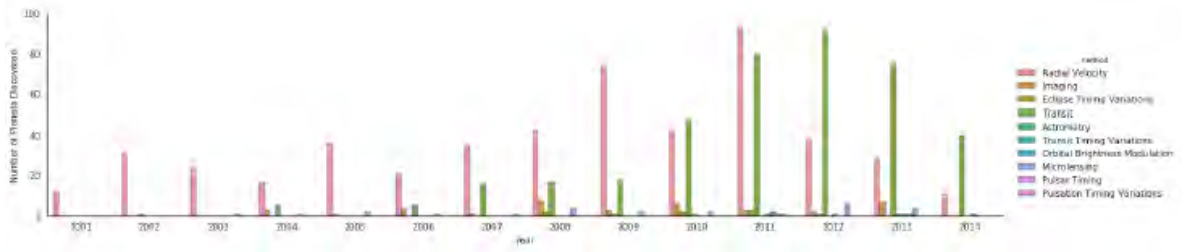

Figura 17 Ejemplo de gráfico de barras

Tomado de "*Python Data Science Handbook: Essential Tools for Working with Data*", por VanderPlas, 2016.

• Histograma: Permite representar la distribución de frecuencias de una variable cuantitativa. Brinda información visual acerca de las medidas de tendencia central, dispersión y asimetría.

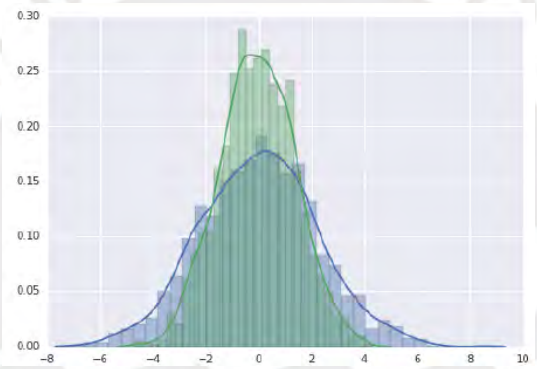

Figura 18 Ejemplo de histograma

Tomado de "*Python Data Science Handbook: Essential Tools for Working with Data*", por VanderPlas, 2016.

• Diagrama de caja: Es una forma de representar estadísticamente la distribución de datos. En esta se muestran: el rango intercuantil, el rango sin considerar outliers (permite identificar valores atípicos), los cuantiles (incluye la mediana), y los valores mínimos y máximos.

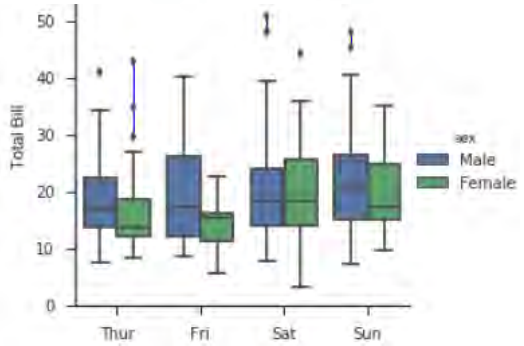

Figura 19 Ejemplo de diagrama de caja

Tomado de "*Python Data Science Handbook: Essential Tools for Working with Data*", por VanderPlas, 2016.

• Mapa de calor: Muestra la visualización de datos con tres dimensiones en un gráfico bidimensional. Para esto hace uso de la variación del color por matiz o intensidad para representar una tercera dimensión.

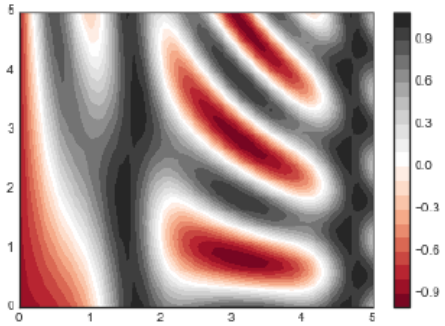

Figura 20 Ejemplo de mapa de calor
Tomado de "*Python Data Science Handbook: Essential Tools for Working with Data*", por VanderPlas, 2016.

• Histograma de dos dimensiones: Permite visualizar mejor la relación de dos variables, mostrando sus respectivos histogramas y el gráfico de puntos.

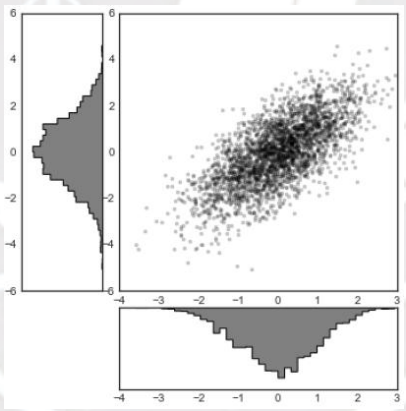

Figura 21 Ejemplo de histograma de dos dimensiones
Tomado de "*Python Data Science Handbook: Essential Tools for Working with Data*", por VanderPlas, 2016.

• Gráfico de pares: Permite visualizar relaciones por pares en un conjunto de datos. Es bastante útil para verificar la correlación entre variables cuando las variables son muchas.

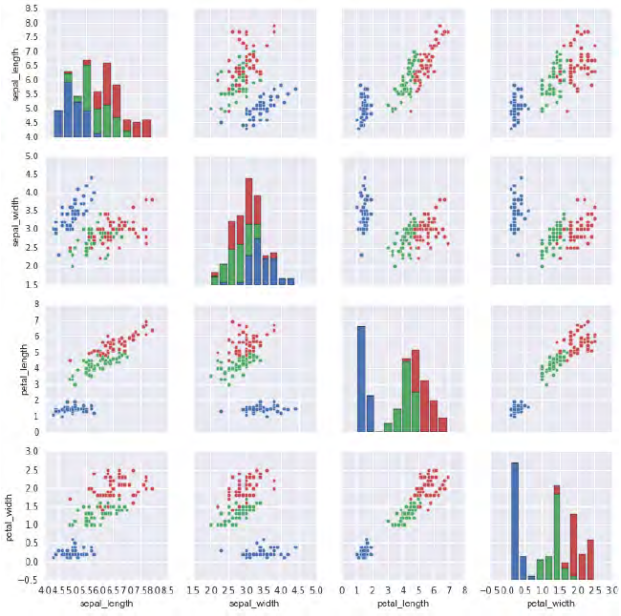

Figura 22 Ejemplo de gráfico de pares

Tomado de "*Python Data Science Handbook: Essential Tools for Working with Data*", por VanderPlas, 2016.

1.5.2. Presentación de datos

Cairo (2012) proporciona una herramienta para pensar en los "sacrificios" o compensaciones de diseño al construir gráficos de información, y llama a esta herramienta la rueda de visualización, que se aprecia en la Figura 23. En esta se aprecia que hay dos polos. El superior representa datos muy complejos que informan a un nivel profundo, mientras que el inferior proporciona un acceso más fácil a los datos, pero solo informa de manera superficial.

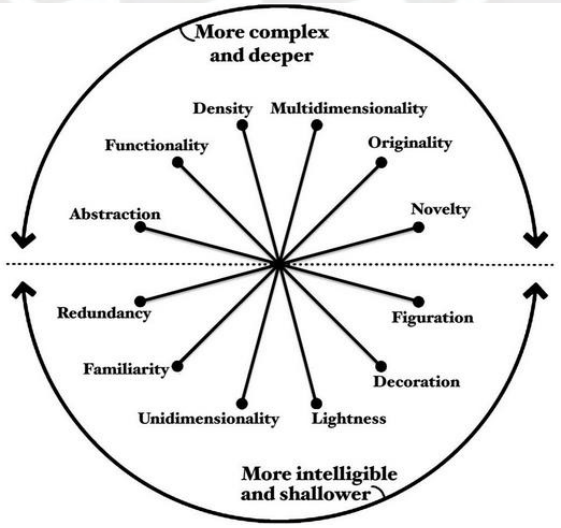

Figura 23 Rueda de visualización de Alberto Cairo

Tomado de "*The Functional Art: An introduction to information graphics and visualization*", por Cairo, 2012.

También se observa que dentro del círculo hay seis dimensiones que describen las compensaciones entre los dos polos. Estas dimensiones serán comentadas a continuación:

- Abstracción – Figuración: El nivel de figuración en un gráfico depende del fenómeno que se representa mediante representaciones físicas, como fotografías o dibujos. A medida que las representaciones pasan de fenómenos reales a menos reales y más conceptuales, el énfasis pasa de la figuración a la abstracción.
- Funcionalidad – Decoración: Un gráfico completamente funcional no tiene decoraciones y es probable que sea una representación directa de los datos. Por otro lado, un gráfico muy decorado presenta más ornamentos. Estas decoraciones pueden aumentar la cantidad de tiempo que un espectador dedica a considerar lo visual; pero, a veces, esto ayuda a formar asociaciones mentales, lo que aumenta la familiaridad y la memorización.
- Densidad – Livianidad: Se relaciona con la cantidad de información que se muestra. Esta dimensión se relaciona con la audiencia a la que se muestre en el gráfico. Así, si la audiencia necesita un entendimiento en profundidad de la información, la densidad es importante. Por otro lado, si la audiencia solo necesita información general, es preferible la livianidad.
- Multidimensional – Unidimensional: Un gráfico multidimensional explica un fenómeno considerando un mayor número de dimensiones y esto permite a la audiencia explorar diferentes aspectos. Por otro lado, un gráfico unidimensional se centra en una sola o pocas dimensiones con el fin de explorar el fenómeno de manera focalizada.
- Originalidad – Familiaridad: Relacionado con el tipo de gráfico a usar para mostrar la información. Se debe evaluar cuál es el mejor gráfico a usar teniendo en cuenta la audiencia a la que va dirigida. Así, por ejemplo, si la audiencia posee pocos conocimientos estadísticos será preferible usar gráficos más simples.

• Novedad – Redundancia: La redundancia se refiere a mostrar la misma información de muchas formas distintas; y la novedad, a mostrar la información de una sola forma. La redundancia causa que los gráficos sean más complejos y puede confundir a los lectores. Por ejemplo, usar gráficos de barras con distintos colores para representar cuál es más alta. En este ejemplo, el color es redundante, ya que no se usa (el dato del más alto está dado intrínsicamente por las barras).

Es importante mencionar un punto que recalca Cairo (2012) en las 6 dimensiones mencionadas anteriormente. Esto es que para la presentación de datos siempre se debe tener en consideración la audiencia a la que esta irá dirigida, vinculado al rol o puesto que tienen los miembros de la audiencia, así como con su formación profesional. Las dimensiones que primarán en los gráficos dependerán de la audiencia a la que va dirigida con la finalidad de lograr una buena transmisión de la información.

1.6. Modelado de datos

De acuerdo a Tagliaferri, Morales, Birbeck y Wan (2019), Machine Learning es un subcampo de la inteligencia artificial cuya meta es entender una estructura de datos y ajustarlos a modelos que puedan ser entendidos y utilizados. Así también, Müller y Guido (2016) mencionan que Machine Learning, también llamado analítica predictiva o aprendizaje estadístico, es un campo que resulta de la interacción entre la estadística, inteligencia artificial y ciencias de la computación, cuyo fin es extraer conocimiento a partir de datos y que en los últimos años ha desarrollado más aplicaciones en distintas industrias: filtros de spam en correos electrónicos, detección de rostros en dispositivos electrónicos, clasificación de secuencias de ADN, clasificación de células cancerosas como malignas o benignas, predicción de ventas, sistemas de recomendación de productos en páginas de ventas, recomendación de tratamientos médicos personalizados, recomendación de películas y música en plataformas de streaming, entre muchas otras aplicaciones en distintos campos que se desarrollan continuamente.

Los problemas de Machine Learning se clasifican en dos categorías basado en cómo se realiza el aprendizaje y la retroalimentación al sistema de aprendizaje. Estas dos categorías son el aprendizaje supervisado, en el cual se requiere datos de entrada etiquetados previamente para entrenar los modelos; y el aprendizaje no supervisado, en el cual no se provee datos etiquetados al sistema y este debe encontrar una estructura en los datos provistos (Awad y Khanna, 2015).

1.6.1. Aprendizaje supervisado

Es uno de los tipos de aprendizaje automático más utilizados y exitosos. El propósito de este método es que el algoritmo aprenda comparando la salida real con las salidas enseñada para encontrar errores y modificar el modelo para mejorarlo (Tagliaferri et al., 2019).

Existen dos tipos de problemas de aprendizaje supervisados llamados clasificación y regresión. En la clasificación el objetivo es predecir una variable categórica, es decir, un

número finito de posibilidades. Por su parte, en la regresión el objetivo es predecir una variable cuantitativa continua. Existen diversos algoritmos asociados cada uno de estos tipos, algunos de los cuales serán mencionados a continuación junto con unas breves descripciones basadas en Müller et al. (2016) y VanderPlas (2016).

• k-nearest neighbors: Es posiblemente el algoritmo de aprendizaje automático más simple. Para hacer una predicción de un nuevo punto de datos, el algoritmo busca los puntos de datos más cercanos en el conjunto de datos de entrada (vecinos más cercanos). Este algoritmo puede ser usado en sus dos variantes: clasificación y regresión. En la figura 24 se observa un ejemplo de clasificación, donde las áreas sombreadas indican que los nuevos datos que se localicen dentro de las regiones serán clasificados en la categoría asociada a la región.

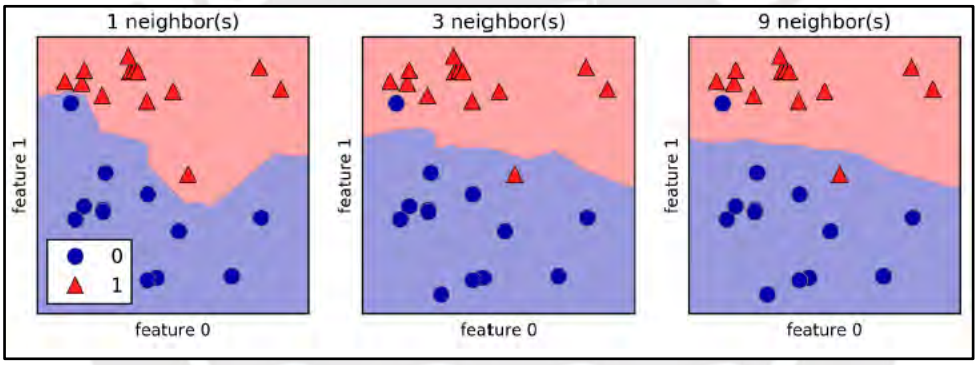

Figura 24 Ejemplo de aplicación de k-nearest neighbors

Tomado de "*Introduction to Machine Learning with Python: A Guide for Data Scientists*", por Müller, 2016.

• Regresión lineal: Es el método lineal más simple y clásico de regresión. Esta minimiza el error cuadrático medio entre las predicciones y los verdaderos objetivos de la regresión, en los datos de entrada. El error cuadrático medio es la suma de las diferencias cuadráticas entre las predicciones y los valores verdaderos. El algoritmo de regresión lineal no tiene parámetros internos, lo cual es un beneficio por su baja simpleza, pero como desventaja no tiene forma de controlar la complejidad del modelo.

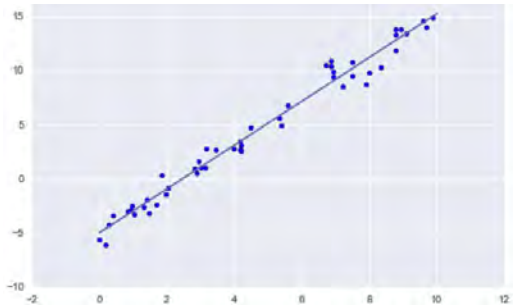

Figura 25 Ejemplo de aplicación de regresión lineal

Tomado de "*Python Data Science Handbook: Essential Tools for Working with Data*", por VanderPlas, 2016.

- Regresión ridge: Es un modelo lineal de regresión regularizado que añade una penalidad en la minimización el error cuadrático medio. Esta penalidad tiene el efecto de restringir la magnitud de los coeficientes del modelo y hacer que adquieran valores cercanos a cero.

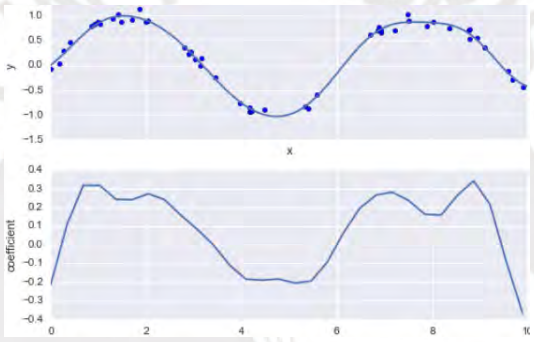

Figura 26 Ejemplo de aplicación de regresión ridge

Tomado de "*Python Data Science Handbook: Essential Tools for Working with Data*", por VanderPlas, 2016.

- Regresión lasso: Es también un modelo de regresión lineal regularizado que añade una penalidad en la minimiza el error cuadrático medio, la cual causa que algunos coeficientes del modelo sean ceros. Es útil cuando se quiere desarrollar un modelo lineal que considere solo las variables que más expliquen la variabilidad de los datos.

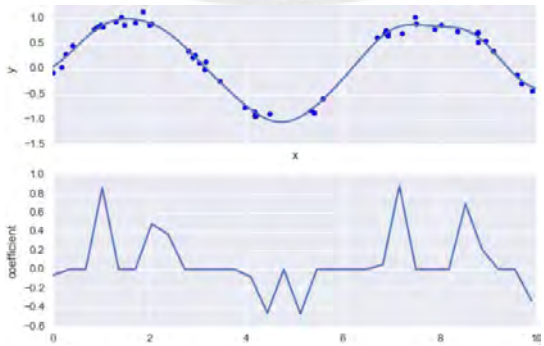

Figura 27 Ejemplo de aplicación de regresión lasso

Tomado de "*Python Data Science Handbook: Essential Tools for Working with Data*", por VanderPlas, 2016.

• Regresión logística: A pesar del nombre, este es un método de clasificación. Para ello hace uso de una función sigmoide para transformar sus valores iniciales predictores (producto de una función lineal) en valores de probabilidad para determinar la categoría a la que pertenece el dato. También se pueden usar las mismas penalidades como en la regresión Ridge y Lasso. En la Figura 28 se observa un ejemplo de regresión logística en el que la línea recta separa las dos áreas que corresponden a cada categoría.

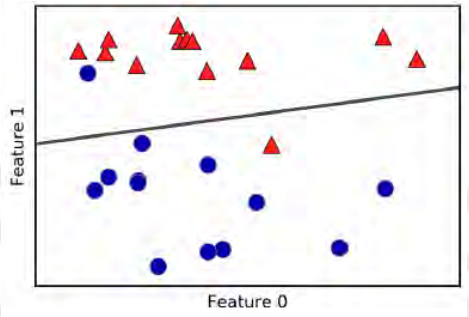

Figura 28 Ejemplo de aplicación de regresión logística
Tomado de "*Introduction to Machine Learning with Python: A Guide for Data Scientists*", por Müller, 2016.

• Support vector machines: Es un tipo poderoso y flexible de algoritmos supervisados, tanto para clasificación como para regresión. Es flexible ya que el algoritmo incorpora algunos parámetros internos. Así también, permite la incorporación de kernels, que puede entenderse como transformaciones a los datos, lo cual lo convierte en un método de bastante poder.

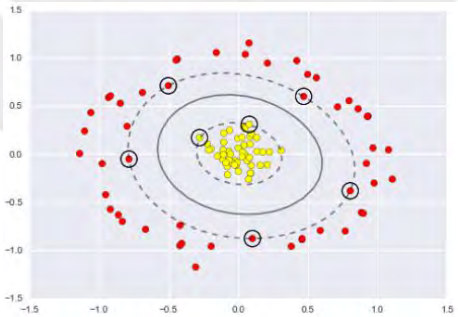

Figura 29 Ejemplo de aplicación de support vector machines
Tomado de "*Python Data Science Handbook: Essential Tools for Working with Data*", por VanderPlas, 2016.

• Árboles de decisiones: Son modelos ampliamente utilizados para tareas de clasificación y regresión. Esencialmente, aprenden una jerarquía de preguntas condicionales (si/si no), lo que lleva a una decisión. Si bien es un método bastante simple, presenta ventajas sobre los otros

modelos mencionados anteriormente, como son su facilidad de visualización e interpretación, y permitir trabajar con una gama amplia de distintos tipos de datos.

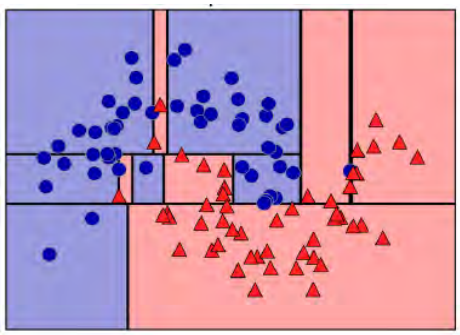

Figura 30 Ejemplo de aplicación de redes neuronales

Tomado de "*Introduction to Machine Learning with Python: A Guide for Data Scientists*", por Müller, 2016.

• Redes neuronales: Son una familia de algoritmos que han formado las bases para el campo del Deep Learning. Estos algoritmos son bastante poderosos y flexibles, ya que incorporan más parámetros y esto le permite capturar estructuras de datos complejas. Aunque, como contraparte, necesita un mayor poder de cómputo.

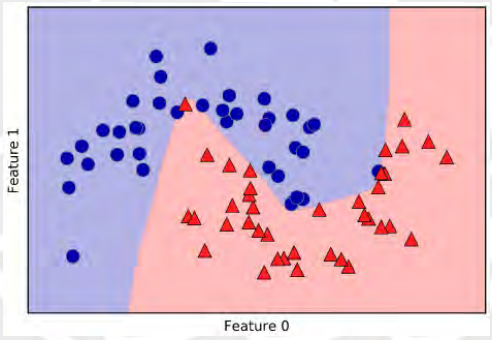

Figura 31 Ejemplo de aplicación de redes neuronales

Tomado de "*Introduction to Machine Learning with Python: A Guide for Data Scientists*", por Müller, 2016.

1.6.2. Aprendizaje no supervisado:

En el aprendizaje no supervisado, al algoritmo de aprendizaje solo se le muestran los datos de entrada y se le pide que extraiga conocimientos de estos datos. El objetivo del aprendizaje no supervisado puede ser tan sencillo como descubrir patrones ocultos dentro de un conjunto de datos, pero también puede tener el objetivo del aprendizaje de características, que permite descubrir automáticamente las representaciones necesarias para clasificar datos (Tagliaferri et al., 2019).

Existen dos tipos de problemas de aprendizaje no supervisados llamados transformaciones de datos y agrupamientos (clustering). Las transformaciones de datos son algoritmos que crean una nueva representación de los datos que podrían ser más fáciles de entender para las personas u otros algoritmos de aprendizaje en comparación con la representación original de los datos. Por su parte, los algoritmos de agrupación dividen los datos en distintos grupos de elementos similares basados en patrones o características de los datos. Existen diversos algoritmos asociados cada uno de estos tipos, algunos de los cuales serán mencionados a continuación junto con unas breves descripciones basadas en Müller et al. (2016) y VanderPlas (2016).

• Principal component analysis (PCA): Es un método de transformación de datos que rota el conjunto de datos de tal manera que las variables rotadas no están correlacionadas estadísticamente. Esta rotación suele ir seguida de la selección de solo un subconjunto de las nuevas variables, según su importancia para explicar los datos. Este modelo es usualmente usado para explorar y visualizar datos con el fin de comprender relaciones entre variables o posteriores agrupamientos.

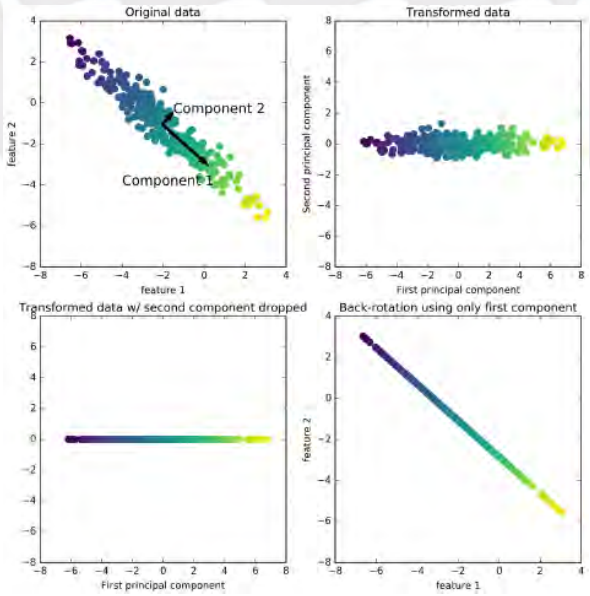

Figura 32 Ejemplo de aplicación de PCA

Tomado de "*Introduction to Machine Learning with Python: A Guide for Data Scientists*", por Müller, 2016.

• Non-negative matrix factorization (NMF): Es otro algoritmo que tiene como objetivo extraer características útiles. Funciona de manera similar al PCA y también se puede utilizar para la reducción de dimensionalidad. Pero mientras que el PCA busca explicar la mayor variación de los datos posible, el NMF busca que los componentes y los coeficientes sean no negativos; es decir, que tanto los componentes como los coeficientes sean mayores o iguales a cero.

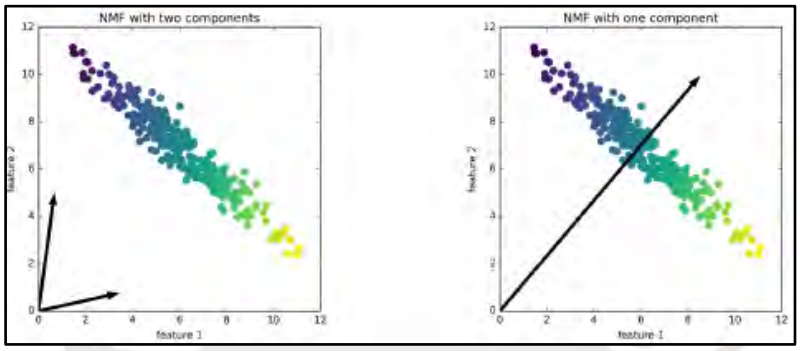

Figura 33 Ejemplo de aplicación de NMF

Tomado de "*Introduction to Machine Learning with Python: A Guide for Data Scientists*", por Müller, 2016.

• Manifold learning: También es un algoritmo de trasformación de datos. Es bastante eficiente para encontrar estructuras en pocas variables a partir de una base de datos compuesta por muchas variables. Así también, es bastante útil para descubrir relaciones no lineales en los datos, lo cual no es tan factible con los métodos mencionados anteriormente.

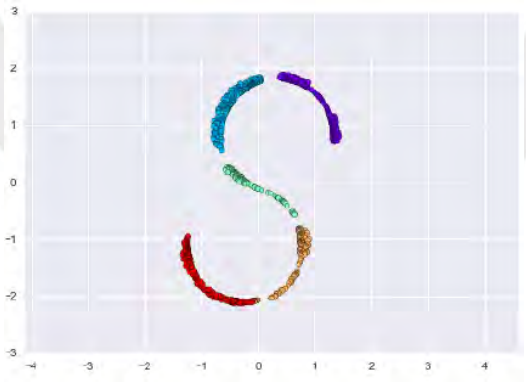

Figura 34 Ejemplo de aplicación de manifold learning

Tomado de "*Python Data Science Handbook: Essential Tools for Working with Data*", por VanderPlas, 2016.

• k-means clustering: Es uno de los algoritmos de agrupación en clúster más simples y más utilizados. Intenta encontrar centros de grupos que sean representativos de determinadas

regiones de los datos. El algoritmo alterna entre dos pasos: asignar cada punto de datos al centro del grupo más cercano y luego establecer cada centro del grupo como la media de los puntos de datos que se le asignan. Tiene el inconveniente de no funcionar eficientemente cuando los datos tienen distribuciones de forma globular.

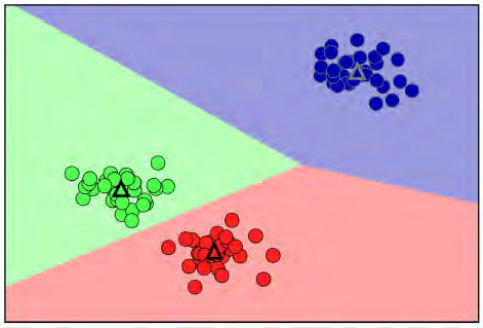

Figura 35 Ejemplo de aplicación de k-means clustering

Tomado de "*Introduction to Machine Learning with Python: A Guide for Data Scientists*", por Müller, 2016.

• Agglomerative clustering: Es otro algoritmo de agrupamiento que se basa en los mismos principios que el k-means clustering, pero en este caso el algoritmo comienza declarando cada punto como su propio grupo y luego fusiona los dos grupos más similares hasta que se cumple algún criterio de detención definido previamente. Tiene como ventaja que permite visualizar niveles jerárquicos de agrupamiento que pueden ser visualizados mediante dendrogramas.

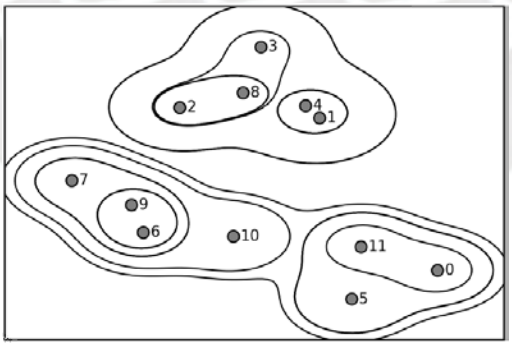

Figura 36 Ejemplo de aplicación de agglomerative clustering

Tomado de "*Introduction to Machine Learning with Python: A Guide for Data Scientists*", por Müller, 2016.

• DBSCAN: Los principales beneficios de DBSCAN son que no requiere que el usuario establezca el número de grupos a priori, como sucede con los modelos de agrupamiento mencionados anteriormente, ya que puede capturar grupos de formas complejas y puede

identificar puntos de datos que no forman parte de ningún grupo, por lo cual es más robusto ante outliers.

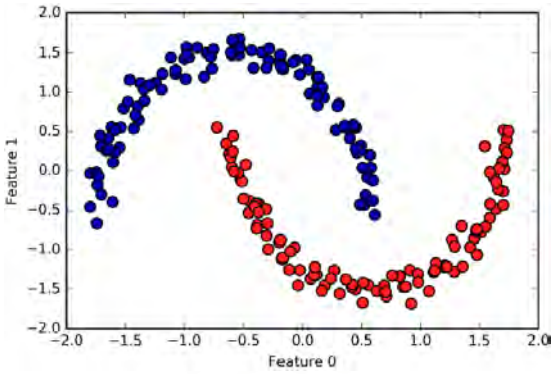

Figura 37 Ejemplo de aplicación de DBSCAN

Tomado de “Introduction to Machine Learning with Python: A Guide for Data Scientists”, por Müller, 2016.

Capítulo 2. Estado de arte

En este capítulo se hará mención de casos de aplicación de Data Science en distintas industrias, vinculándolos con las actividades desarrolladas durante el proceso, así como con los resultados que se lograron con su aplicación.

2.1. Servicios de salud

2.1.1. Detección de nódulos pulmonares y cáncer por tomografía computarizada

El trabajo de Rubin (2015) hace mención de técnicas alternativas de visualización de los nódulos pulmonares para mejorar la eficacia de la interpretación de cáncer de pulmón. Así, primero se recapitulan resultados obtenidos de algunos trabajos como con el de Horeweg et al. (2014), en el que se realiza un estudio con 34 pacientes con cáncer de pulmón haciendo uso de una detección con un ensayo NELSON. En este estudio, el 50% de los casos no fueron diagnosticados debido a errores de detección e interpretación radiológica. Entre estos 17 pacientes en los que la falla en la detección del cáncer de pulmón, 13 se atribuyeron a un error de detección, dos a un error humano y dos a un error de interpretación.

Por otro lado, en el estudio de Xu et al. (2014) se realiza una revisión de exámenes anuales de detección por tomografía computarizada que se realizaron en 104 pacientes con cáncer de pulmón y se indica que: los cánceres no fueron visibles en el 23% de los casos y el 58% de los casos fueron visibles mas no detectados.

Entonces, estos resultados con baja eficacia son debidos a la sensibilidad limitada y variada para la detección de nódulos pulmonares. Es por ello que Rubin (2015), hace mención de la detección asistida por computadora (CAD), la cual emplea algoritmos de Machine Learning que extraen y analizan los datos de las tomografías computarizadas para aislar y resaltar regiones donde es probable que se produzcan anomalías. Dentro de las ventajas se tienen que el CAD facilita la detección de nódulos más pequeños, lo cual reduce la probabilidad

que estos nodos no sean vistos por los radiólogos. Así, se realizó una comparativa entre resultados de detección solo con siete radiólogos, y otra de los mismos radiólogos asistidos por CAD, de esta se observó que la precisión aumentó del 77% y el 72% en la evaluación inicial al 84% y 80% con CAD (Jeon et al., 2012).

Por ende, de este caso se aprecia la utilidad del Machine Learning en el ámbito de la detección de cáncer pulmonar, lo cual posteriormente conlleva a un mejor tratamiento y un incremento en la tasa de supervivencia al cáncer producto de una detección temprana.

2.1.2. Árboles de Decisión para análisis de decisión de riesgos de infarto

Cleophas y Zwinderman (2013) muestra una aplicación de los árboles de decisiones para predecir los riesgos y las mejoras para la salud, para esto se refiere a un estudio para predecir riesgos de infarto. Para ello se cuenta con datos de 1004 pacientes de factores de riesgo de infarto de miocardio. Dentro de las variables a usar se tuvieron: nivel de colesterol, edad, peso, tabaquismo, entre otros. Con estos datos se utiliza un modelo llamado interacción automática chi-cuadrado (CHAID) para el análisis. En la Figura 38 se muestra el árbol de decisión resultante, donde el nivel de colesterol es el mejor predictor de la clasificación de infarto. Para los pacientes con colesterol bajo, el nivel de colesterol es el único predictor necesario para su clasificación; mientras que para los pacientes con nivel de colesterol medio y alto el siguiente predictor a usar es la edad del paciente. En los pacientes ancianos con colesterol medio, el tabaquismo contribuye considerablemente al riesgo de infarto, ya que el 100% de los pacientes ancianos con colesterol medio que fuman presentan riesgo de infarto. Por otro lado, en los pacientes jóvenes con colesterol alto, el tener un peso normal implica un riesgo mayor de infarto a tener un peso normal (un incremento del 11.5% de pacientes que presentan riesgo).

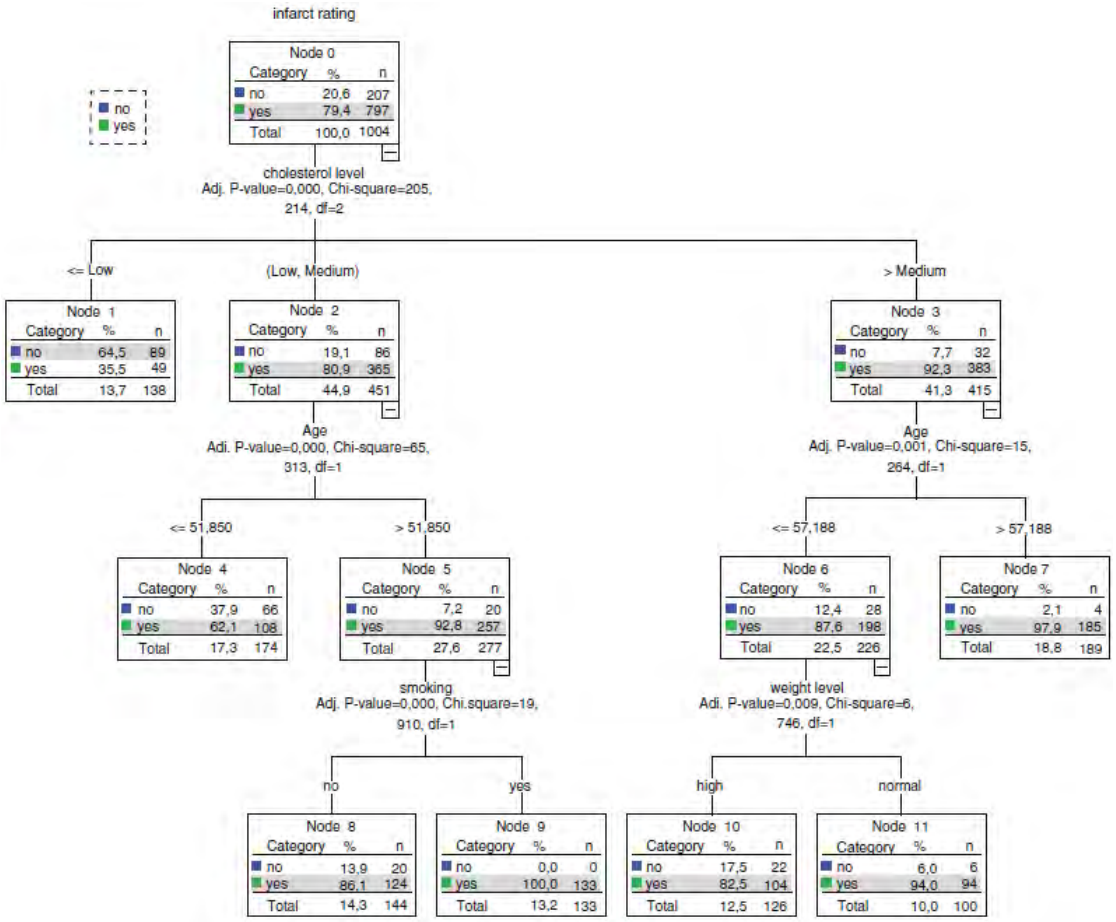

Figura 38 Árbol de decisión para predecir riesgo de infarto

Tomado de "*Machine Learning in Medicine: Part Three*", por Cleophas y Zwinderman, 2013.

Este modelo de clasificación de riesgo tuvo como resultado que el 16.6% de los casos presentan una categoría de infarto predicha incorrecta y 83,4% de los casos presentan una predicción correcta. Pero, algo a tener a consideración es que, dado que es una aplicación médica, es más relevante la medida de evaluación sensibilidad frente a la exactitud de la modelo mencionada anteriormente, ya que se busca penalizar los falsos negativos pues constituyen el peor escenario en el contexto médico. Entonces, en este modelo se obtuvo una sensibilidad del 93.8%, lo cual es un valor alto pues constituye solo un 6.2% de falsos negativos. Este modelo es un ejemplo del gran avance que presentan los modelos de aprendizaje supervisado, como en este caso son los árboles de decisión, que pueden ser usado para discernir el riesgo potencial de enfermedades en pacientes y con ello servir de apoyo para análisis médicos.

2.1.3. Regresión logística para predicción de mortalidad de pacientes

Hsu et al. (2020) realizan un estudio donde compara tres sistemas de puntuación de uso común en medicina para predecir la mortalidad a corto y largo plazo, con la finalidad de construir un modelo de predicción de mortalidad para pacientes enfermos críticos. Estos tres sistemas de puntuación son la Evaluación de la Salud Crónica y Fisiología Aguda II (APACHE II), el Índice de Comorbilidad de Elixhauser (ECI) y el Índice de Comorbilidad de Charlson (CCI).

En este estudio se usó la Base de Datos Nacional de Investigación sobre Seguros de Salud de Taiwán (NHIRD), la cual incluye datos de pacientes como son: edad, peso, comorbilidades, insuficiencias orgánicas, presencia de cirugías, número de ingresos a emergencias en el año, uso de ventilador en emergencias, entre otros. De esta base de datos se usaron los registros de los pacientes que ingresaron a la Unidad de Cuidados Intensivos (UCI) en el periodo del 1 de enero al 31 de diciembre de 2012, y solo se tomó en cuenta el primer ingreso para evitar que un pequeño grupo de pacientes sesgue las características de la población. En total se usaron los datos de 1608 pacientes en el estudio.

Para evaluar el riesgo de mortalidad, se desarrollaron modelos de regresión para identificar los predictores de mortalidad más fuertes, que luego se ingresaron en múltiples modelos de regresión logística para evaluar el desempeño general del modelo y predecir el riesgo de mortalidad. Posteriormente, para medir el desempeño del modelo se usó la medida del área bajo la curva característica operativa del receptor (AUROC), la cual penaliza la presencia tanto de falsos positivos como falsos negativos, y se consideran como valores superiores a aquellos por encima de 0.7. En la Figura 39 se observan los resultados de las medidas AUROC de las distintas puntuaciones médicas evaluadas en los modelos de regresión logística.

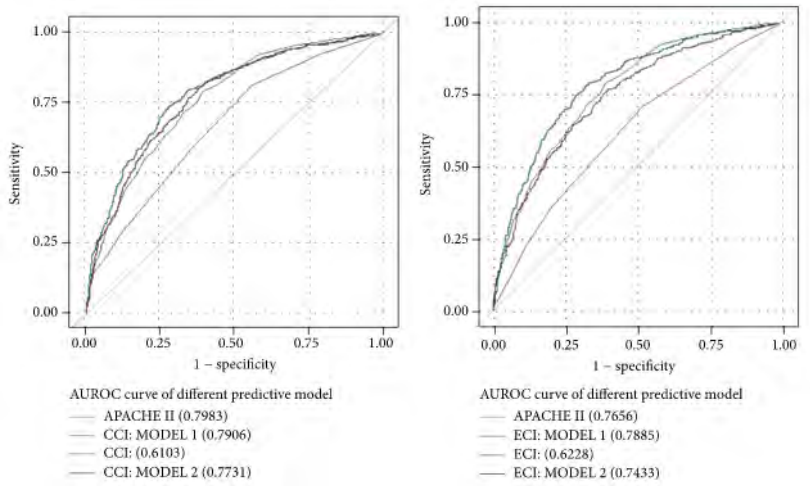

Figura 39 AUROC de modelos de regresión logistica para predicción de mortalidad
Tomado de "*Administrative and Claims Data Help Predict Patient Mortality in Intensive Care Units by Logistic Regression: A Nationwide Database Study*", por Hsu et al., 2020.

Como resultado se obtuvo que la puntuación APACHE II tuvo el mejor poder predictivo de mortalidad hospitalaria con un valor cercano al 0.8, y por su parte, las puntuaciones CCI y el ECI presentan poderes predictivos más pobres. En conclusión, este modelo, usado conjuntamente con la puntuación APACHE II, sirven como un modelo aceptable para predecir riesgo de muerte en los pacientes ingresados a la UCI, y es de utilidad para brindar cuidado temprano a los pacientes de riesgo y con ello incrementar la tasa de supervivencia de pacientes que ingresan a UCI anualmente.

2.1.4. Diagnóstico de lesiones de melanoma maligno aplicando support vector machines

En el estudio de Jaworek-Korjakowska (2016) se desarrolla un modelo de detección de melanoma maligno mediante el análisis de imágenes, el cual es una forma de cáncer de piel menos común pero más peligrosa y que está presentando un crecimiento de casos debido a la exposición excesiva a los rayos ultravioleta sol. Este estudio se dividió en cuatro etapas: preprocesamiento (mejora de la imagen), segmentación de las lesiones cutáneas, extracción y selección de características, y clasificación.

En la mejora de la imagen se refinan las imágenes y se reduce el efecto del ruido debido

a las fuentes de interferencia de tal forma que las características de la imagen sean más fáciles de percibir y detectar. En esta etapa se incluye el alisado de burbujas de aire y la homogenización de vellos en la imagen. Por su parte, la segmentación de las lesiones cutáneas divide las imágenes en regiones homogéneas en términos de intensidad de píxeles, y en este estudio se hace uso del algoritmo de crecimiento de la región sembrada, el cual busca grupos de píxeles que tengan intensidades similares. Este algoritmo tiene la ventaja de ser es capaz de segmentar correctamente regiones que tienen las mismas propiedades y que están separadas espacialmente. En la Figura 40 se aprecia el resultado de la aplicación de este algoritmo.

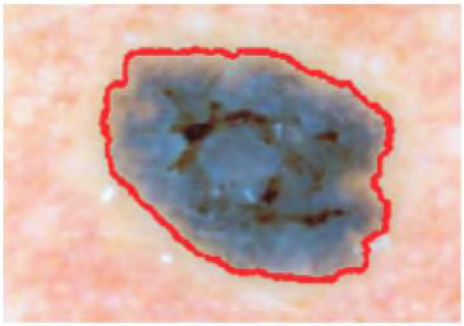

Figura 40 Segmentación de las lesiones cutáneas
Tomado de "*Computer-Aided Diagnosis of Micro-Malignant Melanoma Lesions Applying Support Vector Machines*", por Jaworek-Korjakowska, 2016.

Respecto a la extracción y selección de características, se debe analizar la imagen y reconocer los parámetros críticos que pueden indicar melanoma. Estas características incluyen características de forma (asimetría, variación de color, intensidad, morfología, etc.) y características de textura (uniformidad, regularidad, etc.). Así, para el estudio se eligieron las siguientes características para el proceso de clasificación: compacidad, solidez, distancia de simetría, características de variación de color, variación de entropía, contraste, disimilitud y variación de elipse. Finalmente, para la clasificación se usó el modelo de aprendizaje supervisado de support vector machines.

Como resultado del estudio se obtuvieron buenas medidas de desempeño, donde se puede mencionar una sensibilidad del 90% (que penaliza los falsos negativos) y un AUROC del

93.24% (que penaliza falsos positivos y falsos negativos). Se aprecia que se obtiene una clasificación de los melanomas de forma muy precisa y se concluye que el algoritmo propuesto se puede utilizar para el diagnóstico de lunares cutáneos muy pequeños, contribuyendo de esta forma a diagnosticar un melanoma en una etapa temprana y a la reducción de la tasa de mortalidad relacionada con el melanoma.

2.1.5. Semáforo epidemiológico frente a la pandemia de COVID-19

Más recientemente, uno de los trabajos de Data Science de mayor impacto en el sector salud frente a la pandemia del presente año en el Perú ha sido el estudio de Burhum (2020), en el cual se realiza un análisis del impacto de la pandemia en las regiones en términos de número de personas fallecidas y contagiadas por COVID-19, todo esto a lo largo de los meses que duró la cuarentena desde marzo hasta julio. A partir de este análisis del impacto por regiones se permitió inferir que el efecto de la pandemia estaba más acentuado en algunas regiones donde la cantidad de contagiados y fallecidos eran mayores, por lo cual se deberían aplicar distintas políticas de salubridad en base a las regiones y su posición frente a la pandemia. En la Figura 41 se aprecia la cantidad de fallecidos por COVID-19 hasta el 29 de mayo de 2020.

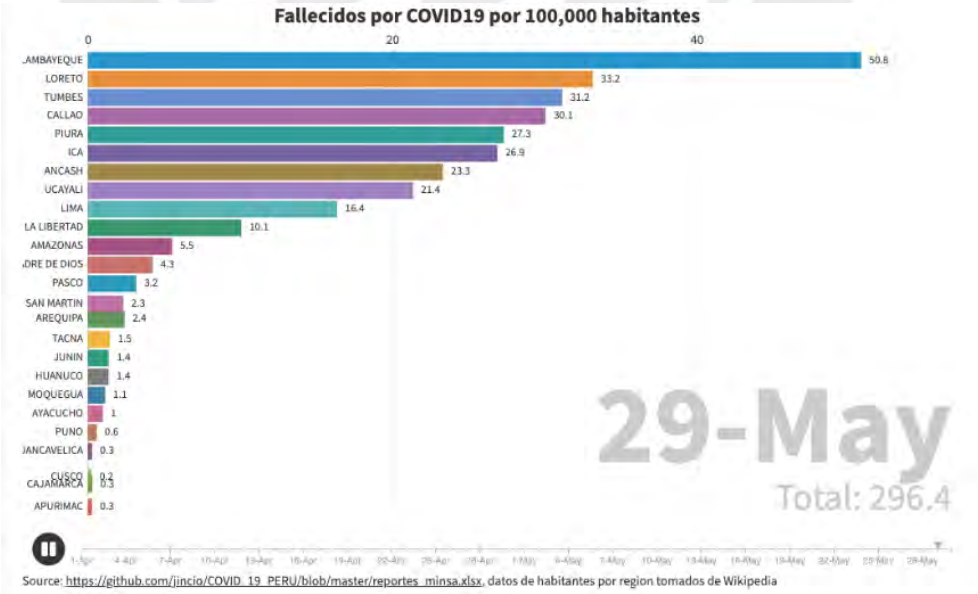

Figura 41 Cantidad de fallecidos por COVID-19 por cada 100 mil habitantes por regiones
Tomado de "*Un Semáforo para el Huayno*", por Burhum, 2020.

Además, a partir de este análisis por regiones se plantea el desarrollo de un semáforo epidemiológico, el cual es un sistema de monitoreo que mide el impacto de la pandemia teniendo en cuenta la población de regiones o poblaciones. Consiste en medir la tendencia de un indicador de salubridad en los últimos 7 o 15 días de la localidad o región; los valores de esa tendencia definen el nivel de riesgo y con ello, la estrategia que se deben utilizar en cada región o localidad. Dentro de estos indicadores se pueden mencionar la carga hospitalaria, casos nuevos de COVID-109, etc.

En el trabajo de Burhum (2020) se desarrolló el semáforo epidemiológico con el indicador de casos positivos de COVID-19 por cada 100 mil habitantes (cp100k), el cual fue medido semanalmente. Este semáforo constó de 5 colores: blanco (0 cp100k), que indica una región sin casos de COVID-19; verde (0 – 22 cp100k), que indica una región tranquila; amarillo (22 -100 cp100k), que indica una región donde se deben aplicar restricciones medianas; naranja (100 - 600 cp100k), que indica una región donde se deben aplicar restricciones altas; y rojo (mayor a 600 cp100k), que indica una región de prioridad de contención. En la Figura 42 se aprecia el estado de algunas regiones del Perú al 25 de mayo de 2020.

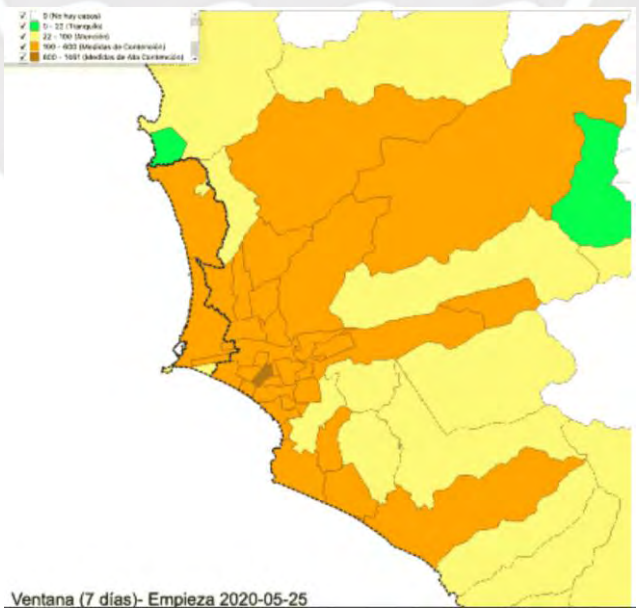

Figura 42 Aplicación del semáforo epidemiológico en las regiones del Perú

Tomado de "*Un Semáforo para el Huayno*", por Burhum, 2020.

Este trabajo es un importante ejemplo del uso de Data Science en materia de salud pública, en el cual se generó nuevo conocimiento a partir del análisis de grandes volúmenes de datos que eran actualizados semanalmente. A partir de esto se desprendió una guía que permitió priorizar a algunas regiones del país por su grado de impacto por la pandemia. Además, este estudio tuvo tal relevancia en su momento, que fue tomado en consideración por el Gobierno del Perú para definir políticas públicas durante la pandemia.

2.2. Transporte

2.2.1. Control de señales de tráfico adaptativo mediante redes neuronales

Castro, Hirakawa y Martini abordan su estudio con el fin de brindar una solución para el control de señales de tráfico. Esta tarea es una de las principales prioridades en las grandes metrópolis debido al impacto en el orden de la sociedad; sin embargo, es una tarea compleja pues el tráfico urbano es un ecosistema complejo y dinámico. Así, ha habido distintas propuestas de solución, desde métodos de optimización mediante tiempos verdes fijos para semáforos con el fin de reducir el tiempo medio de viaje de los vehículos, hasta el uso de redes neuronales de inspiración biológica más recientemente. Estos últimos han resultado en una gran mejoría del comportamiento dinámico de los modelos y, además, no demandan grandes volúmenes de datos y tiempos de entrenamiento para representar adecuadamente el comportamiento del sistema de tráfico urbano.

Entonces, Castro, Hirakawa y Martini estudian un escenario conformado por una única intersección con cuatro fases de semáforo y propone un modelo de red bio-neuronal donde cada fase de semáforo está representada por una neurona excitadora con plasticidad intrínseca, que es el mecanismo de adaptación del modelo. El proceso de adaptación se realiza mediante la plasticidad intrínseca que facilita la transición entre neuronas activas aumentando la activación de neuronas inactivas y disminuyendo la activación de neuronas activas. El esquema de este

modelo se muestra en la Figura 43.

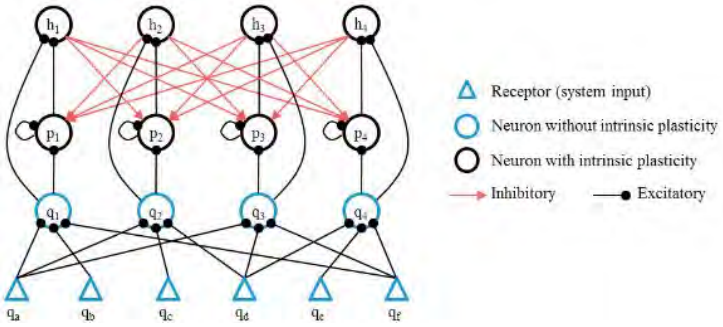

Figura 43 Modelo de red bio-neuronal para el control de señales de tráfico

Tomado de "*Adaptive traffic signal control based on bio-neural network*", por Castro et al., 2017.

Esta rede neuronal propuesta se implementó en MatLab y se evaluaron distintos escenarios que concluyeron que este modelo de control era capaz de soportar una demanda de 3086 vehículos por hora durante tres horas sin saturarse, lo cual representó un 14,3% más que otros métodos de control sensible al tráfico evaluados.

2.2.2. Predicción del flujo de tráfico

La predicción del flujo de tráfico es importante para la planificación de rutas, repercutiendo esto en la congestión vehicular, cantidad de accidentes de tráfico, tiempos de viaje de los vehículos, entre otros. Dado la importancia que tiene esto han surgido diferentes modelos para la predicción, siendo de los más usados los modelos de media móvil integrada autorregresiva de Box-Jenkins (ARIMA); sin embargo, estos modelos presentan la desventaja de necesitar una base de datos sólida para su construcción. Es por esto que Kumar y Vanajakshi (2015) realizan un estudio en el que usan un modelo Seasonal ARIMA (SARIMA), de tal forma que se realiza una predicción a corto plazo del flujo de tráfico utilizando datos de entrada limitados.

Para el estudio se seleccionó un tramo de tres carriles de la carretera Rajiv Gandhi en Chennai (India) y se utilizaron datos de flujo de solo tres días consecutivos para el desarrollo del modelo SARIMA. Se tomaron datos de un sensor de tráfico automatizado en una sola dirección del tráfico cada diez minutos, es decir, se contabilizó el número total de vehículos en

intervalos de tiempo de diez minutos. Luego de recabados los datos, estos se dispusieron en una serie temporal para examinar las características de tendencia y estacionalidad, y realizar transformaciones a los datos de ser necesario. En la Figura 44 se muestra la serie temporal de los datos observados y los predichos.

Una vez desarrollado el modelo se validó y se obtuvo como resultado un error porcentual absoluto medio (MAPE) entre el flujo observado y el predicho de 9.22, el cual al ser menos del 10% se considera altamente preciso; además, comparativamente con otros estudios presenta mejor rendimiento, pues estos han presentado un MAPE mayor en el rango de 10 a 20%. En base en los resultados se concluye que el modelo SARIMA es adecuado para predecir el flujo de 24 horas con una precisión aceptable en comparación a otros modelos.

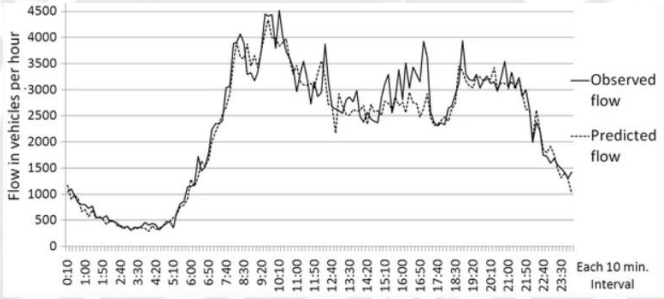

Figura 44 Modelo SARIMA para predicción del flujo de tráfico

Tomado de "*Short-term traffic flow prediction using seasonal ARIMA model with limited input data*", por Kumar y Vanajakshi, 2015.

2.2.3. Clasificación de vehículos por atributos geométricos y de apariencia

Una tarea que está cobrando relevancia debido al aumento de la población y con ello el número y tipo de vehículos, es la necesidad de clasificar los vehículos de manera eficiente, ya que presenta relevancia para la planificación de la infraestructura y la gestión del tráfico como pueden ser: vigilancia del tráfico, pago de peaje, prevención de la congestión del tráfico, prevención de accidentes vehiculares, etc. Algunas dificultades que se presenten en este ámbito son el incremento de la cantidad de modelos y tamaños de vehículos; además, la oclusión, sombra e iluminación contribuyen en la dificultad del problema.

Moussa, G. (2014) realiza un estudio donde presentan dos enfoques de clasificación de vehículos usando el modelo de support vector machines, estos dos enfoques son: enfoque basado en geometría y enfoque basado en la apariencia. Así, la clasificación se divide en dos tareas: clasificación de vehículos multiclase y clasificación de vehículos intraclase. En la primera clasificación los vehículos se clasifican por tamaños en tres clases: pequeño, mediano y grande; por su parte, en la segunda clasificación los vehículos se clasifican por subclases: camioneta, vehículo deportivo utilitario y furgoneta.

Este modelo se desarrolló en MatLab y se evaluó con una técnica de validación cruzada de 10 veces, y se usó como medida de evaluación la sensibilidad. En la clasificación multiclase se obtuvo una sensibilidad de 0.967 y 0.958 para los enfoques basados en la geometría y la apariencia, respectivamente. Por su parte, en la clasificación intraclase se obtuvo una sensibilidad de 0.553 y 0.896 para los enfoques basados en la geometría y la apariencia, respectivamente. Entonces, para la clasificación multiclase, ambos enfoques brindan resultados muy similares; pero, para la clasificación intraclase es recomendable el enfoque basado en la apariencia. Estos resultados muestran el potencial del uso de modelos de Machine Learning para la clasificación de vehículos, que tienen un gran ámbito de aplicaciones en el mundo real descritos anteriormente.

2.2.4. Mapa de riesgo de accidentes de tránsito

Otro problema de importancia en el ámbito del sector transporte es hacer frente el creciente número de accidentes de tránsito. Para ello es indispensable comprender las causas de los accidentes de tránsito con el fin de brindar soluciones. Dentro de las posibles causas relacionadas con los accidentes de tránsito se pueden mencionar el comportamiento del conductor, el clima, las condiciones de la carretera, congestión vehicular, etc. El problema vinculado a estos factores es su dinamismo, sobre todo aquellos vinculados intrínsecamente

con el conductor.

En base a esto; Chen, Song, Yamada y Shibasaki (2016) plantean un estudio que busca estimar el riesgo de accidentes de tráfico a través de datos de ubicación en tiempo real. Para esto se usan 300 mil registros de accidentes de tránsito y registros de GPS de 1.6 millones de usuarios, datos recopilados en Japón en un periodo de 7 meses. Luego estos datos fueron procesados y discretizados en dimensiones espaciales y temporales para finalmente obtener matrices de frecuencias, donde es más factible realizar un mapeo entre la movilidad humana y el nivel de riesgo. Posteriormente, se utiliza un modelo autoencoder de eliminación de ruido apilado (SdAE), el cual permite extraer una representación jerárquica de características.

Finalmente, se evalúa el modelo desarrollado en diferentes momentos del día en las ciudades de Tokio y Yokohama. Los resultados obtenidos se comparan con otros modelos desarrollados en otros trabajos que hacen uso de árboles de decisión, regresiones logísticas y support vector machines; y se encuentra que el modelo SdAE presenta un mejor desempeño pues presenta un menor error de predicción. En la Figura 45 se muestra la visualización de los datos de movilidad de las personas y los resultados del modelo SdAE, donde se muestra un mapa de riesgo de accidentes de tráfico que puede ser usado para predecir zonas de riesgo en tiempo real y así controlar los riesgos en el tránsito vehicular.

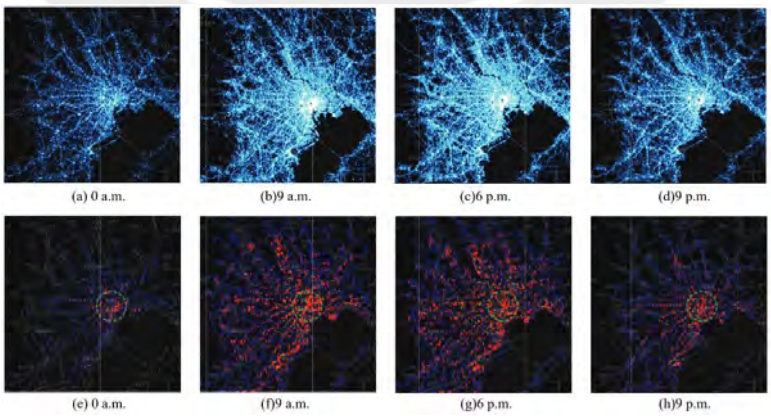

Figura 45 Mapa de riesgo de accidentes de tráfico en Tokio y Yokohama

Tomado de "*Learning Deep Representation from Big and Heterogeneous Data for Traffic Accident Inference*", por Chen et al., 2016.

2.2.5. Predicción de demanda de viajes

La predicción de la demanda de viajes busca estimar el número de usuarios del transporte público. Es uno de los principales problemas de transporte, ya que muchos otros modelos de transporte utilizan la demanda de pasajeros como datos de entrada. Es así que Davis, Raina y Jagannathan (2016) realizan un estudio que busca realizar la predicción de la demanda de pasajeros de una empresa de taxis en la ciudad de Bengaluru en la India. Para esto se recolectaron datos en un periodo de más de dos meses, con lo cual se recopiló la fluctuación de más de 14 millones de viajes en dicho periodo. Con los datos recopilados se analizó la tendencia y estacionalidad para posteriormente realizar el procesamiento de datos, donde se aplicó una transformación Box-Cox a los datos de tal forma que se estabilice su varianza.

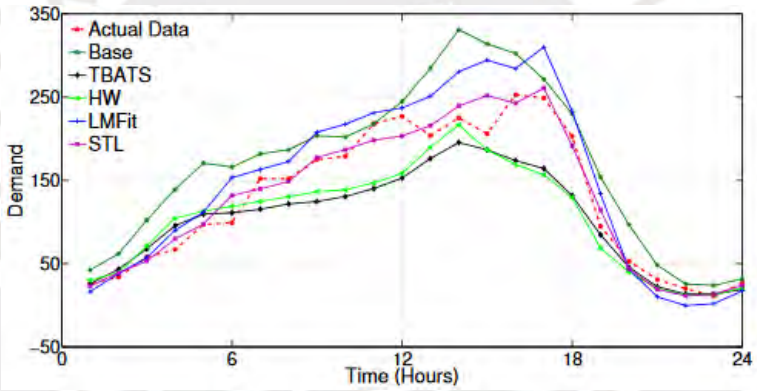

Figura 46 Predicción de demanda de pasajeros con distintos modelos de predicción
Tomado de "*A multi-level clustering approach for forecasting taxi travel demand*", por Davis et al., 2016.

Posteriormente se realizó el modelado de los datos, donde se aplican distintos modelos de predicción entre los que están: modelo de línea base (Base), regresión lineal con tendencia y estacionalidad (LMFit), descomposición estacional y de tendencias usando Loess (STL), modelo TBATS, y el modelo Holt Winters (HW). Así también, con el fin de mejorar el rendimiento del modelo se usó una técnica de clustering multinivel con el que se logra incrementar el rendimiento de los modelos reduciéndose el MAPE en un 5% en promedio. Finalmente se alcanza una precisión del 89% en un área de 1 km^2, lo cual constituye una precisión aceptable en base a tener como rendimiento el MAPE.

2.3. Finanzas

2.3.1. Optimización de carteras de inversión

Aldridge (2019) hace un recuento de estudios realizados sobre optimización de carteras de inversión, en el que detalla métodos usados tradicionalmente y las mejoras que traen consigo el uso de Data Science y, específicamente, el Big Data. Así, se menciona que, como principio fundamental, las inversiones deben diversificarse de modo que, si el valor de una inversión disminuye, las otras aumentan o al menos contrarrestan el valor total de la cartera de inversiones.

Los métodos tradicionales que se usan para esta tarea se dividen en 2 enfoques: bayesianos y no bayesianos. Aquí se puede mencionar el modelo de optimización tradicional Markowitz o también llamado optimización de la varianza media (MVO), la cual busca aumentar los rendimientos medios mientras que simultáneamente disminuye la varianza en las carteras. Es cierto que cuando la cartera de inversiones es relativamente pequeña y estable este modelo puede funcionar bien; sin embargo, para carteras más grandes, como fondos mutuos y fondos de cobertura con activos valorados en miles de millones de dólares, la diversificación se ve afectada por matrices de varianza-covarianza inestables, grandes costos de transacción y restricciones de liquidez.

Ante estas problemáticas, Aldridge (2019) usa una técnica de Big Data, la descomposición espectral y optimización de la inversa de la matriz de correlación de activos. Utiliza esta técnica para realizar comparaciones con los métodos tradicionales de optimización de portafolios en dos experimentos en un periodo de 20 años (1998 - 2017): comparación de las técnicas de gestión de la cartera principal y comparación de las técnicas de gestión de carteras en 1000 carteras con 50 o más acciones cada una.

Con este estudio se demostró que con el uso de la técnica de Big Data propuesta se es

más sensible a las perturbaciones, lo cual afecta las estrategias de asignación de la cartera de inversión. Así también, con esta técnica se mantuvo el valor total de la cartera de inversión en un margen de 2% de variación en el periodo de 20 años, pero con un menor número de acciones que con las otras técnicas tradicionales, lo cual implica una mejor optimización de la diversificación de las inversiones.

2.3.2. Detección de fraudes

Uno de los casos de aplicación donde el Data Science está bastante desarrollada es en la prevención de fraudes financieros. Aquí, por ejemplo, se puede mencionar el fraude con tarjetas de crédito, donde el Machine Learning ha provisto de grandes avances en la solución de este problema. Awoyemi, Adetunmbi y Oluwadare (2017) realizan un estudio donde comparan tres modelos de clasificación usados para la detección de fraudes con tarjetas de crédito: Naive Bayes, k-nearest neighbors y regresión logística.

Para esto se recabaron datos de transacciones con tarjeta de crédito de titulares europeos en setiembre de 2013, con lo cual se tuvieron 284,807 transacciones. Luego, se realizó la selección y reducción de características con el modelo PCA, lo cual resultó en una selección de 28 componentes principales. Posteriormente se entrenaron los tres modelos mencionados y se evaluaron con distintas métricas de evaluación como son la precisión, exactitud y sensibilidad. Los resultados de la evaluación se muestran en la Figura 47.

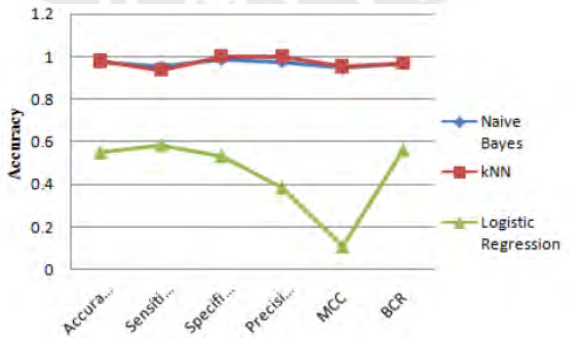

Figura 47 Resultados de métricas de evaluación en tres modelos de clasificación de fraude
Tomado de "*Credit card fraud detection using Machine Learning Techniques: A Comparative Analysis*", por Awoyemi et al., 2017.

De estas métricas, la más importante para esta aplicación de fraude es la sensibilidad, pues se busca penalizar la cantidad de falsos negativos, ya que estos conllevan los más altos costos de producirse. En base a esta métrica, el modelo k-nearest neighbors presentó el mejor desempeño de los tres modelos evaluados con un valor de 88.35% de sensibilidad (aunque, en realidad, mostró los mejores resultados en todas las métricas), seguido por el modelo Naive Bayes y la regresión logística, respectivamente.

2.3.3. Predicción del tipo de cambio de divisas

Otro campo de aplicación del Data Science en las finanzas se da en el pronóstico de tipos de cambio de divisas. Esto es de vital importancia para las empresas que buscan realizar inversiones u obtener financiamiento externo en divisas extranjeras, ya que contar con un buen pronóstico en el corto o mediano plazo ayuda a elegir la opción de menor riesgo financiero.

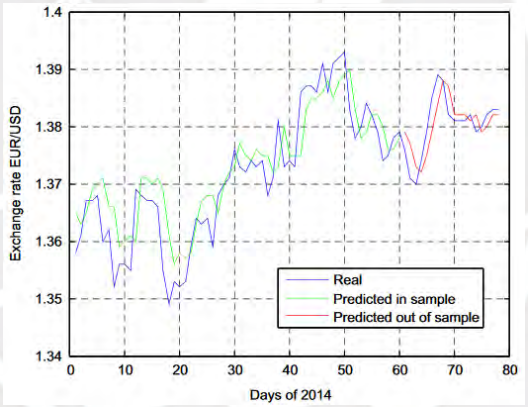

Figura 48 Predicción del tipo de cambio EUR/USD con redes neuronales
Tomado de "*Neural networks performance in exchange rate prediction*", por Galeshchuk, 2016.

Galeshchuk (2016) realiza un estudio en el que afrontan este problema de predicción. Para esto recopilan datos de forma diaria sobre los tipos de cambio EUR/USD, GBP/USD y USD/JPY en el periodo comprendido entre el 1 de enero de 2014 y el 25 de abril de 2014. Así también, recopilan datos de forma mensual de los mismos tipos de cambio mencionados anteriormente en el periodo comprendido desde mayo de 2009 hasta mayo de 2014. Posteriormente, estos datos son modelados mediante redes neuronales y este modelo es evaluado en base al error de predicción. En líneas generales, el método de predicción a corto

plazo proporcionó una buena precisión de la predicción, con errores promedios de 0.2% para los tipos de cambió EUR/USD y GBP/USD, y un error promedio de 0.3% para el tipo de cambio USD/JPY. Entonces, se puede decir que este tipo de modelos sí se puede utilizar en casos prácticos para predecir el tipo de cambio en empresas que realizan negocios internacionales a través de importaciones, exportaciones o inversión extranjera.

2.3.4. Predicción de quiebra empresarial

La predicción de quiebra empresarial constituye uno de los ámbitos de investigación más importantes en el sector financiero. Esto es relevante tanto para las propias empresas como para las financieras. Para las primeras, esta predicción permite examinar su salud financiera para tomar decisiones empresariales que mejoren la solvencia y liquidez con antelación; por su parte, a las financieras les permite evaluar el riesgo de insolvencia empresarial a las empresas que brindan créditos y de esta forma reducir pérdidas por quiebra.

Alaminos (2018) realiza un trabajo donde realiza la predicción de quiebra empresarial mediante redes neuronales. Para esto, inicia recopilando datos de 220 empresas tanto en situación legal de quiebra como no quebradas, en el periodo comprendido entre 1990 y 2013, y en tres continentes (Asia, Europa y Norte América). Dentro de estos datos recopilados se tienen: ingresos netos, activos corrientes, pasivos corrientes, capital, deudas totales, ingresos netos, utilidades antes de intereses e impuestos, entre otros. Posteriormente, se realiza el procesamiento de los datos, donde se realiza una normalización de los datos; así también, se realiza un análisis exploratorio de datos, donde se evalúa la correlación de las variables para evitar contar con variables redundantes en el modelo. Luego se realizó el modelado de los datos, donde se utiliza un modelo de redes neuronales con un total de 12 variables de entrada. En la evaluación de este modelo se obtiene una exactitud promedio del 95.26% para las empresas europeas, 98.45% para las asiáticas y 89.35% para las americanas, y en promedio se

consiguió una exactitud del 94.33% para el total de las empresas evaluadas.

2.3.5. Predicción del precio del oro

Otro uso del Data Science está vinculado al pronóstico de commodities como son el oro, petróleo, plata, cobre, entre otros. Es importante la predicción de sus precios pues los commodities conforman inversiones en las cuales las empresas pueden enfocar sus carteras de inversión. Es por ello que esta información es de vital importancia para reducir el riesgo en este tipo de inversiones en las empresas que invierten en commodities.

Villada, Muñoz y García-Quintero (2016) realizaron un estudio donde predicen el comportamiento del precio del oro mediante un modelo de redes neuronales. Para esto recopilaron datos del precio de cierre diario del oro en el mercado de Londres en 2015. Con estos datos se modelaron los datos en MatLab utilizando el algoritmo de aprendizaje Levenberg Marquardt, el cual se caracteriza por presentar una rápida convergencia. Finamente, en la evaluación del modelo se obtuvo un MAPE de 0.71, el cual es bastante alto considerando la volatilidad de los precios del oro en el periodo evaluado. Este resultado muestra la aplicabilidad del Data Science en el mercado de commodities, lo cual permite obtener un buen estimado del precio del oro del día siguiente para planear compras y ventas de este metal de tal forma que se disminuya el riesgo financiero y se incrementen las ganancias.

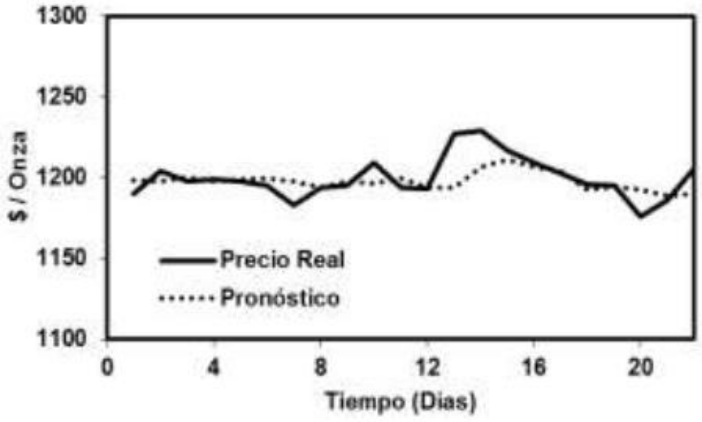

Figura 49 Predicción del precio del oro mediante redes neuronales

Tomado de "*Redes Neuronales Artificiales aplicadas a la Predicción del Precio del Oro*", por Villada et al., 2016.

2.4. E-commerce

2.4.1. Predicción de ventas

El sector de e-commerce o comercio en línea se ha desarrollado recientemente en las últimas décadas debido a la proliferación del uso diario de internet y redes sociales por parte de las personas, lo cual ha abierto un nuevo medio para desarrollar comercios sin necesidad de necesitar una tienda física. Una tarea importante en el e-commerce consiste en la previsión de ventas de los productos de tal forma que las empresas no se queden sin stock de productos; además, una adecuada anticipación de las ventas conlleva a una mejor planificación y reducción de costos de compras y costos logísticos.

Wei, Geng, Ying y Shuaipeng (2014) realizan una investigación en este ámbito, en el que busca predecir las ventas de prendas de vestir. Para esto se recopila datos de búsqueda web y volumen de ventas de la web china de ventas en línea llamada Taobao en el periodo comprendido entre el 29 de junio de 2011 al 27 de setiembre de 2011. Dentro de los datos se tiene información de precios de productos, calidad del producto, origen del producto, opiniones de consumidores sobre el producto, el historial crediticio del comprador, entre otros. Posteriormente se realizó el procesamiento de los datos y el modelado, para este último se utilizó un modelo de series de tiempo con ajuste a la tendencia y estacionalidad del volumen de ventas. Para esto se usó un factor de tendencia calculado mediante regresión lineal y un coeficiente estacional calculado como el promedio del coeficiente estacional del mismo día en diferentes semanas, donde este coeficiente resulta de la relación entre las ventas y el factor de tendencia.

Para la evaluación del modelo se usó el MAPE, teniendo el modelo como resultado un valor de 4.84% en el periodo de una semana de predicción. Este es un resultado bastante bueno y muestra que los modelos de predicción basados en datos de búsqueda web pueden reflejar

los cambios de factores externos a través del comportamiento de búsqueda web de los compradores. Con lo cual, estos modelos presentan una buena capacidad para predecir fluctuaciones debido a factores intrínsecos al comprador.

2.4.2. Análisis de opiniones o reviews de productos

Las revisiones de comentarios, opiniones o reviews de productos en las plataformas web proporcionan feedback sobre productos y/o servicios. Con esto se puede obtener información para mejorar las falencias de los productos; así también, se puede obtener información de las características del producto o servicio que son más valoradas por los consumidores. De esta forma se puede desplegar una mejora continua para incrementar la calidad de los productos y que estos sean más atractivos para los consumidores.

Una aplicación de esta tarea se observa en el trabajo de Mubarok, Adiwijaya y Aldhi (2017), quienes realizan un análisis de sentimiento para realizar una clasificación de las opiniones en restaurantes. Para esto utilizan datos del International Workshop on Semantic Evaluation 2014, donde recaban datos de 3714 reviews de restaurantes. Se realiza el procesamiento de estos datos para obtener la mayor información posibles sobre el precio pagado, ambiente del restaurante, calidad del servicio, anécdotas, entre otros. Esto constituye la parte más laboriosa de la tarea pues es necesario un cuidadoso procesamiento de los datos que incluye tratamiento de palabras en mayúsculas, eliminación de palabras vacías, segmentación o tokenización de palabras, entre otros; de tal forma que se tenga una lista de las palabras usadas en todos los reviews.

De la lista del total de palabras se seleccionaron 219 de estas, que constituye una lista simplificada con las palabras que son más significativas para diferenciar los reviews positivos de los negativos. Posteriormente se usó el modelo de Naive Bayes para la clasificación de los reviews, donde se usaron datos de 3618 reviews para el entrenamiento del modelo y los

restantes 96 para la evaluación del mismo.

Para la evaluación se usó la métrica F1, la cual combina la precisión y sensibilidad en una solo medida y penaliza tanto tos falsos negativos como los falsos positivos. Se obtuvo así un valor de F1 del 78.12%, el cual es aceptable, aunque se menciona que existen estudios que han presentado valores de esta misma métrica del 88.57% haciendo uso de modelos más complejos de Machine Learning. Con estos resultados se puede apreciar el potencial de los análisis de reviews u opiniones de los consumidores para identificar oportunidades de mejora, y cómo el avance del Data Science permite esto con modelos relativamente sencillos como es el Naive Bayes.

2.4.3. Promociones personalizadas

Otro aspecto importante en el e-commerce que se ha podido desarrollar es el tema de las promociones personalizadas. Así pues, mediante el análisis de datos históricos de compras de los usuarios es posible determinar las mejores ofertas para cada comprador, de tal forma que le resulte beneficioso la compra del producto; ya que, mientras que para algunos usuarios es más relevante un descuento en el precio, para otros puede ser un regalo por la compra, y para terceros puede ser prioridad envíos gratis de productos. Entonces, resulta importante diferenciar estas prioridades para cada consumidor de tal forma que se despliegue promociones personalizadas y eficientes que beneficien a las empresas y compradores.

Zhao, Zhang, Friedman y Tan (2015) realizan una investigación que busca la construcción de un sistema que ofrezca promociones personalizadas a los consumidores para aumentar la posibilidad de compra. Para esto, primero recaban información de Amazon sobre productos para el cuidado de la piel. Los datos que se recabaron son el precio de lista, precio de venta, marca del producto, descuento, valoración en reviews, popularidad de la marca, tiempo que dura la compra, entre otros. Luego, estos datos son tratados de tal forma que no se

consideran aquellos registros donde la compra fue menor a cinco minutos o donde el descuento fue demasiado alto.

Con esta información de descubren los productos de mejor valor, es decir, los productos que las personas consideran que vale la pena comprar. En base a estos productos más relevantes se despliegan recomendaciones de productos en base a la premisa “la persona que compró este producto también compró estos otros", con la finalidad de generar promociones atractivas al comprador. Posteriormente, se desplegaron 339 ofertas con estos productos y se evaluaron estas en 54 compradores, donde se usó el beneficio del vendedor como indicador de efectividad, esto resultando de la diferencia entre el precio de venta final menos el costo de producción o adquisición del producto. Finalmente, se menciona que con el trabajo desarrollado se maximiza la ganancia esperada, ya que esta se ve incrementada en un 10% mediante una mejor personalización de los descuentos a los compradores.

2.4.4. Sistemas de recomendación de productos

Con el desarrollo del comercio electrónico, las empresas que deciden usar este medio para sus ventas presentan la necesidad de desarrollar plataformas de ventas que sean de fácil uso y con información necesaria de los productos para los potenciales compradores. Esto no es tan complicado de lograr si se trata solo de una tienda en línea básica; sin embargo, con la necesidad de incrementar las ventas surge la necesidad de desarrollar plataformas inteligentes que permitan que los compradores encuentren rápidamente productos de su interés, que eviten la sobrecarga de información y que realicen recomendación de productos en que los compradores pueden estar interesados. Este tipo de sistemas es llamado sistema de recomendación y muchas empresas los usan en la actualidad, como son: Amazon, Spotify, Netflix, entre otros. Así, Amazon hace uso de estos sistemas para recomendar productos similares o productos complementarios a una compra realizada anteriormente; Spotify los usa

para recomendar canciones o artistas que pueden interesar a los usuarios basado en sus gustos musicales; Netflix lo usa para recomendar nuevas películas basado en el tipo de películas que una persona suele ver y en las valoraciones o reviews que hizo a películas vistas (Sánchez-Corcuera et al., 2020).

Guo, Wang y Li (2017) realizan una investigación donde proponen un algoritmo A priori para la generación de un sistema de recomendación. Para esto usa una base de datos de ventas de prendas de vestir femenina provenientes de la web china de ventas Taobao. Luego se realiza el procesamiento de los datos donde se eliminan los registros incompletos y los outliers, para posteriormente transformar los datos en forma de matriz de tal forma que el modelado de datos sea más eficiente. Respecto a la evaluación del modelo, se usa la exactitud como métrica y se obtiene una precisión del 91.23% y se menciona que el algoritmo propuesto presenta la ventaja de mejorar el tiempo consumido conforme el número de transacciones aumenta. Esto es una ventaja frente a otros algoritmos pues permite la disminuir el tiempo de uso del sistema por parte de los usuarios sin sacrificar la precisión.

Conclusiones

Con base en el desarrollo del presente trabajo, es posible concluir lo dictado en los siguientes enunciados.

- El trabajo de investigación sirve en parte como una guía de aprendizaje de los conocimientos estadísticos, matemáticos e informáticos que un profesional debe adquirir con para desempeñarse en el área de Data Science. Donde los conocimientos estadísticos son de relevancia para el análisis exploratorio de datos, y también sirven de base para el entendimiento de modelos y/o algoritmos de Machine Learning. Por su parte, los conocimientos matemáticos son de utilidad para el entendimiento de algoritmos y también en lo referente a transformaciones de tipos especiales datos, como pueden ser datos acústicos e imágenes. Por último, respecto a los conocimientos informáticos, se debe enfatizar el aprendizaje de distintos lenguajes de programación y plataformas, pasando desde lenguajes más usados en Data Science, como R y Python, hasta sistemas de gestión de bases de datos relacionales y no relacionales, y también, plataforma destinadas específicamente para visualización de datos.
- De acuerdo a lo consultado, si bien todas las actividades de un Data Science son relevantes, muchos autores coinciden en que la etapa más crucial y en la que se debe ser más minucioso es en la preparación de datos en formatos correctos y ordenados. Esto debido a que es la primera actividad que se realiza, y la cual tendrá un impacto directo en los distintos análisis y modelados de datos realizados posteriormente; si esta actividad no es desarrollada correctamente puede conducir a resultados erróneos y desacreditar todo el trabajo posterior a este.
- Machine Learning está cobrando más relevancia en la actualidad y es uno de los causantes que el Data Science se esté tornando tan popular en los últimos tiempos. Dentro de este campo se han desarrollado distintos algoritmos que, junto con una adecuada colección de datos y una buena capacidad de cómputo, hacen posible desarrollar predicciones complejas que han tenido

aplicación en distintos campos: finanzas, con pronósticos de utilidades; medicina, con personalización de tratamientos en base a historial médico; comercio, con recomendación de productos en páginas webs; telecomunicaciones, con algoritmos de detección de rostros; entre otros.

- De acuerdo a lo consultado sobre las aplicaciones del Data Science en las industrias, se ha corroborado su implicancia en los servicios de la salud, donde su uso ha hecho posible la detección temprana de cánceres de piel y de pulmón; así también, se han revisado trabajos donde su uso ha contribuido a la predicción de mortalidad de pacientes que ingresan a UCI y a la predicción de riesgos de infarto en pacientes. Es importante recalcar uno de los trabajos más importantes en el contexto de la pandemia COVID, donde se desarrolló un semáforo epidemiológico para priorizar políticas o medidas se salubridad en algunas regiones del Perú por su grado de impacto por la pandemia. En resumen, el Data Science aplicado en el sector conlleva a un mejor tratamiento y a un incremento en la tasa de supervivencia de enfermedades producto de una detección temprana.
- Respecto a la aplicación en el sector transporte, el Data Science permite un mejor control del tráfico y planificación de rutas, repercutiendo esto en una menor congestión vehicular, menor cantidad de accidentes de tráfico y disminución de tiempos de viaje de los vehículos. En estos casos de uso se corroboraron las aplicaciones de modelos como las redes neuronales, support vector machines, regresiones lineales, entre otros más complejos. Aquí, uno de los casos de aplicación más importantes y que se debería investigar más en el Perú es de la predicción de accidentes de tránsitos, los cuales han demostrado predecir zonas de riesgo en tiempo real y así controlar riesgos de accidentes en el tránsito vehicular.
- Respecto a las aplicaciones en finanzas, se comprueba la aplicación para la detección de fraude, donde se obtuvo una buena sensibilidad de un modelo k-nearest neighbors con un valor de 88.35%. También se ratifica la aplicación para la predicción a corto plazo del tipo de cambio

de divisas haciendo uso de redes neuronales donde se obtuvo bajos errores de pronósticos con un margen de 0.2% a 0.3%. Así también, se muestra un caso donde se usan las redes neuronales para la predicción de quiebra empresarial, logrando una exactitud del 94.33%. En líneas generales, se comprueba que la aplicación del Data Science ayuda a la reducción del riesgo financiero en las inversiones y a incrementar las probabilidades de ganancias.

- Por otro lado, respecto a las aplicaciones en e-commerce, se ha recabado casos de aplicación para predicción de ventas haciendo uso de modelos de series de tiempo con resultados satisfactorios. También, se corrobora la aplicación para análisis de opiniones de productos con el uso del modelo Naive Bayes, lo cual brinda oportunidades de mejora continua para incrementar la calidad de los productos. Además, se hace mención de una de las aplicaciones de más impacto en este sector, que son los sistemas de recomendación de productos, los cuales son cada vez más usados pues permiten realizar recomendación de productos en que los compradores pueden estar interesados y de esta forma incrementar las ventas.
- Por último, al tener el área de Data Science un estrecho vínculo y dependencia con las tecnologías de información, pues esto ha sido uno de los factores que ha impulsado su popularización en la actualidad. Un Data Scientist debe estar en constante aprendizaje de nuevas tecnologías que se van desarrollando y que mejoran la eficiencia de su trabajo. Esto va desde el aprendizaje de nuevos lenguajes de programación, hasta algoritmos más complejos que permiten un mejor modelado de datos.

Bibliografía

Alaminos, D. (2018) *Un Modelo Global de Predicción de Quiebra con Redes Neuronales*. Tesis para obtener el título de doctor en finanzas. Universidad de Málaga, Málaga.

Albon, C. (2018) *Python Machine Learning Cookbook.* California, Estados Unidos: O'Reilly Media.

Aldridge, I. (2019) Big Data in Portfolio Allocation: A New Approach to Successful Portfolio Optimization. *Journal of Financial Data Science*, 1 – 20.

Awad, M., Khanna, R. (2015) *Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers.* Nueva York, Estados Unidos: Springer.

Awoyemi, J., Adetunmbi, A. & Oluwadare, S. (2017) Credit card fraud detection using Machine Learning Techniques: A Comparative Analysis. *International Conference on Computing Networking and Informatics (ICCNI)*, 1 – 9.

Behazin, E., Misra, M. & Mohanty, A. (2017) Compatibilization of toughened polypropylene/biocarbon biocomposites: A full factorial design optimization of mechanical properties. *ScienceDirect*, 61, 364-372.

Blasina, F. (2019) *Procesamiento de señales acústicas aplicado al monitoreo de procesos.* Tesis para obtener el título de magister en ingeniería eléctrica. Universidad de la República, Montevideo.

Brito, P. & Diday, E, (1990) Pyramidal representation of symbolic objects. *NATO ASI Series*, 61, 3-16.

Burhum, R. (2020) Un Semáforo para el Huayno. Disponible 11 de noviembre de 2020. Recuperado de https://medium.com/@rburhum/un-sem%C3%A1foro-para-el-huayno-33588eb4db4b

Bustamante, S. (2014) *Algoritmos de procesamiento de imagen aplicados a la detección de figuras geométricas y sus propiedades espaciales.* Artículo, Pontificia Universidad Católica de Valparaíso, Valparaíso.

Cairo, A. (2012) *The Functional Art: An introduction to information graphics and visualization.* California, Estados Unidos: Pearson.

Caro-Hernández, P. & Tobar, J. (2020) Análisis Microbiológico de Superficies en Contacto con Alimentos. *Entramado*, 16(1), 240-249.

Castro, G., Hirakawa, A. & Martini, J. (2017) Adaptive traffic signal control based on bio-neural network. *ScienceDirect*, 109, 1182 – 1187.

Catanedo, F. (2015) *Data Preparation in the Big Data Era.* California, Estados Unidos: O'Reilly Media.

Chen, Q., Song, X., Yamada, H. & Shibasaki, R. (2016) Learning Deep Representation from Big and Heterogeneous Data for Traffic Accident Inference. *Association for the Advancement of Artificial Intelligence,* 338 - 344.

Cleff, T. (2013) *Exploratory Data Analysis in Business and Economics: An Introduction Using SPSS, Stata, and Excel*. Nueva York, Estados Unidos: Springer.

Cleophas, T., Zwinderman, A. (2013) *Machine Learning in Medicine: Part Three*. Nueva York, Estados Unidos: Springer.

Córdova, M. (2003) *Estadística: Descriptiva e Inferencial*. Lima, Perú: Editorial Moshera.

Davis, N., Raina, G., & Jagannathan, K. (2016) A multi-level clustering approach for forecasting taxi travel demand. *IEEE 19th International Conference on Intelligent Transportation Systems (ITSC),* 223 - 228.

DB-Engines (2020) DB-Engines Ranking - Trend Popularity. Disponible 20 de octubre de 2020. Recuperado de https://db-engines.com/en/ranking_trend

Dietrich, D., Heller, B., Yang, B. (2015) *Data Science Y Big Data Analytics.* Nueva Jersey,

Estados Unidos: John Wiley & Sons.

Donoho, D (2017) 50 Years of Data Science. *Journal of Computational and Graphical Statistics,* 26, 745-766. Obtenido de https://www.tandfonline.com/doi/full/10.1080/10618600.2017.1384734

Esteban, M., Bernardo, A., Tuero, E., Cervero, A. & Casanova, J. (2017) Variables influyentes en progreso académico y permanencia en la universidad. *European Journal of Education and Psychology*, 10(2), 75-81.

Finkel, R., Mercuri, E., Darras, B. et al. (2017) Nusinersen versus Sham Control in Infantile-Onset Spinal Muscular Atrophy. *The New England Journal of Medicine*, 377(18), 1723-1732.

Galeshchuk, S. (2016) Neural networks performance in exchange rate prediction. *Neurocomputing,* 172, 446 – 452.

Gartner (2020) Gartner Magic Quadrant for Analytics and Business Intelligence Platforms. Disponible 20 de octubre de 2020. Recuperado de https://www.gartner.com/en/documents/3980852/magic-quadrant-for-analytics-and-business-intelligence-p

Grus, J. (2015) *Data Science from Scratch.* California, Estados Unidos: O'Reilly Media.

Guo, Y., Wang, M. & Li, X. (2017) Application of an improved Apriori algorithm in a mobile e-commerce recommendation system. *Industrial Management & Data Systems*, 117(2), 287 – 303.

Herranz, R. (2014) *Bases de Datos NoSQL: Arquitectura y Ejemplos de Aplicación.* Tesis para obtener el título de ingeniero informático. Universidad Carlos III de Madrid, Madrid.

Horeweg, N., Scholten, E., de Jong, P. & et al. (2014) Detection of lung cancer through low-dose CT screening (NELSON): a prespecified analysis of screening test performance

and interval cancers. *The Lancet Oncology,* 15(12), 1341 - 1350.

Hsu, Y., He, Y., Ting, C., Tsou, M., Tang, G. & Pu, C. (2020) Administrative and Claims Data Help Predict Patient Mortality in Intensive Care Units by Logistic Regression: A Nationwide Database Study. *BioMed Research International,* 10, 1-10.

Irizarry, R. (2019) *Introduction to Data Science: Data Analysis and Prediction Algorithms with R*. Florida, Estados Unidos: CRC Press.

Jaworek-Korjakowska, J. (2016) Computer-Aided Diagnosis of Micro-Malignant Melanoma Lesions Applying Support Vector Machines. *BioMed Research International*, 6, 1 - 8.

Jeon, K., Goo, J., Lee, C. & et al. (2012) Computer-aided nodule detection and volumetry to reduce variability between radiologists in the interpretation of lung nodules at low-dose screening computed tomography. *Investigative Radiology*, 47(8), 457 - 61.

Komorowski, M., Marshall, D., Salciccioli, J., Crutain, Y. (2016) *Exploratory Data Analysis.* Nueva York, Estados Unidos: Springer.

Kouro, S. & Musalem, R. (2002) *Tutorial introductorio a la Teoría de Wavelet*. Artículo, Universidad Técnica Federico Santa María, Valparaíso.

Kumar, S. & Vanajakshi, L. (2015) Short-term traffic flow prediction using seasonal ARIMA model with limited input data. *European Transport Research Review,* 7, No 21.

Levine, D., Krehbiel, T., Berenson, M. (2012) *Estadística Descriptiva*. México D.F.: Pearson.

Martínez, C. (2012) *Estadística y muestreo*. Bogotá, Colombia: ECOE Ediciones.

Mailund, T. (2017) *Beginning Data Science in R: Data Analysis, Visualization, and Modelling for the Data Scientist*. Nueva York, Estados Unidos: Apress.

McCallum, Q. Ethan (2012) *Bad Data Handbook: Cleaning Up The Data So You Can Get Back To Work.* California, Estados Unidos: O'Reilly Media.

McKinney, W. (2017) *Python for Data Analysis*. California, Estados Unidos: O'Reilly Media.

Moussa, G. (2014) Vehicle Type Classification with Geometric and Appearance Attributes.

International Journal of Architectural and Environmental Engineering, 8(3), 273 - 278.

Mubarok, S., Adiwijaya & Aldhi, M. (2017) Aspect-based sentiment analysis to review products using Naive Bayes. *AIP Conference Proceedings*, 1867, 1-8.

Müller, A., Guido, S. (2016) *Introduction to Machine Learning with Python: A Guide for Data Scientists*. California, Estados Unidos: O'Reilly Media.

Osborne, J. (2013) *Best Practices in Data Cleaning: A Complete Guide to Everything You Need to Do Before and After Collecting Your Data*. California, Estados Unidos: SAGE Publications.

Peng, R., Matsui, E. (2016) *The Art of Data Science*. Estados Unidos: Lulu Press.

Pérez, A., Escobar, C. & Toledo, M. (2017) Modelo de predicción de la deserción estudiantil de primer año en la Universidad Bernardo O´Higgins. *Educação e Pesquisa*, 44.

Pierson, L. (2017) *Data Science for Dummies.* Nueva Jersey, Estados Unidos: John Wiley & Sons.

Pimpler, E. (2018) *Data Visualization and Exploration with R: A practical guide to using R, RStudio, and Tidyverse for data visualization, exploration, and data science applications*. Texas, Estados Unidos: CreateSpace.

Resendiz, G. (2017) *Filtros Wavelet óptimos para detección de ruido en imágenes SD y HD*. Tesis para obtener el título de ingeniero en telecomunicaciones. Universidad Nacional Autónoma de México, México D.F.

Rubin, G. (2015) Lung Nodule and Cancer Detection in CT Screening. *Journal of Thoracic Imaging,* 30(2), 130 - 138.

Sabater, Y., Molero, R. & Pla, L. (2010) Análisis Descriptivo de las Características de los Contactos de Menores con sus Familias Biológicas en los Acogimientos en Familia Ajena. *International Journal of Developmental and Educational Psychology*, 2(1), 229-236.

Sahinaslan, E. (2019) *Review of the Most Popular Data Science Programs Used Today: Python and R.* Artículo, Maltepe University, Estambul.

Sánchez-Corcuera, R., Casado-Mansilla, D., Borges, C. & López-de-Ipiña, D. (2020) Persuasion-based recommender system ensambling matrix factorisation and active learning models. *Personal and Ubiquitous Computing*, 1 – 11.

Selvamuthu, D., Das, D. (2018) *Introduction to Statistical Methods, Design of Experiments and Statistical Quality Control*. Nueva York, Estados Unidos: Springer.

Serov, V. (2017) *Fourier Series, Fourier Transform and Their Applications to Mathematical Physics.* Nueva York, Estados Unidos: Springer.

Singh, D. & Singh, B. (2019) Investigating the impact of data normalization on classification performance. *Applied Soft Computing,* No 105524.

Skiena, S. (2017) *The Data Science Design Manual.* Nueva York, Estados Unidos: Springer.

Stack Overflow (2017) The Incredible Growth of Python. Disponible 20 de octubre de 2020. Recuperado de https://stackoverflow.blog/2017/09/06/incredible-growth-python/

Stark, H. (2005) *Wavelets and Signal Processing***.** London, England: Pearson.

Tagliaferri, L., Morales, M., Birbeck, E., Wan, A. (2019) *Python Machine Learning Projects*. Nueva York, Estados Unidos: DigitalOcean.

Taylor, A. (2013) *SQL for Dummies*. Nueva Jersey, Estados Unidos: John Wiley & Sons.

Tukey, J. (1977) *Exploratory Data Analysis*. London, England: Pearson.

VanderPlas, J. (2016) *Python Data Science Handbook: Essential Tools for Working with Data.* California, Estados Unidos: O'Reilly Media.

Villada, F., Muñoz, N. & García-Quintero, E. (2016) Redes Neuronales Artificiales aplicadas a la Predicción del Precio del Oro. *Información Tecnológica*, 27(5), 143 – 150.

Wei, D., Geng, P., Ying, L. & Shuaipeng, L. (2014) A prediction study on e-commerce sales based on structure time series model and web search data. *The 26th Chinese Control*

and Decision Conference (2014 CCDC), 5346 – 5351.

Xu, D., Yip, R., Smith, J., Yankelevitz, D., Henschke, C. & et al. (2014) Retrospective Review of Lung Cancers Diagnosed in Annual Rounds of CT Screening. *AJR American journal of roentgenology,* 203(5), 965 - 972.

Zhao, Q., Zhang, Y., Friedman, D. & Tan, F. (2015) E-commerce Recommendation with Personalized Promotion. *Association for Computing Machinery*, 219–226.