Instituto Tecnológico y de Estudios Superiores de Monterrey
Escuela de Ingeniería y Ciencias
Ingeniería en Ciencias de Datos y Matemáticas

Gerardo Barajas Sánchez - *A00828760*

Miguel Alejandro Salas Reyna - *A00827219*

Iker Ledesma Durán - *A01653115*

Paola Balbuena Almanza - *A01652413*

Ethan Enrique Verduzco Pérez - *A01066955*

Profesores: Laura Hervert Escobar y María de los Angeles Constantino González

Socio Formador: Amador Lopez Hueros y Federico Valdez

Monterrey, Nuevo León
30 de abril de 2021

Índice

Resumen

La inteligencia de negocios permite que las empresas tomen decisiones correctas y oportunas, implementar este conjunto de procesos puede hacer la diferencia para que una empresa sea exitosa. CEMEX Ventures es una empresa que busca revolucionar el sector de la construcción y entre sus objetivos busca optimizar el proceso de fabricación de soportes de forma en que se reduzcan los costos de energías empleadas manteniendo una buena tasa de producción y de calidad. Para ello se implementó la metodología CRISP-DM para construir un modelo que en base a características de la fabricación del material permita predecir la calidad y el costo de su producción y así poder seleccionar aquellas que arrojen mejores resultados. Por último, se desarrolló una página web dónde se implementa el modelo Random Forest Classifier, que se espera funcione como guía para el área de producción de la empresa.

Palabras claves: Optimización, Energías, Machine Learning, Algoritmos de Aprendizaje automático, Minería de Datos, Ciencia de Datos.

1. Introducción

CEMEX Ventures es el capital riesgo corporativo de CEMEX, con un enfoque global, invierten en innovadoras startups de construcción para impulsar la revolución de la industria de la construcción. De la misma manera, buscan un mejor futuro a través de la industria de la construcción, al reunir los principales actores del ecosistema, como startups, emprendedores, universidades y otros grupos de interés. Uno de las principales ventajas competitivas que tiene CEMEX Ventures, es que emplean como uno de sus recursos principales de optimización, el uso de distintas técnicas de aprendizaje automático, con el fin de ampliar su panorama sobre sus producciones y poder determinar las mejores soluciones y alternativas para el negocio. [**cemex**]

El aprendizaje automático es la disciplina científica que se centra en cómo las computadoras aprenden de los datos. Surge en la intersección de la estadística, que busca aprender las relaciones a partir de los datos, y la informática, con su énfasis en los algoritmos informáticos eficientes. Este matrimonio entre las matemáticas y la informática está impulsado por los desafíos computacionales únicos de construir modelos estadísticos a partir de conjuntos de datos masivos, que pueden incluir miles de millones o billones de puntos de datos. Los tipos de aprendizaje utilizados por las computadoras se subclasifican convenientemente en categorías tales como aprendizaje supervisado y aprendizaje no supervisado.

2. Problema

Durante el proceso de fabricación, CEMEX Ventures se enfrenta a distintas problemáticas que lo limitan en función de obtener la mayor cantidad de ganancias, porque se encuentran en un dilema entre minimizar el uso de recursos y sobretodo los gastos por el costo de la energía calórica y energía eléctrica durante el proceso de manufactura, necesarios para cumplir los estándares de calidad en función con lo establecido por la industria. Por lo que la problemática principal que se enfrentan es encontrar la configuración óptima

del uso y gasto de energía eléctrica y calórica, tomando en cuenta las ventajas y desventajas durante el proceso de manufactura de cada una y sus respectivos costos, sabiendo que el precio de la energía calórica es aproximadamente 0.724 veces lo de la energía eléctrica.

2.1. Valoración de la situación

- Que conocemos:

  En este reto analizaremos un conjunto de datos reales para construir un modelo predictivo el cual definirá las condiciones óptimas de producción para minimizar el gasto energético, usando técnicas de aprendizaje automático.

  El proceso de producción actual de CEMEX se basa en la combinación de dos tipos de fuentes de energía para el método de creación:

  - Energía calórica: Obtenida mediante la quema de basura comprada por CEMEX de otras compañías.
  - Energía eléctrica: Generada por el uso de coque de petróleo, el cual el un residuo creado a partir de la producción de la gasolina.

  Ambas energías son diferentes en cuanto a eficiencia y costo. Siendo la energía eléctrica la mas cara y la calórica la mas barata.

- Que desconocemos:

  En este punto desconocemos la base de datos a utilizar y desconocemos el método de producción del polvo de cemento.

3. Antecedentes

3.1. Gerardo - Learning from Machine Learning in Accounting and Assurance.

En el libro "Learning from Machine Learning in Accounting and Assurance"muestran a la ciencia de datos como un algoritmo, como una secuencia de pasos de procesamiento estadístico. Cho, Soohyun; Vasarhelyi, Miklos A.; Sun, Ting; Zhang, ChanyuanEl presentan al Machine Learning como un método computacional el cual nos permite aprender sobre diversos patrones generados a partir de una base de datos. Estos procesos de aprendizaje y/o entrenamiento nos permiten hacer predicciones sobre eventos futuros. Hoy en día, existen miles de ejemplos de Machine Learning a nuestro alrededor. Asistentes digitales, sitios web, reproductores de música. Con cada interacción que tenemos con nuestros dispositivos electrónicos o con cada nuevo dato registrado de nuestro proyecto, la información se vuelve más precisa. Lo cual permite una mayor eficiencia en nuestra vida personal y laboral. [3]

3.2. Paola - Keeping a factory in an energy-optimal state

En el artículo Keeping a Factory in an Energy-optimal State los autores Wahren, S., Colangelo, E., Sauer, A., Mandel, J. y Siegert, J. presentan el proyecto ECOMATION por el que las empresas manufactureras buscan adoptar modelos enérgicamente eficientes y sostenibles de producción en base a un enfoque holístico que les permite analizar el rendimiento de máquinas individuales y procesos periféricos. La implementación de dichos modelos les permite no solo alcanzar un estado de energía óptimo sino también reducir costos. [5]

3.3. Iker - Statics review 14: Logistic Regression.

En el artículo "Logistic Regresion"de BiomedCentral, los autores Bewick, V. Cheek, L. y Ball, J. Destacan que la regresión logística proporciona un medio útil para modelar la dependencia de una variable de respuesta binaria en una o más variables explicativas, donde estas últimas pueden ser categóricas o continuas. Y que el ajuste del modelo resultante se puede evaluar mediante otros métodos. Este algoritmo se puede utilizar para varios problemas de clasificación, como la detección de spam, predicción de la diabetes, si un cliente determinado comprará un producto en particular o si se irá con la competencia, hay muchos más ejemplos en donde se puede aplicar este algoritmo. Por su parte la Regresión Logística lleva el nombre de la función utilizada en el núcleo del método, la función logística es también llamada función Sigmoide. Esta función es una curva en forma de S que puede tomar cualquier número de valor real y asignar a un valor entre 0 y 1. Si la curva va a infinito positivo la predicción se convertirá en 1, y si la curva pasa el infinito negativo, la predicción se convertirá en 0. Si la salida de la función Sigmoide es mayor que 0.5, podemos clasificar el resultado como 1 o SI, y si es menor que 0.5 podemos clasificarlo como 0 o NO. Por su parte si el resultado es 0.75, podemos decir en términos de probabilidad como, hay un 75 de probabilidades de que el paciente sufra cáncer. La técnica también se puede usar en ingeniería, sobre todo para la predicción de la probabilidad de fallo de un proceso, sistema o producto dado. También se utiliza en aplicaciones de marketing como la predicción de la propensión de un cliente a comprar un producto o suspender una suscripción, etc. En economía se puede usar para predecir la probabilidad de que una persona elija estar en la fuerza laboral, y una aplicación comercial sería predecir la probabilidad de que un propietario no pague una hipoteca . Los campos aleatorios condicionales, una extensión de la regresión logística a datos secuenciales, se utilizan en el procesamiento del lenguaje natural. [2]

3.4. Miguel - Optimización mediante técnicas de minería de datos del ciclo de recocido de una línea de galvanizado.

Francisco Javier Martínez de Pisón Ascacíbar nos presenta en su tesis la búsqueda que tienen las empresas por reducir los costos de producción y aumentar la calidad del producto sin reducir en gran magnitud la tasa de producción. Por lo que a lo largo de los años, se han desarrollado métodos y algoritmos que nos permiten

cumplir con esta función, como es el caso de Data Mining, método utilizado para la optimización del ciclo de producción de una línea de galvanizado. En dicha tesis, mediante técnicas de análisis multivariante y Minería de Datos, se intenta demostrar que con dichas técnicas es posible determinar las razones de los fallos en maquinaria y la línea de producción. [4]

3.5. Ethan - Workload Optimization and Energy Consumption Reduction Strategy

De acuerdo con el análisis de los datos históricos de los sistema de nube privada por parte de distintas empresas y organizaciones a nivel global en la industria manufacturera y de construcción, Xiaoqin Wang determinó que la nube privada en la industria manufacturera no solo podría hacer frente a los problemas de capacidad limitada sino también a los problemas de baja utilización de recursos y al excesivo uso de energías consumida en los procesos de producción, causado por la baja optimización de las herramientas y la falta de análisis correlacional de los atributos necesarios para la manufactura de los productos. A través de este artículo el autor explicó cómo, basado en la creación de la base de datos de gestión de la configuración (CMDB) y el sistema de monitoreo, y combinado con las características del aprovisionamiento de recursos de la nube privada, trató de proponer diferentes estrategias para optimizar el proceso de costos y gastos de los recursos de energía. Además, propone a través de la optimización, obtener valores económicos sólidos, como la mejora de la carga de trabajo, la reducción del consumo de energía y un ciclo de uso más largo. [6]

4. Objetivos

CEMEX Ventures, una empresa dedicada a la industria de la construcción, después de examinar los costes históricos en la fabricación de cemento, se ha puesto reducir el gasto del coste de la energía necesaria, es decir la suma de los costes asociados a Energía Eléctrica y Energía Calórica. Será necesario construir un clasificador entrenado en un determinado conjunto de datos óptimo (a construir igualmente), que devuelva la calidad estimada según la configuración en energía necesaria EE y EC, dureza, aspiración y tasa de producción deseada para producir las necesidades diarias de demanda de fabricación.

4.1. Objetivos del negocio

Después de llevar a cabo una plática de los antecedentes con el socio formador y de examinar los valores históricos de los costos en la manufactura y fabricación de los soportes, concluimos que el objetivo es reducir de manera significativa el gasto del coste de la energía utilizada, teniendo en cuenta que las 2 energías utilizadas son la eléctrica y la calórica, siendo esta última la energía de uso primario por su menor costo, sin embargo, causa un daño al medio ambiente. Por lo que es necesario optimizar su uso, de manera que se

encuentre la mejor configuración de energía con los datos proporcionados

A partir de dicho contexto, se tiene que llevar a cabo el diseño y construcción de un modelo a partir de algún método de aprendizaje supervisado que nos permita ingresar valores de entrada, tales como los atributos manejados en la bases de datos como la dureza, la aspiración, la tasa de producción, y la cantidad de energía empleada, para obtener como resultado la calidad esperada, a través de conjuntos de entrenamiento y prueba optimizados, que nos permitan modificar la cantidad de energía necesaria para cumplir con los estándares de calidad de producción y que nos permita reducir los costos de producción, utilizando la menor cantidad posible de recursos y maximizar las ganancias.

4.2. Objetivos de Data Mining

En relación a los objetivos de Data Mining, la meta principal de la elaboración de este reporte es, a través de técnicas de analítica, emplear funciones y estrategias de minería de datos utilizando como herramienta principal el lenguaje de programación Python, con el fin de poder diseñar y construir un modelo final, a partir de una comparación de distintos modelos de aprendizaje supervisado, que nos permitan primeramente tener un panorama más detallado de la problemática, a través de técnicas de agrupación o clasificación, para después poder llevar a cabo predicciones sobre posibles alternativas y soluciones de optimización de los costos por gastos de los dos usos de energía principales por parte de Cemex, los cuales son la energía calórica y la energía eléctrica, a través de nuevos valores de entrada que entreguen dichos resultados gracias a un entrenamiento y prueba de conjuntos de valores establecidos en el modelo.

- Crear un modelo computacional en Python que prediga la calidad del polvo de cemento.
- Obtener los datos de los materiales utilizados en la producción del polvo de cemento.
- El porcentaje de cada combustible primario y secundario que es utilizado para la producción del polvo de cemento.
- Minimizar los costos de la energía eléctrica y calórica.
- Encontrar la correlación existente entre las variables (dureza, tasa de producción, aspiración).

5. Enfoque

Para la elaboración de este reporte, se analizará un conjunto de datos reales con el propósito de poder llevar a cabo la construcción de un modelo predictivo finalizado a definir las condiciones óptimas de producción para minimizar el gasto energético, a través de distintas técnicas de aprendizaje automático.

Para poder implementar dicho modelo con técnicas de aprendizaje supervisado es necesario, primero que nada, llevar a cabo un proceso de limpieza de datos, en el que se identificarán valores atípicos y la tendencia que tienen los datos, con el fin de facilitar el uso de la base de datos. Después se llevará a cabo el preprocesamiento de datos, en donde se dividirá la base de datos en conjuntos de entrenamiento y prueba, con el fin de poder generar diferentes modelos de aprendizaje supervisado y no supervisado que nos permitan agrupar, clasificar y predecir nuevos valores de la base de datos a través de nuevos valores de entrada. Finalmente, se comparan los modelos utilizados en función de su eficiencia, con el fin de poder determinar qué modelo es el que mejor se adapta a nuestra base de datos y al tipo de información con el que se maneja la problemática.

5.1. Plan del proyecto

Para implementar un plan ideal es necesario conocer el proceso que deseamos optimizar, así como los datos disponibles para su evaluación y posterior análisis. Según la información proporcionada hasta el momento, se utilizan dos tipos de maquinaria, una con diésel y otra con energía eléctrica, los materiales cuentan con cierta dureza, aspiración, calidad y tasa de producción. Una vez comprendido el procedimiento, es necesario hacer la implementación del modelo que nos permitirá optimizar el proceso de producción para mejorar la calidad y disminuir el gasto.

Figura 1: Estructura del proyecto

6. Método

6.1. Entendimiento de los datos

6.1.1. Preparación de los datos

El conjunto de datos proporcionados por CEMEX Ventures necesitaba ser preprocesado antes de utilizarlo para construir modelos. Los objetivos de este proceso fueron hacer frente a valores perdidos, corregir valores erróneos, seleccionar atributos importantes y adaptar la base de datos a Python.

6.1.2. Descripción de los datos

- Tiempo: Variable categórica que es usada como índice y refleja el día en el que fueron producidos los productos.
- Dureza: Variable numérica que determina la resistencia final obtenida del producto
- Tasa Prod: Variable numérica que representa la tasa de producción.(un día en este caso)
- Asp: Variable numérica que representa el nivel de aspiración durante el proceso.
- Energía calórica: Variable numérica obtenida mediante la quema de basura comprada por CEMEX de otras compañías.
- Energía eléctrica: Variable numérica generada por el uso de coque de petróleo, el cual es un residuo creado a partir de la producción de la gasolina.
- Calidad: Variable numérica que no debe bajar del 0.039 por ciento para poder utilizar el producto.

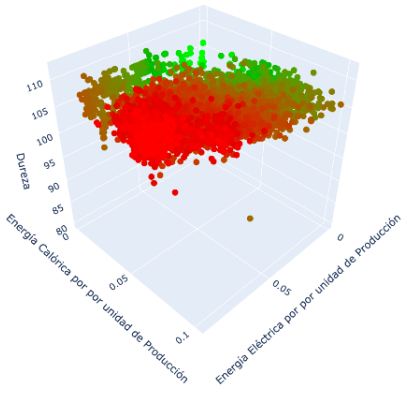

Figura 2: Energía eléctrica y calórica vs Dureza

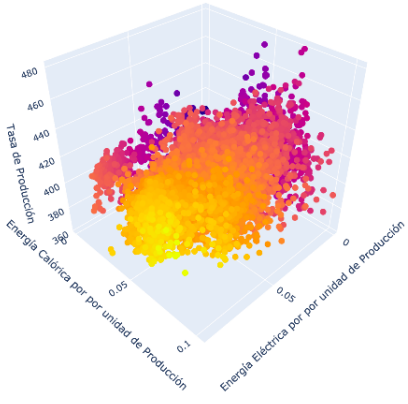

Figura 3: Energía eléctrica y calórica vs Tasa de Producción

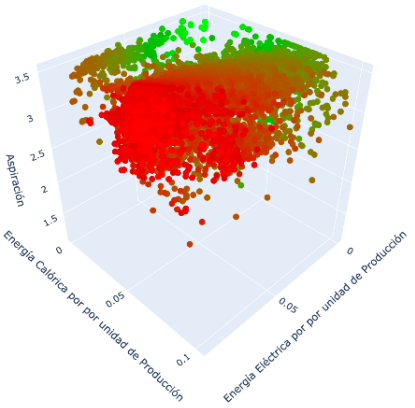

Figura 4: Energía eléctrica y calórica vs Aspiración

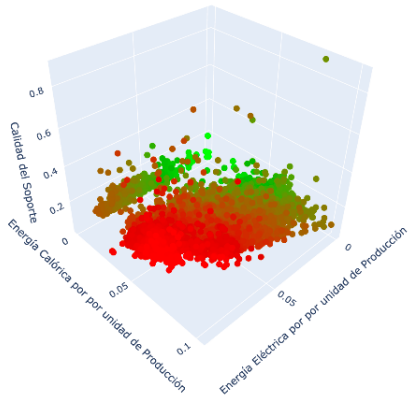

Figura 5: Energía eléctrica y calórica vs Calidad

6.1.3. Exploración de los datos

Una vez que ya contábamos con los valores listos e importados dentro de nuestro script de Python, se llevó a cabo el proceso de exploración de los datos, en los que se identificaron el tipo de datos con el que estábamos trabajando, cuántos valores nulos teníamos por columna y las dimensiones de la base de datos en general, a través del uso de funciones preestablecidas para el conocimiento de los datos del lenguaje de programación utilizado. Seguido de esto, se obtuvieron medidas estadísticas de los datos de cada uno de los atributos o columnas de la base de datos, de todas aquellas que cumplieran con el requisito de tener datos de tipo numérico, tales como medidas de tendencia central (la media, la desviación estándar), medidas de dispersión (mínimo, máximo) y medidas de posición no-central (cuartiles, valores atípicos).

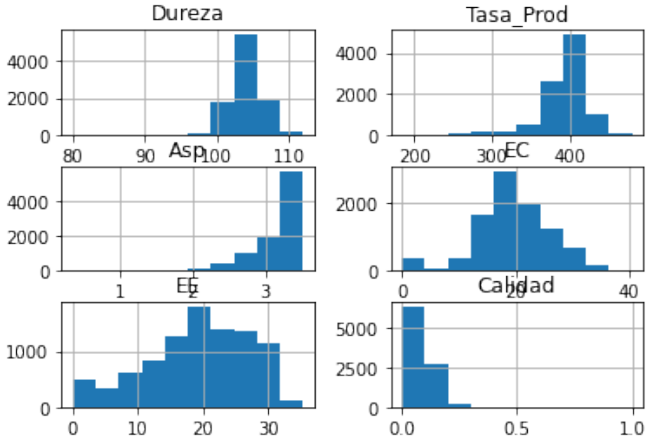

Figura 6: Histograma de los atributos de la base de datos

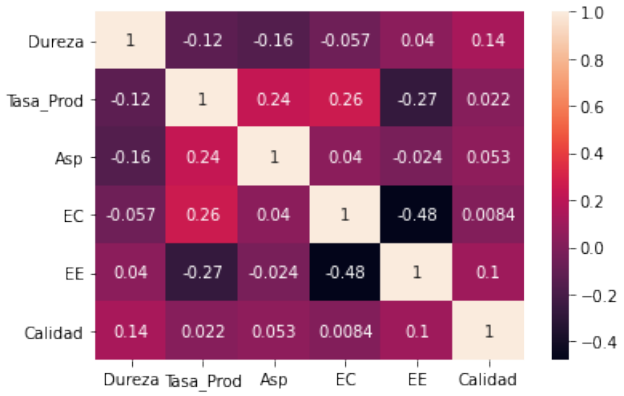

Figura 7: Matriz de correlación de los atributos de la base de datos

6.2. Preprocesamiento de los datos

6.2.1. Consolidacion de Datos

En este apartado se accedió y se recopilaron todos los datos que fueran necesarios y significativos, haciendo uso de distintas herramientas computacionales como el lenguaje de programación Python y distintas librerías con las que cuenta como Pandas, enfocadas en la importación y en el procesamiento de información y datos a nuestro script. Después se seleccionaron y filtraron los datos, en función de las medidas de estadística obtenidas anteriormente en la exploración de los datos, para finalmente integrar y unificar la información en un sola base de datos a tratar.

6.2.2. Limpieza de datos

Lo que se realizó en esta sección fue manejar los valores faltantes en los datos, tales como aquellos datos que aparecieran como nulos, como aquellos que fueron descritos como valor '0', pero no coincide con el resto de tipo de registros. En el caso de los valores nulos, en la base de datos se encontraron 2 registros con valores nulos, uno en los registros 60 y 62, los cuales les hacía falta el valor de Dureza y Aspiración respectivamente. Lo que se decidió hacer fue eliminar esos registros completos, porque no representaban una cantidad significa dentro de la base de datos. En relación con la tasa de producción, dentro de los la base de datos se encontraron en 68 registros donde el valor establecido era 0, por lo que una vez que fueron identificados esos registros también fueron eliminados de nuestro dataframe con todos los atributos.

En función a la detección de valores atípicos dentro de la base de datos, lo que se llevó a cabo fue establecer cuartiles para los atributos más importantes como la tasa de producción y la calidad esperada, en específico se seleccionaron su primer cuartil con el 30 % de la información total y el tercer cuartl con el 70 % respectivamente, para poder establecer un rango intercuartil, el cual nos permitió eliminar valores atípicos

a cierta distancia de los valores reales, para determinar con qué valores no queríamos contar, que hicieran nuestra base de datos más significativa y cumplieran con las condiciones de tener una calidad esperada mayor a 0.039, cumpliendo con los estándares de la empresa. Finalmente, se comprobó que sí se eliminaran dichos valores atípicos a través del uso de diagramas de caja de los atributos, que nos mostraran el antes y después de llevar los procesos de limpieza de nuestra base de datos, repitiendo de la misma manera algunos de los pasos llevados a cabo en la exploración de los datos para tener un entendimiento más detallado de la nueva información final.

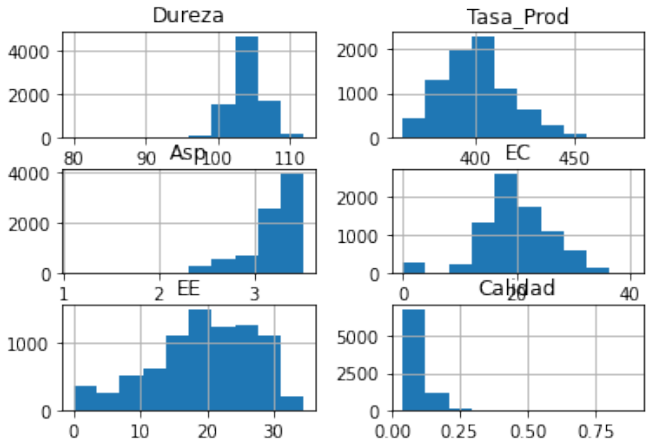

Figura 8: Histograma de los atributos después de la limpieza de los datos

6.2.3. Transformacion de datos

Durante el proceso de transformación de datos se calculó el costo de producción. EL cual consistió en la suma de los costos de energía, tanto eléctrica como calórica. Debido a que el costo de ambas energías es diferente, fue necesario establecer una relación entre ambos tipos. De esta manera, el cálculo del costo es $CostoTotal = EE + 0.724EC$. A continuación, se buscó una relación entre la tasa de producción y el costo, la energía calórica y la energía eléctrica. Por lo que se crearon tres columnas, las cuales corresponden al costo ponderado, la energía eléctrica ponderada y la energía calórica ponderada. El cálculo consistió en dividir la columna original entre la tasa de producción. Por ejemplo, la energía calórica ponderada se calcula $EEPonderado = EE/TasaProduccion$. Por último se añadió un nuevo atributo derivado de la variable Calidad, en este se transforman los datos numéricos de Calidad a categóricos y como resultado se obtienen 4 clases, Aceptable, que abarca valores de 0.04 a menores de 0.1, Regular, que abarca valores mayores a 0.1 y menores a 0.12, Bueno, que abarca valores entre 0.12 y 0.15, y Excelente, que abarca valores mayores a 0.15.

6.2.4. Reducción de datos

Realizamos un reducción de registros de diferentes valores atípicos por medio del establecimiento de cuartiles para los atributos más importantes como la tasa de producción y la calidad esperada, en específico se

seleccionaron su primer cuartil con el 30 % de la información total y el tercer cuartil con el 70 % respectivamente. Y por el método de valores nulos. El cual tan solo eliminó dos registros de nuestra base de datos. Esto volvió a nuestra base de datos más significativa y permitió que cumpliera con las condiciones de tener una calidad esperada mayor a 0.039, cumpliendo con los estándares de la empresa.

Después, se redujó la base de datos aún más, para que nos quedara el menor costo ponderado, uno de los nuevos atributos derivados, porque eso significaba que la tasa de producción era más alta. La manera en la que se llevó a cabo esto fue a través del uso de cuartiles nuevamente, filtramos los datos y decidimos quedarnos con los primeros 3 cuartiles de datos, porque todavía hasta ahí la media de los datos era muy parecida al valor máximo al 75 % de información, por lo que seguiríamos manteniendo la esencia de la base de datos y sus valores, pero reduciéndola al mismo tiempo de manera significativa, lo que será de mucha ayuda a la hora de llevar a cabo la modelación de los datos.

Posteriormente se hizo un análisis de correlación de los datos y se graficaron diagramas de dispersión de los diferentes tipos de energía contra las demás variables para detectar posibles relaciones entre los atributos de la base de datos.

7. Modelación de los datos

Se optó por usar modelos de clasificación para establecer la configuración adecuada de energías mientras se preserva una buena calidad y una alta tasa de producción. Para ello se estableció como variable objetivo Calidad y por lo tanto los valores se tuvieron que discretizar. Las categorías derivadas fueron 'Aceptable', 'Regular', 'Bueno' y 'Excelente'. Cada una de ellas se establecieron con un rango de calidad especifica.

Rangos de Calidad		
Categoría	Mínimo	Máximo
Aceptable	0.039	0.1
Bueno	0.1	0.12
Regular	0.12	0.15
Excelente	0.15	

Una vez creado este nuevo atributo se estableció como variable objetivo, y las variables de Dureza, Tasa de Producción, Aspiración, Energía Calórica y Energía Eléctrica se establecieron como las características que ayudarán a predecir las clases.

7.1. Técnicas de clasificación

Los algoritmos de aprendizaje automático supervisados definen modelos que capturan las relaciones entre los datos. La clasificación nos permite predecir la pertenencia de los datos a diferentes grupos. Es un área de aprendizaje automático supervisado que intenta predecir a qué clase o categoría pertenece una entidad, en función de sus características. Lo cual nos permite entrenar a los datos y tratar de predecir la clasificación de su grupo.

Las características o variables pueden adoptar una de dos formas:

Las variables independientes, también llamadas entradas o predictores, no dependen de otras características de interés (o al menos lo asume para el propósito del análisis). Las variables dependientes, también llamadas salidas o respuestas, dependen de las variables independientes.

Tenemos varios modelos de clasificación:

- DecisionTreeClassifier: El análisis del árbol de clasificación es cuando el resultado predicho es la clase (discreta) a la que pertenecen los datos.
- RandomForestClassifier: RandomForest es un algoritmo de aprendizaje supervisado. Se puede utilizar tanto para clasificación como para regresión.
- XGBClassifier: XGBClassifier consta de una colección de parámetros que se pueden aplicar en el ámbito global.
- SVC: La implementación se basa en libsvm. El tiempo de ajuste se escala al menos de forma cuadrática con el número de muestras y puede resultar poco práctico más allá de decenas de miles de muestras.
- LogisticRegression: LogisticRegression es un método estadístico de clasificación de objetos. Este capítulo ofrecerá una introducción a la regresión logística con la ayuda de algunos ejemplos.

7.2. Generacion de pruebas del diseño

De los 6445 registros de la base de datos el 60 % se usaron como datos de entrenamiento, 20 % como datos de prueba y el 20 % restante se utilizó como set de validación. Modelo por modelo los datos fueron ajustados e índices de precisión fueron calculados comparando las predicciones hechas por el modelo con la variable objetivo de los datos de prueba. Los datos se dividieron según el parámetro cv. Cada muestra pertenecía exactamente a un conjunto de pruebas y de predicciones para cada modelo. Con esa información se calculaba con un estimador ajustado de entrenamiento correspondiente.

7.3. Especificación de parametros a utilizar en el modelo

Para cada uno de los modelos anteriormente mencionados se hizo una selección de sus parámetros para optimizar su desempeño. Es necesario parametrizar un modelo para que su comportamiento se ajuste al problema que se está tratando. Para realizar dicha evaluación se utilizó el método de Búsqueda Grid. Esta estrategia crea y evalúa modelos con todas las posibles combinaciones de parámetros especificados y arroja la mejor.

Para el modelo de Regresión Logística los mejores parámetros fueron 'C': 1.0, 'dual': False y 'max iter': 100. El parámetro 'C' es un valor inverso a la fuerza de regularización que se aplicará en el modelo, es decir mientras más pequeño sea el valor mayor será la fuerza de regularización. El parámetro 'dual' hace formulación dual. El parámetro 'max iter' permite establecer el número máximo de iteraciones para que los solucionadores converjan.

Para el modelo de Árbol de Decisión los mejores parámetros fueron 'criterion': 'entropy', 'max depth': 3, 'max features': 2, 'min samples leaf': 4. El parámetro 'criterion' hace referencia al método que se utilizará para evaluar la calidad de las divisiones. El parámetro 'max depth' permite establecer la profundidad máxima deseada del árbol. El parámetro 'max features' permite establecer el número máximo de características a considerar cuando se busca elegir la mejor división. El parámetro 'min samples leaf' permite establecer el mínimo número de muestras necesarias para estar en un nodo hoja.

Para el modelo de Random Forest los mejores parámetros fueron: 'criterion': 'gini', 'max depth': 6, 'max features': 'auto', 'n estimators': 50. El parámetro n estimators construye diferentes regresores a partir del árbol de decisiones con un valor predeterminado de 100.

Para el modelo de Vectores de Soporte los mejores parámetros fueron 'C': 0.5, 'gamma': 'scale', 'kernel': 'linear'. El parámetro 'kernel' permite especificar el tipo de kernel que se desea utilizar en el algorítmo. El parámetro 'gamma' hace referencia al coeficiente del kernel, cuando se le da el valor de 'scale' se asigna como coeficiente el valor de la división de 1 entre el producto del número de características por la varianza del conjunto de características.

7.4. Evaluación de los modelos

Ya una vez habiendo realizado las pruebas de los distintos modelos utilizados, se realizó una comparación sobre cuál se acopla de una manera más óptima a la base de datos, todo esto basándonos en la precisión final obtenida porque el resto de métricas para llevar a cabo la evaluación eran muy similares entre todos los modelos. De acuerdo a su eficiencia o precisión obtenido por cada modelado, sin haber optimizado los hiper parámetros, en orden en el que se fueron aplicando los resultados obtenidos fueron: El modelo de Support Vector Classifier, obtuvo una precisión de 0.6757, de la misma manera, Logistic Regression tuvo una precisión

de 0.6758, a este le siguió el modelo de Decision Tree Classifier con un puntaje de 0.6703. Después se obtuvo el puntaje de precisión del modelo de Extreme Gradient Boost Classifier (XGB Classifier) con 0.6734. y finalmente lo mismo para el modelo de Random Forrest, el cual obtuvo el puntaje de precisión más alto con 0.6587.

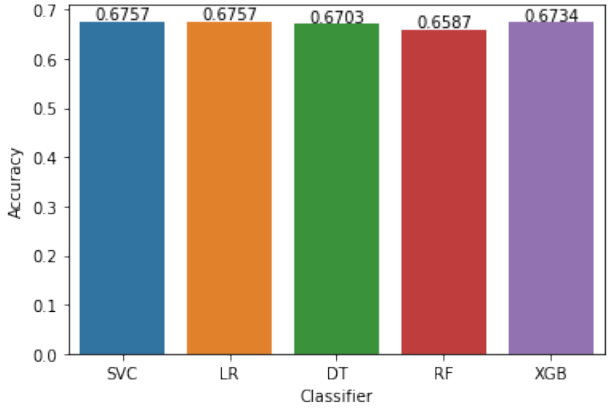

Figura 9: Precisión de modelos de clasificación antes de optimización de hiper parámetros

Después de que se llevó a cabo los cambios necesarios para optimizar los hiper parámetros, las evaluaciones dieron un cambio sumamente significativo, porque ahora que ya contábamos con la mejor versión de cada modelo según los datos utilizados, el clasificador que terminó teniendo la mejor precisión fue el Random Forest, con una precisión de 0.6861. A continuación se mostrará una gráfica que representa los resultados de precisión final de los modelos de clasificación utilizados, dejando de lado al XGB Classifier por la falta de información obtenida para encontrar una optimización de sus hiper parámetros y obteniendo resultados menores a lo de los otros modelos:

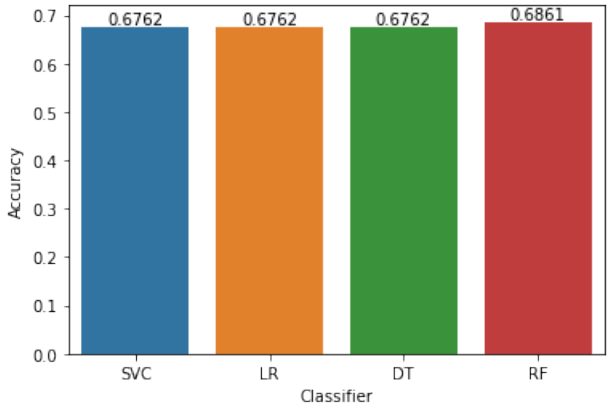

Figura 10: Precisión de modelos de clasificación después de optimización de hiper parámetros

De la misma manera, se utilizó el reporte de clasificación el cual es un visualizador de informes de clasificación muestra las puntuaciones de precisión, recuperación, F1 y soporte del modelo. Para facilitar la interpretación y la detección de problemas, el informe integra puntuaciones numéricas con un mapa de calor

codificado por colores. Todos los mapas de calor están en el rango para facilitar la comparación de modelos de clasificación en diferentes informes de clasificación [7]. Además, se hace uso de la matriz de confusión, la cual es una herramienta que permite visualizar el desempeño de un algoritmo de aprendizaje supervisado. Cada columna de la matriz representa el número de predicciones de cada clase, mientras que cada fila representa a las instancias en la clase real, es decir, en términos prácticos nos permite ver qué tipos de aciertos y errores está teniendo nuestro modelo a la hora de pasar por el proceso de aprendizaje con los datos [1]. A continuación se mostrarán los resultados de dichos procesos de evaluación:

Super Vector Classifier				
Calidad	precision	recall	f1-score	support
Aceptable	0.68	1.00	0.81	871
Bueno	0.00	0.00	0.00	3135
Excelente	0.00	0.00	0.00	82
Regular	0.00	0.00	0.00 201	
accuracy			0.68	1289
macro avg	0.17	0.25	0.20	1289
weighted avg	0.46	0.68	0.54	1289

Logistic Regression				
Calidad	precision	recall	f1-score	support
Aceptable	0.68	1.00	0.81	871
Bueno	0.00	0.00	0.00	135
Excelente	0.00	0.00	0.00	82
Regular	0.00	0.00	0.00	201
accuracy			0.68	1289
macro avg	0.17	0.25	0.20	1289
weighted avg	0.46	0.68	0.54	1289

Decision Tree				
Calidad	precision	recall	f1-score	support
Aceptable	0.68	1.00	0.81	871
Bueno	0.00	0.00	0.00	135
Excelente	0.00	0.00	0.00	82
Regular	0.00	0.01	0.03	201
accuracy			0.68	1289
macro avg	0.27	0.25	0.20	1289
weighted avg	0.46	0.68	0.54	1289

Random Forest				
Calidad	precision	recall	f1-score	support
Aceptable	0.68	1.00	0.81	871
Bueno	0.00	0.00	0.00	135
Excelente	1.00	0.02	0.05	82
Regular	0.12	0.00	0.01	201
accuracy			0.67	1289
macro avg	0.45	0.26	0.22	1289
weighted avg	0.54	0.67	0.55	1289

8. Implantacion de la solución

8.1. Planeación de la implantación

La empresa especificó que como interfaz de usuario se desarrollara una página web utilizando Flask, por lo que se decidió a través de un sistema interactivo con el usuario que necesite obtener predicciones de los valores de calidad del proceso, en relación a ciertos parámetros o atributos de entrada, los cuales son la energía eléctrica, energía calórica, dureza, aspiración y la tasa de producción deseada. Para evitar errores de clasificación, en la página web se presentará una descripción de los atributos a elegir, algunos ejemplos de cómo actuaría el modelo en relación a costos y calidad para cada atributo, además de tener limitado el rango de posibles entradas para cada valor, de acuerdo a los valores mínimos y máximos históricos otorgados a nosotros en la base de datos original. Finalmente, la página web otorgará resultados de clasificación de la calidad estimada, el costo calculado y la precisión del modelo empleado.

8.2. Ejecución de la implementación

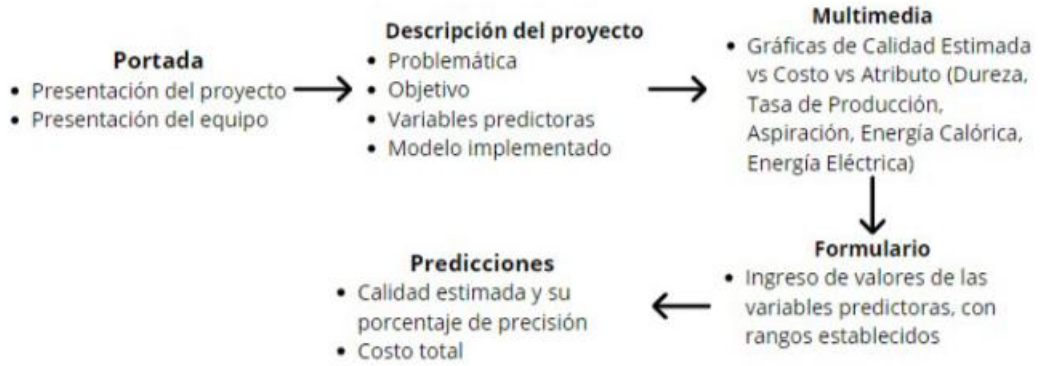

Figura 11: Estructura de la página web

8.3. Monitoreo de la implementación

Con cada avance o con cada nuevo dato presentado en nuestra base de datos, es necesario actualizar el programa debido a las incógnitas que pueden producir. En casos en los cuales la calidad tenga otro estándar mínimo para ser aceptable, que no nos basemos en la calidad como variable objetivo sino en el costo. Estos casos que afectan tanto nuestro programa, como el funcionamiento y propósito de nuestra pagina. Para una página web, mantener su operatividad en todo momento es indispensable. El tiempo de inactividad en tu sitio web puede traerte problemas tanto monetarios, como de productividad. Estos errores pueden darse por factores previsibles; como software y hardware obsoletos, o por cuestiones incontrolables. Si, nosotros como desarrolladores de la pagina de una empresa. Esto podría generar: Inconformidad en los usuarios, Desprestigio en la imagen de tu empresa, Pérdida de dinero e información, etc. Al monitorear tu sitio web, podrás evitar el tiempo de inactividad y las consecuencias que esto conlleva.

9. Conclusiones

La inteligencia de negocios permite acceder de forma interactiva a los datos, manipularlos y analizarlos, así como a situaciones y rendimientos tanto históricos como actuales para convertirlos en información valiosa que permita a las organizaciones tomar decisiones más informadas y mejores. Los datos se transforman a información, luego a decisiones y lo ideal es que en base a estas se lleven a cabo acciones que optimicen los procesos del negocio. En este proyecto se pretendía optimizar el uso de energías en la producción de soportes sin escatimar en la tasa de producción o en la calidad del material, pero este proceso de analítica de datos puede ser empleado en cualquier área de la empresa. El entorno empresarial puede ser muy complejo y dinámico, se ha convertido en una necesidad saber responder y adaptarse a un entorno tan cambiante y hacer uso de la analítica y de la ciencia de datos para ello es la mejor manera en que una empresa puede evolucionar y sobrevivir.

9.1. Gerardo Barajas

La ciencia de datos y los modelos de clasificación junto con el ámbito del aprendizaje automático (Machine Learning) y de la Inteligencia Artificial que partiendo de una base de datos, crea diagramas de construcciones lógicas que nos ayudan a resolver problemas. A pesar de las precisiones similares entre todos los modelos de clasificación, al final el Random Forest fue el ganador después de que se optimizaron los hiperparametros. Esto sucedió gracias al funcionamiento del modelo. Distintos árboles ven distintas porciones de los datos. Ningún árbol ve todos los datos de entrenamiento. Esto hace que cada árbol se entrene con distintas muestras de datos para un mismo problema. De esta forma, al combinar sus resultados, unos errores se compensan con otros y tenemos una predicción que generaliza mejor.

9.2. Paola Balbuena

Ha sido muy interesante ver cómo la ciencia de datos se aplica en el entorno empresarial. Para este proyecto revisamos diferentes modelos con los que puedes tratar una problemática, cada uno de ellos con diversas características y también consideramos diversas métricas de evaluación y métodos en las que los hiperparámetros de los modelos pueden ser optimizados. Para este caso se implementó un modelo de Random Forest Classifier después de hacer una búsqueda grid para considerar un modelo más eficiente, su precisión resultó ser del 68 %. Cada situación necesita especificaciones distintas de acuerdo a sus objetivos y al trabajar para cumplir los de este proyecto no solo aprendimos sobre procesos de aprendizaje automático sino también sobre los pasos a seguir, establecidos en la metodología CRISP-DM, y lo necesario para implementar inteligencia de negocios.

9.3. Miguel Salas

Mediante una estrategia basada en el modelo de negocios CRISP-DM para proyectos de minería de datos fue posible realizar el entendimiento, planeación, modelación e implementación del problema presentado por CEMEX Ventures. Para ello fue necesario realizar un entendimiento de los datos y su limpieza. Durante este proceso nos dimos cuenta que los datos presentados no eran los mejores para relizar una predicción aceptable por lo que en la etapa de modelación hubo muchos problemas para encontrar el mejor modelo. En el caso presente, destacó el modelo Random Forest que nos permite desarrollar una variedad de árboles de decisión para elegir al final el más aceptable para predecir resultados, teniendo una presición de casi el 70 %. Considero que con el tiempo tenido para la realización del proyecto se obtuvo un resultado aceptable, por lo que con un poco más de tiempo y sobre todo, mejores y más datos, se puede tener un mejor resultado.

9.4. Iker Ledesma

Se puede afirmar que, de los mejores métodos para poder realizar un proyecto de negocios es el CRISP-DM. Ya que con este se busca entender a detalle el negocio en el que se va a participar, conocer y preparar tus datos para que a la hora de modelarlos puedas obtener las mejores predicciones posibles. Por ello, es indispensable utilizar más de un modelo de predicción, ya que así puedes comparar cuál es el más útil para la meta que quieres lograr. En nuestro caso fue el Random Forest, el cuál nos brindó una precisión de .68 mientras que las demás no. Quizás si hubieras intentado utilizar algunos métodos adicionales se hubiera logrado encontrar uno con una mayor precisión. Sin embargo, a falta de tiempo no pudo ser posible. A pesar de eso, se pudo satisfacer las necesidades del cliente con los métodos utilizados y por lo tanto desarrollar y concluir este proyecto de manera exitosa.

9.5. Ethan Verduzco

En relacion la investigacion e implementacion de diferentes metodos de aprendizaje automático, sin duda alguna puedo afirmar que me llevo una experiencia sumamente enriquecedora tanto en mi crecimiento personal como profesional, dado que mi conocimiento previo sobre este tipo de metodos era practicamente nulo, por lo que toda la investigacion realizada y la informacion obtenida significaba un proceso muy emocionante y retador, pero sobre todo la implementacion practica con datos y problematicas reales es lo que terminaron de conjugar todas estas experiencias como uno de mis primeros proyectos relacionados a los metodos de agrupamiento de datos en especifico. Por otra parte, en terminos mas especificos relacionados a la programacion y manejo de base de datos, considero que este proyecto me a ser de muchisima utilidad en un futuro, porque poco a poco me permite ir completando un portafolio con investigacion, implementaciones de codigo y reportes escritos sobre los temas relacionados a los metodos de aprendizaje automático, los cuales aumentan en importancia y uso dia con dia por lo precisos que son muy modelos y la facilidad de emplearlos una base que se fundamentan las bases sobre su utilidad. De igual manera, me parece sumamente interesante e importante el uso y su correcta implementación de la metodología CRISP, definitivamente significativamente a estructurar y definir un orden más claro a la hora de llevar a cabo la resolución del proyecto.

10. Conclusión de plática de Philip Evans: Cómo los datos van a transformar los negocios

En los inicios de la inteligencia de negocios se tenía la idea que dos factores afectan el éxito de una empresa. Uno de ellos era la escalabilidad, a mayor volumen y concentración de masas, mayor recepción de ingresos y experiencia. El otro, las transacciones, es decir, el valor de la cadena de suministro que propiciaba una fortaleza entre los agentes que pertenecían a ella. La llegada del internet que trajo consigo una revolución digital y de la información nos hizo darnos cuenta que eso no era así. Uno podría simplemente moverse entre las cadenas de suministros e incluso, algunas partes de la cadena fueron eliminadas. Muchos negocios se vieron afectados por la revolución tecnológica que trajo el internet donde los primeros años causó una deconstrucción de la inteligencia de negocios y su segunda etapa destacó por la participación de los usuarios, donde cada uno podía ser un creador de contenido en la web. Al principio del milenio, hubo una explosión en la creación de información, lo que ha permitido que la tercera etapa del internet se caracterice por el uso de los datos. El uso de grandes cantidades de datos nos ha dado la posibilidad de reducir costos y encontrar patrones que no habíamos visto antes. Es increíble como hace 15 años, la codificación del genoma humano costaba millones de dólares y actualmente se reduce a unos cientos cientos de dólares. El Big Data es un gran paso no solo en los negocios, sino en todos los campos. Los avances científicos y tecnológicos que habrá en los siguientes años será una revolución.

Apéndice A Reflexiones Tqueremos - Etiquetas que estorban

A.1 Reflexión Ethan Verduzco

Uno de los mayores obstáculos que tenemos hoy en día como sociedad es lo polarizados que nos encontramos, por la creación y el manejo inadecuado de etiquetas, las cuales te fuerzan a pertenecer o a dejar a formar parte de algo por criterios de terceros y en la gran mayoría de los casos circunstancias ajenas a las posibilidades de uno. La mayor reflexión que me llevo de la plática de "Tqueremos- Etiquetas que estorban", es que tenemos que dejar de lado todos estos prejuicios que nos obligan a etiquetarnos y a ejercer nuestro criterio de cierta forma, hay que aprender a valorarnos a nuestro entorno por lo que somos y entender que funcionamos mejor unidos, que separados bajo estándares o etiquetas.

A.2 Reflexión - Miguel Salas

Muchas veces tenemos prejuicios de las personas y nos hacemos una idea de cómo son sin siquiera conocerlas. Suele ser común que al preguntarle a un amigo sobre una persona nos de una descripción con la cual tenemos una percepción de la persona en cuestión. Sin embargo, muchas veces estas ideas son fuera de la realidad y son más bien, la percepción que nuestro amigo se formó. Por lo cual, al conocer a la persona, nos damos cuenta que es realmente diferente a cómo nos la imaginábamos. Por lo cual, las etiquetas que le pusimos a esa persona son totalmente erróneas. Otro ejemplo es cuando vamos a presentar un examen o resolver una tarea que recibimos el comentario que la tarea en cuestión es muy difícil, lo cual nos causa miedo y estrés. Cuál es nuestra sorpresa que al momento de presentar el examen nos damos cuenta que realmente no era tan difícil, por lo cual nuestro estrés y preocupación fue exagerada ante la verdadera dificultad de la tarea. Es recomendable dejar un poco de lado los prejuicios y ser nosotros mismos ante un verdadera encuentro quienes formemos nuestra opinión y percepción. De esta manera nos damos la oportunidad de conocer mejor a las personas o por lo menos de conocerlas de una forma diferente.

A.3 Reflexión - Iker Ledesma

En nuestra vida debido a experiencias u opiniones de terceros, solemos tener ideas erróneas de situaciones, gente o lugares. Otras personas nos pueden decir su punto de vista sobre algo y debido a la falta de información que tenemos nos quedamos con lo que nos dijeron y creamos etiquetas que en ocasiones no son precisas. Por ello, la mayor reflexión que me llevo de la plática es la importancia de tener un criterio propio. No dejar que las etiquetas que tú mismo le pusiste a algo te impida poder disfrutarlo o en todo caso enfrentarlo. Siempre hay que tener el coraje y la valentía necesaria para poder desmentirlas y no conformarnos con lo que "sabemos". Ya que si vivimos gobernados por estas, nunca podremos disfrutar la vida de una manera que nos deje satisfechos. Es por eso que, siempre hay que estar dispuestos a intentar cosas nuevas con una

mentalidad abierta al cambio.

A.4 Reflexión - Paola Balbuena

Estamos acostumbrados a etiquetar personas, cosas o situaciones muchas veces sin haber convivido o experimentado anteriormente con ellas. Cuando llega el momento de enfrentarnos a ellas, puede ser que lo hagamos de forma limitada por los juicios erróneos que nos hayamos hecho de ellas. Los prejuicios son barreras que estorban y que no permiten crear relaciones ni convivencias sanas ni enriquecedoras. Es importante aprender a darle un nuevo enfoque a las cosas en el que removamos esas etiquetas que muchas veces no nos permiten disfrutar lo que hacemos o estar agradecidos por lo que tenemos.

A.5 Reflexión - Gerardo Barajas

Una persona crea estereotipos sin pensarlo al instante de conocer a una nueva persona. Esto puede llegar a depender de la apariencia, nacionalidad, cultura, etc. Estos prejuicios que generamos cada día de nuestras vidas tienen consecuencias y crean escenarios negativos los cuales no podemos evitar. Esto me hace pensar hasta que punto hemos llegado y no podemos recordar lo que nos hace iguales, pero si lo que nos hace diferentes. Nuestro propósito de vida debe estar alineado a la construcción de una sociedad más diversa e inclusiva, donde quepamos todos, donde mis derechos y el del otro sean aceptados.

Apéndice B Aportaciones

B.1 Ethan Verduzco

Dentro de las aportaciones del autor se encuentran la introducción, Problema, antecedentes (artículo Workload Optimization and Energy Consumption Reduction Strategy), Objetivos, Enfoque, exploración de los datos, consolidación de los datos, evaluación de modelos y planeación de la implantación.

B.2 Gerardo Barajas

Dentro de las aportaciones del autor se encuentran la Valoración de situación, antecedentes(articulo Learning from Machine Learning in Accounting and Assurance), Creación de tablas de resultados de modelos Técnicas de Clasificación, Monitoreo de implementación, Especificación de parametros a utilizar en el modelo ,Técnicas de clasificación, Conclusión personal.

B.3 Iker Ledesma

Dentro de las aportaciones del autor se encuentran: Antecedentes (artículo Logistic Regression), Evaluación de Modelos, conclusión personal.

B.4 Paola Balbuena

Dentro de las aportaciones del autor se encuentran abstract, concluión general, diagramas en plan de proyecto e implantación, Especificación de parámetros a utilizar en el modelo, artículo Keeping a factory in an enegy-optimal state, limpieza y transformación de datos.

B.5 Miguel Salas

Dentro de las aportaciones del autor se encuentran: Antecedentes (artículo Optimización mediante técnicas de minería de datos del ciclo de recocido de una línea de galvanizado), la modelación de los datos, transformación de datos y conclusión de la plática de Philips Evans.

Referencias

[1] Juan Arce. *La matriz de confusión y sus métricas – Inteligencia Artificial –*. 2021. URL: https://www.juanbarrios.com/la-matriz-de-confusion-y-sus-metricas/.

[2] V. Bewick. "Logistic regression." En: *Statistics review 14:* (2015), pág. 112. URL: https://doi.org/10.1186/cc3045.

[3] C Cho S Zhang. "Learning from Machine Learning in Accounting and Assurance." En: *Journal of Emerging Technologies in Accounting* (2020), págs. 17-49. URL: https://0-doi-org.biblioteca-ils.tec.mx/10.2308/jeta-10718.

[4] F Martínez de Pisón. "Optimización mediante técnicas de minerı́a de datos del ciclo de recocido de una l ınea de galvanizado (Doctorado". En: *Universidad de la Rioja* (2017). URL: https://dialnet.unirioja.es/servlet/tesis?codigo=81.

[5] S Wahren y E Colangelo. "Logistic regression." En: *Procedia CIRP* (2016), págs. 40-55. URL: https://doi.org/10.1016/j.procir.2016.01.053.

[6] Xiaoqin Wang, Ming Lu y Youyan Wang. "Workload Optimization and Energy Consumption Reduction Strategy of Private Cloud in Manufacturing Industry." En: *2020 IEEE 11th International Conference on Software Engineering and Service Science (ICSESS), Software Engineering and Service Science (ICSESS), 2020 IEEE 11th International Conference on* (2020), págs. 440-444. ISSN: 978-1-7281-6578-3.

URL: http://0-search.ebscohost.com.biblioteca-ils.tec.mx/login.aspx?direct=true&db=edseee&AN=edseee.9237662&lang=es&site=eds-live&scope=site.

[7] YellowBrick. *Classification Report*. 2020. URL: https://www.scikit-yb.org/en/latest/api/classifier/classification_report.html.