**Proyecto de curso Intro. Ciencia de Datos**

**Docente:**

**Lyda Peña**

**Integrantes:**

**Juan Camilo Javela Valencia - 2176484**

**Juan Manuel Perdomo -2170338**

**Sebastián Sanchez – 2175528**

**Karen Arcila Mejía - 2195005**

**Universidad Autónoma de Occidente**

**Facultad de ingeniería**

**2021**

**Dataset principal:**

**Dataset 1:** La base de datos conocida como Faostat, en la cual se registran estadísticas de cultivos para 173 productos, que cubren las siguientes categorías: cultivos primarios, cultivos de fibra primarios, cereales, grano grueso, cítricos, frutas, yute, fibras similares al yute, equivalente de tortas oleaginosas, cultivos oleaginosos primarios, legumbres, raíces y tubérculos, nueces de árbol y Verduras y Melones. Los datos se expresan en términos de área cosechada, cantidad de producción y rendimiento. El objetivo es cubrir de manera integral la producción de todos los cultivos primarios para todos los países y regiones del mundo. Cereales: Los datos sobre la superficie y la producción de cereales se refieren a cultivos cosechados únicamente para grano seco. Por lo tanto, se excluyen los cultivos de cereales recolectados para heno o recolectados en verde para alimento o utilizados para pasto. Los datos de superficie se relacionan con el área cosechada. Algunos países informan solo el área sembrada o cultivada.

**Dataset complementario:**

**Dataset 2:** La FAO tiene como objetivo principal ser pionero en las actividades internacionales que están encaminadas a erradicar el hambre en el mundo. En el dataset “Cultivos y productos de ganadería” se recopila, procesa y difunde el conjunto de datos sobre el comercio agrícola y alimentario de acuerdo con la metodología estándar de estadísticas del comercio internacional de mercancías (IMTS). Los datos se expresan en base a 4 elementos seleccionados en términos de las importaciones y exportaciones, los cuales son (Importaciones-cantidad, importaciones valor, exportaciones-cantidad, exportaciones valor ). Esta fuente de datos se verifica en busca de valores atípicos, los datos de los socios comerciales se utilizan para los países que no informan o las celdas faltantes, y los datos sobre la ayuda alimentaria se agregan para tener en cuenta los flujos comerciales transfronterizos totales.

**Preprocesamiento de los datos.**

Primero se descargaron los datasets de la página web FAOSTAT, posteriormente se cargaron al Power BI. Se puso en formato UTF-8 para que reconociera tildes y la letra “ñ” del abecedario español. En la siguiente imagen se da un ejemplo del dataset de producción cargado en el Power BI.

Tabla

Descripción generada automáticamente con confianza media

Puesto que la información se busca orientar hacia un público que se desempeña en el mercado alimenticio, se quitaron todas las columnas innecesarias para el trabajo, que no tuvieran relevancia en el análisis. En total se eliminaron 10 columnas que tenía información repetida o que no tuviera algún significado relevante. En la siguiente imagen se muestra el cambio al eliminar dichas columnas irrelevantes.



****

Se muestra también un ejemplo de columnas que no aportan valor al dataset.

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Esta columna, se elimina ya que solo contiene el valor de QI para toda la columna, haciendo que no aporte nada para el análisis. Como esta, hay otras 9 que solo tienen valores que se repetían.

Adicionalmente, ya que se tienen más de 50 productos, se decidió escoger 9 que tuvieran relevancia para el país de Colombia y que también se cultivaran en otros países de América para hacer posteriores análisis. Se muestra a continuación los productos seleccionados para el análisis.

Tabla

Descripción generada automáticamente

También se modificaron las columnas que tuvieran un tipo de dato que no funcionara para el análisis, se muestra entonces en la siguiente imagen, la comparación del cambio de Año.

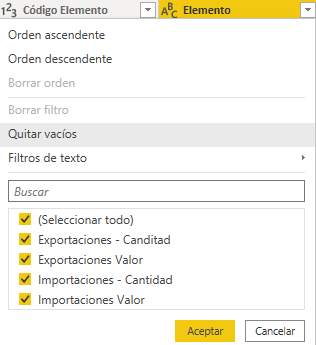




Se cambio el tipo numérico a texto, para que en las graficas no lo tomara como un valor al que se le pueda hacer un cálculo, si no que fuera como un valor único.

.

Por último, en el segundo dataset, además de aplicar las mismas técnicas anteriores, este tenía los datos de exportaciones e importaciones, por valor y cantidad, puestos en el mismo dataset, por lo que se decidió separar estos atributos en tablas distintas para facilitar la manipulación de estos.



Esto se hizo copiando y pegando la misma tabla y filtrando cada selección. Se muestra un ejemplo de tabla creada a partir de la filtración de Exportaciones - Cantidad

Interfaz de usuario gráfica, Texto, Aplicación

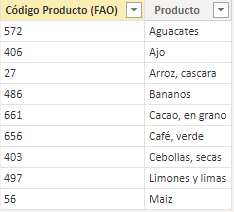
Descripción generada automáticamente

Así se hizo con las otras 3 tablas generando, el modelo que se muestra a continuación.

Interfaz de usuario gráfica, Diagrama, Aplicación

Descripción generada automáticamente

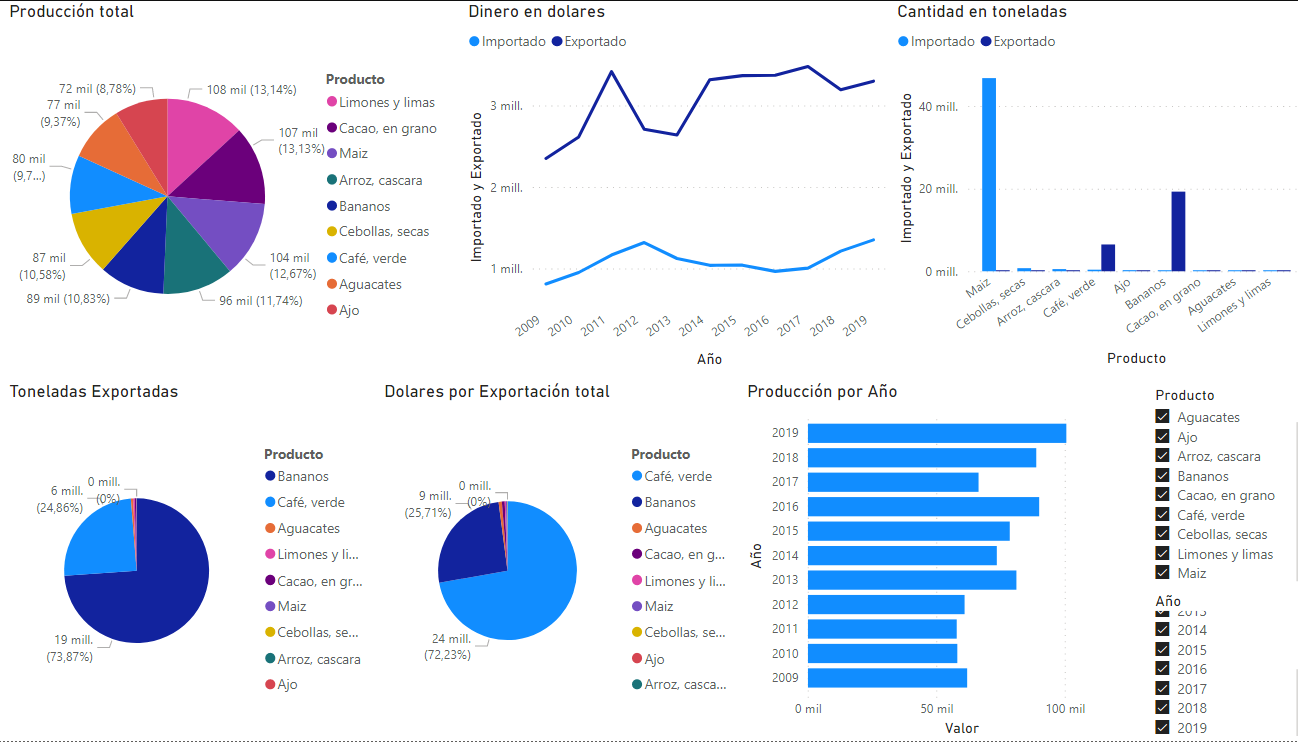
Se creó, por último, 2 tablas adicionales para que el modelo se pudiera unir y se creara las graficas que se mostraran en el dashboard. Se muestra a continuación la tabla que contiene los productos con su ID único y la tabla que contiene los años.

Imagen que contiene luz, calle, reloj, estacionado

Descripción generada automáticamente

**Análisis predictivo.**

Ahora, se muestra el dashboard creado para esta primera parte del proyecto. se muestra el dashboard creado para esta primera parte del proyecto. Este dashboard puede estar dirigido a diferentes personas que trabajan en el sector de la alimentación. Se puede hacer comparaciones de alimentos producidos versus alimentos exportados y con esto poder impulsar a campesinos a que produzcan este alimento que más se exporta, para así poder apoyar a los cultivadores.



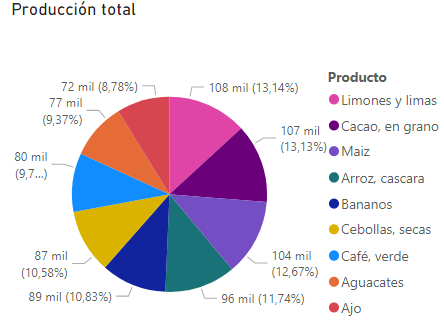
Este dashboard contiene 6 gráficas, 3 de pastel ,1 histogramas, 2 gráficos de barras y 2 filtradores de datos, 1 para los productos y otro para los años. Los seleccionadores son para hacer análisis para datos más específicos o comparaciones que se requieran.

Interfaz de usuario gráfica, Gráfico, Aplicación

Descripción generada automáticamente

Por ejemplo, se muestra en la figura de arriba, que solo se seleccionaron los 3 productos menos producidos durante un periodo de 5 años (2014-2019). Con esto se puede hacer múltiples selecciones y hacer análisis dependiendo de a quien vaya dirigido la información.

El objetivo de este primer gráfico, es el de mostrar la producción de los productos escogidos. Este gráfico se muestra en toneladas



Complementando al anterior, se muestra el siguiente gráfico, y tiene como objetivo comparar la producción anual que han tenido estos alimentos durante los 10 años escogidos.

Gráfico, Gráfico de barras

Descripción generada automáticamente

Ahora se tienen las exportaciones en toneladas de los productos. En la siguiente imagen puede verse que el Banano y el Café se exportan mucho a comparación con los otros productos, pero con el seleccionador, podemos quitarlo para ver como se comportan los otros productos.

Gráfico, Gráfico circular

Descripción generada automáticamente

Complementario al anterior, este gráfico muestra una comparación entre la cantidad de dólares exportados por estos mismos productos. (Nótese, que, aunque se exporte mucho más banano, el dinero recibido en proporción es mucho mayor el del café, el cual se puede decir que tiene un valor en dólares mucho más alto)

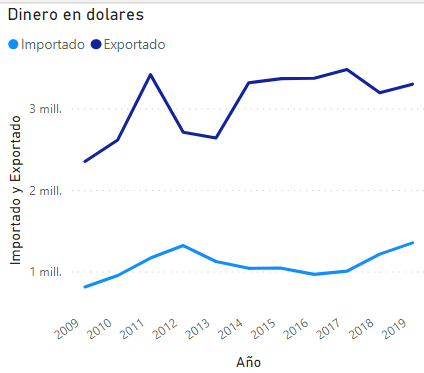
Gráfico, Gráfico circular

Descripción generada automáticamente

Luego, el objetivo de este gráfico, es el de comparar la cantidad en toneladas que se importado y exportado de estos alimentos.



Luego, se tiene este gráfico, que tiene como objetivo mostrar el costo en dólares de lo importado vs lo exportado



Por otro lado, también podemos analizar que alimentos se importan más, para crear planes de alimentación basados en estos productos que más se traen a Colombia o hasta incentivar esta producción para dejar de importar y comenzar a producir más.

Las aplicaciones son muchas y le sirven a cualquiera que trabaje con alimentos.

Se muestra a continuación el storytelling creado a partir de los datos y graficos anteriormente mencionados.

**Imagen que contiene Gráfico

Descripción generada automáticamente**

Nótese que al final del storytelling, se habla de predicciones hechas a partir de los datos, eso se hablara a continuación en el análisis predictivo.

**Análisis predictivo.**

Para el análisis predictivo se hicieron un total de 11 modelos ML, donde 1 esta orientado en la regresión lineal y los otros 10, en clasificación. Esto se hizo ya que los datos que se encontraron se ajustan muy bien a un modelo de regresión (más adelante se mostraran los resultados) pero para modelos de clasificación si se dificulto.

* Problema de regresión lineal: Para este problema, se escogió un nuevo dataset, que se relaciona con los anteriores del análisis descriptivo. En este caso, es un dataset que contiene todos los países de América y del caribe, los cuales muestran la producción de aguacate en cada año, durante un periodo de 10 años (2009-2019). Se escogió este problema de regresión lineal para predecir el comportamiento a futuro de algún producto, que en este caso fue el aguacate. Se muestra a continuación el dataset cargado y usado en la aplicación Knime.

Tabla

Descripción generada automáticamente

Con esta tabla se tiene la información suficiente para predecir el comportamiento que tendrá durante un próximo año, la producción del aguacate en estos países.

Se muestra ahora el modelo creado a partir del dataset mostrado:

Gráfico

Descripción generada automáticamente

Se carga entonces en la primera caja, el dataset como se mostro anteriormente, luego se hace una limpieza de datos, esto consta de las 3 siguientes tablas que se ven en la imagen anterior:

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

Para valores perdidos, se cambian los registros que no estén, por la media de la misma columna. En la imagen anterior, se muestra la configuración se que hizo para lograr sacar el dato de la media.

Luego, se aplico un filtro y cambio de valores atípicos. Para esto, se usa la caja de Numeric Outliners que a partir de cuartiles (Segundo y Cuarto cuartil) los datos deben estar. Los datos que no cumplan la condición, se le aproximara al rango más cercano. Se muestra a continuación la configuración de la caja:

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

Con el k = 1.5, se configura el rango máximo que tendrán los cuartiles para considerar un dato atípico o no. Luego los que no cumplen la condición, se le cambiara al dato permitido más cercano.

Por último, se filtran las columnas que no se utilizaran para la regresión:

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

En este caso se filtran columnas no numéricas (String) y columnas que solo tienen datos numéricos, pero para identificación (Código de área y Producto). Con esto se explica entonces el preprocesamiento aplicado para la parte de la regresión lineal. Ahora se hablará del modelo ML.

Primero, se tiene una caja de partición de los datos, esta caja particiona en 70-30 para datos de entrenamiento y datos de prueba respectivamente. La configuración es la siguiente:

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Se usa una semilla fija para que las pruebas siempre las mismas.

Para la regresión lineal, se aplicó la siguiente configuración:

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Para el objetivo (target) se coloco el ultimo año del registro que es lo que se quiere predecir para los siguientes años.

Luego, se aplico el regresor lineal para que mostrara la predicción, obteniendo la siguiente tabla:

Pantalla de computadora

Descripción generada automáticamente con confianza baja

Vemos entonces que tiene una predicción bastante acertada, ya que, analizando el incremento anual de la producción, se ve que igualmente crece en proporción a lo visto en los anteriores años. Por último, se muestra el puntaje sacado del modelo.

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Vemos que el R2 queda muy cercano a 1, dejando ver que el modelo tuvo una regresión buena y que el MAE% es del 0,055 mostrando que el error fue muy bajo. Podemos decir entonces que la regresión lineal fue acertada por parte del modelo.

* Problema de clasificación: Se introdujo también un problema de clasificación, para poder clasificar a partir de datos futuros, que alimentos pertenecen según las características de esos datos.

Para esto se usó el data set anterior mencionado, pero con todos los productos seleccionados en la parte del análisis descriptivo. Se muestra la tabla con los datos:

Tabla

Descripción generada automáticamente

Se muestra los dos primeros modelos creados para estos datos:

Imagen que contiene Diagrama

Descripción generada automáticamente

En este caso, solo se filtraron las columnas que no tuvieran relevancia para el modelo:

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

Los códigos son solo identificadores y el área pertenece al país, pero como l problema es de clasificación de productos, entonces se saca del modelo.

Luego, la partición es la misma aplicada en el modelo de regresión, 70-30 para entrenamiento y pruebas respectivamente con la misma semilla.

Ahora, se aplicó un modelo de KNN, esto se hizo para poder comparar distintos modelos para clasificación. Se muestra la configuración del modelo KNN. Tras varias pruebas, el mejor k, fue el 20, dando el mejor resultado.

Interfaz de usuario gráfica, Texto

Descripción generada automáticamente

Por otro lado, se probo también el modelo de árboles de decisión el cual tiene la siguiente configuración:

Interfaz de usuario gráfica

Descripción generada automáticamente

Tras varias pruebas, se decidió que la mejor cantidad de datos por nodo son 10. Después pasamos a ver los scores de ambos modelos.

Tabla

Descripción generada automáticamente

Vemos en la primera tabla el puntaje para KNN y la segunda el árbol de decisión. Como vemos, el mejor rendimiento lo tuvo KNN con una precisión mayor y un error más bajo, pero en general, los modelos arrojan clasificaciones pésimas. Debido a esto, se agregaron clasificadores de características y balanceo de datos, para tratar de aumentar la precisión de estos modelos.

1. Seleccionador de características:

Diagrama

Descripción generada automáticamente

Para este caso, se agregó un algoritmo de Forward Feature Selection, el cual tiene por debajo, unos procesos ML y puntajes que seleccionan cuales son las mejores características de acuerdo con los puntajes lanzados. Se obtuvieron las siguientes columnas:

Tabla

Descripción generada automáticamente

Con esto se aplicaron los mismo modelos ML, con los mismos parámetros y se obtuvieron los siguientes resultados:

Tabla

Descripción generada automáticamente

Vemos entonces que, en vez de mejorar el modelo, empeoraron en ambos casos.

1. Balanceo de datos:

Imagen que contiene Gráfico

Descripción generada automáticamente

Vemos entonces que primero hacemos un estadístico de las ocurrencias por cada clase, se logra usando la caja de Statistics y esta arroja este primer resultado:

Tabla

Descripción generada automáticamente

Como vemos, hay demasiadas diferencias entre las clases mayoritarias y las clases minoritarias, haciendo que se afecte el modelo. Aquí es cuando aplicamos SMOTE, que es un algoritmo que, por medio de vectores unidos por un punto y su vecino más cercano, crean datos sintéticos a partir de este vector y las ocurrencias que ya existan en el dataset.

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Vemos entonces la configuración de SMOTE, el cual decimos que, desde un punto, vea 5 vecinos mas cercanos y cree los datos sintéticos. También decimos que solo incremente las clases minoritarias y las iguales a la mayoritaria. Se obtiene el siguiente resultado:

Tabla

Descripción generada automáticamente

Ahora quedaron todas las clases balanceadas. Se aplica entonces los mismos modelos ML con los mismos parámetros y se obtienen los siguientes resultados:

Tabla

Descripción generada automáticamente

Como vemos, se ve una notable mejora para ambos modelos, siendo el KNN superior al árbol de decisión. Sin embargo, con estos datos se sigue clasificando mal.

1. Combinación: Se combinaron los dos algoritmos para ver si mejoraba o empeoraban los resultados:

Imagen que contiene Diagrama

Descripción generada automáticamente

Se aplicaron los mismos parámetros y se obtuvieron los siguientes resultados.

Tabla

Descripción generada automáticamente

Se observa entonces, que, al usar el clasificador y el balanceo, los modelos empeoran.

Finalmente, entonces se decide que para este dataset, el mejor modelo es el KNN con solo balanceo de datos. Pero aun sigue con problemas, ya que la precisión es muy baja. Por tanto, se decidió hacer un ultimo modelo ML pero con un dataset más robusto:

Tabla

Descripción generada automáticamente

Este data set contiene todos los productos de todos los países del mundo y la producción hecha de cada uno de estos productos, desde 1961 hasta el 2019. Se muestra a continuación el modelo creado para este dataset.

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

Para este caso, se filtraron las columnas que no tuvieran relación, por los mismos criterios explicados anteriormente. Aquí aparece una nueva caja, que sirve para seleccionar solo los productos con los que se ha trabajado:

Interfaz de usuario gráfica, Texto

Descripción generada automáticamente

Para esto, solamente se selecciona la columna de la cual queremos filtrar y seleccionamos los productos que queremos. Arroja la siguiente tabla:

Interfaz de usuario gráfica, Tabla

Descripción generada automáticamente

Ahora, vemos que hay muchos datos faltantes. Ya que hay muchas filas que tienen numerosos datos faltantes, se decidió entonces eliminar las que tuvieran datos faltantes. Se muestra la configuración para eliminar datos faltantes:

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

Y se obtiene la siguiente tabla:

Tabla

Descripción generada automáticamente

Una vez que el data set está limpio, se procede a aplicar los algoritmos de ML. Fueron aplicados con los mismos parámetros de los anteriores modelos y se obtuvieron los siguientes resultados:

Tabla

Descripción generada automáticamente

Como vemos, los resultados con este data set, el rendimiento de los modelos mejoro, al tal grado de dar perfecto en arboles de decisión. Con este modelo ya se puede predecir a partir de datos nuevos, a que producto pertenecen los nuevos datos.

La cuestión es: ¿Por qué con el data set trabajado anteriormente, no se logro ni si quiera un 80% de precisión que es lo ideal para decir que un modelo ML es bueno para predecir y porque con el nuevo si funciona?

Esto se responde a varios factores que implican a la data set en sí y no al modelo ni a las técnicas para mejorar el modelo. Por un lado, el primer data set contiene menos datos, para este primero se tiene un total de registros de 254, mientras que el nuevo data set contiene al final después de borrar los valores faltantes, un total de 2009 registros. Esto influye en que se pueden lograr mejores características del modelo y separar mejor cada clase. Por otro lado, en el primer conjunto de datos, solo se tienen registros de los países de América y del Caribe en un periodo de 10 años. En el nuevo, se tienen datos de todo el mundo y por un periodo de 54 años, esto implica mayor separación de características para cada producto y más datos que aportan al modelo.

Conclusiones:

Para mostrar datos de la mejor manera, se debe entender bien a quién van dirigidos, qué datos se quieren mostrar y qué se necesita. Hacer un fuerte análisis al comienzo del proyecto asegura que se pueda lograr un buen trabajo con los datos.

La limpieza de datos es esencial en cualquier parte del proyecto, ya sea para hacer análisis descriptivo o predictivo.

El dashboard, aunque presente gráficas para entender lo que pasa con los datos, resulta en desventaja ante el storytelling, pues tiene mayor capacidad de explicar, ya que se pueden contar mejor los datos y así llegar más fácilmente al cliente.

Los modelos machine learning funcionan para predecir comportamientos y clasificar datos nuevos. Son muy útiles en la actualidad para las empresas o personas tomar decisiones que mejoren aspectos de productividad.

Por último, entre más datos se obtengan, es mucho mejor para poder dar un análisis o una predicción más precisa.