**Reporte Final - Proyecto**

**Presentado por:**
*Sergio Andres Gonzalez Martinez - andgonzalezmar@unal.edu.co*
Sebastian Guerrero Salinas - *sebguerrerosal@unal.edu.co*
Rafael Antonio Salgado Lopez - *rasalgadol@unal.edu.co*

**Profesor:**
Fabio A. González
*fagonzalezo@unal.edu.co*

Lunes 14 de Diciembre del 2020


**Universidad Nacional de Colombia**
**Facultad de Ingeniería**
**Departamento de Ingeniería de Sistemas e Industrial**

## CONTENIDO

## 1. INTRODUCCIÓN

El aguacate es un fruto exotico carnoso, el cual se obtiene del árbol tropical que posee el mismo nombre. Dentro del mercado del aguacate se encuentran de dos maneras, se venden los convencionales y los orgánicos, teniendo sus diferencias uno del otro.

Aunque varios estudios han demostrado que el aguacate orgánico posee una mayor cantidad de vitaminas y grasas buenas para los seres humanos, hay que tener en cuenta que su precio suele ser superior al convencional, entonces podemos encontrar que existen tanto pros como contras así cómo existen sus respectivos grupos de personas que prefieren solamente el uno o el otro.

Dado esto, se planteó diseñar y crear una herramienta de Machine Learning que permita determinar si el aguacate es de tipo orgánico o convencional en base a ciertas variables y características de este cómo el precio, el año, la región, el número total de aguacates vendidos y las ventas por variedad.

## 2. COMPRENSIÓN DEL NEGOCIO

El aguacate se caracteriza por un elevado porcentaje de grasa. Es muy energético y se desaconseja su uso al final de las comidas. Concretamente, 100 g de aguacate aportan entre 128 y 233 kcal. Es una fruta muy rica en minerales, principalmente potasio, hierro y fósforo y muy baja en hidratos de carbono, no llega ni a 2 gramos por cada 100, cuando por ejemplo una manzana llega a 14 gramos. Sin embargo a esta fruta se le conoce popularmente como la "mantequilla vegetal" por ser muy rica en grasa, pero esta grasa se trata de una grasa saludable, vegetal, insaturada y sin colesterol.

**Propiedades nutricionales del aguacate**

| Información nutricional del aguacate | |
|---|---|
| Agua | 67,90% |
| Calorías | 233 Kcal |
| Proteínas | 1,88 gr |
| Grasas | 23,50 gr |
| Carbohidratos | 0,40 gr |
| Fibra | 6,33 gr |
| Calcio | 12,00 mg |
| Hierro | 0,49 mg |
| Magnesio | 30,00 mg |
| Fósforo | 43,00 mg |
| Potasio | 487,00 mg |
| Vitamina A | 12 µg |
| Vitamina C | 6,00 mg |
| Vitamina E | 1,30 mg |
| Vitamina K | 19 µg |
| Vitamina B3 | 31,42 mg |

El aguacate contiene múltiples beneficios para la salud, entre los que se encuentran muchos componentes que el cuerpo necesita y permite la prevención y tratamiento de muchas enfermedades importantes como es el cáncer.

El aguacate es conocido también como el oro verde en muchos países, dado su amplia comercialización en todo el mundo, es por esto que Colombia ha decidido realizar una exploración en otros mercados para empezar a invertir fuertemente en cosechar este fruto a futuro y posicionar este fruto a niveles cercanos a los que tiene en café.

Se ha encontrado en estudios previos que la producción de aguacate es mínima frente al convencional, pero se ha evidenciado una creciente demanda en los mercados. Uno de los mercados donde se tienen una amplia cantidad de datos sobre ventas por variedad es en Estados Unidos, y se observa con interés dada la situación anterior qué tipo de aguacate está teniendo más impacto en el mercado, pero muchas veces no se cuenta con información del tipo de aguacate, observando esta problemática el gobierno de Colombia desea construir un modelo que permita predecir el tipo de aguacate que se tiene dadas algunas características como el número de ventas, la fecha del registro, la región y el número de ventas del mismo y de sus variedades para a futuro tener en cuenta que tipo de aguacate le conviene más al país producir par ser exportado allí.

## 3. COMPRENSIÓN DE LOS DATOS

Para la recopilación de los datos, son datos existentes y tomados en mayo de 2018 en https://hassavocadoboard.com/ por parte de la plataforma Kaggle.
Se cuentan con 18249 registros, y las variables que contiene el dataset consta de:

- Date - La fecha de la observación
- AveragePrice - El precio promedio de un solo aguacate
- type - Tipo del aguacate si es convencional u organico
- year - el año
- Region - La ciudad o región de la observación
- Total Volume - Número total de aguacates vendidos
- 4046 - Número total de aguacates vendidos con PLU 4046
- 4225 - Número total de aguacates vendidos con PLU 4225
- 4770 - Número total de aguacates vendidos con PLU 4770
- Total Bags - Número total de canastas vendidas
- Small Bags - Cantidad de canastas pequeñas vendidas
- Large bags - Cantidad de canastas grandes vendidas

- xLarge bags - Cantidad de canastas extra grandes vendidas

para cada variable se tiene los siguientes tipos de datos

| Variable | tipo |
|---|---|
| Unnamed: 0 | int64 |
| Date | object |
| AveragePrice | float64 |
| Total Volume | float64 |
| 4046 | float64 |
| 4225 | float64 |
| 4770 | float64 |
| Total Bags | float64 |
| Small Bags | float64 |
| Large Bags | float64 |
| XLarge Bags | float64 |
| type | object |
| year | int64 |
| region | object |

El tamaño del dataset es de 18249 filas × 14 columnas

Los datos contienen información acerca de la cantidad de aguacates vendidos en total, la cantidad vendida por PLU (4046, 4225 y 4770) y por el total de canastas vendidas por tamaño (pequeña, mediana o grande) para una determinada fecha, y región de Estados Unidos.

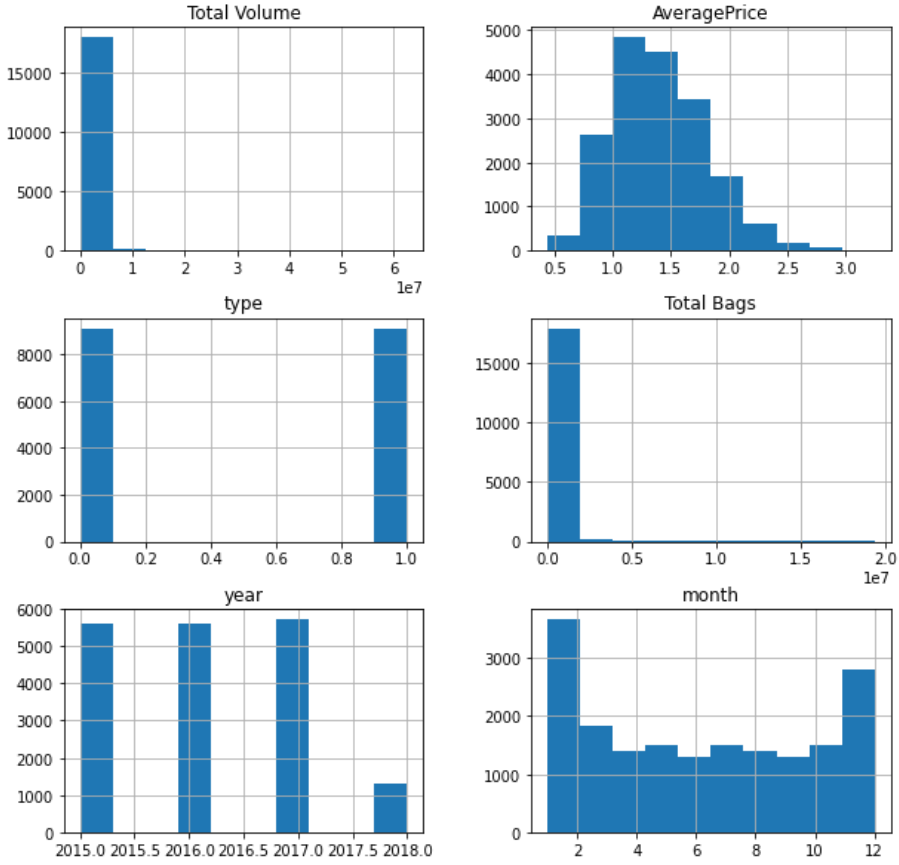
## 4. PREPARACIÓN DE LOS DATOS

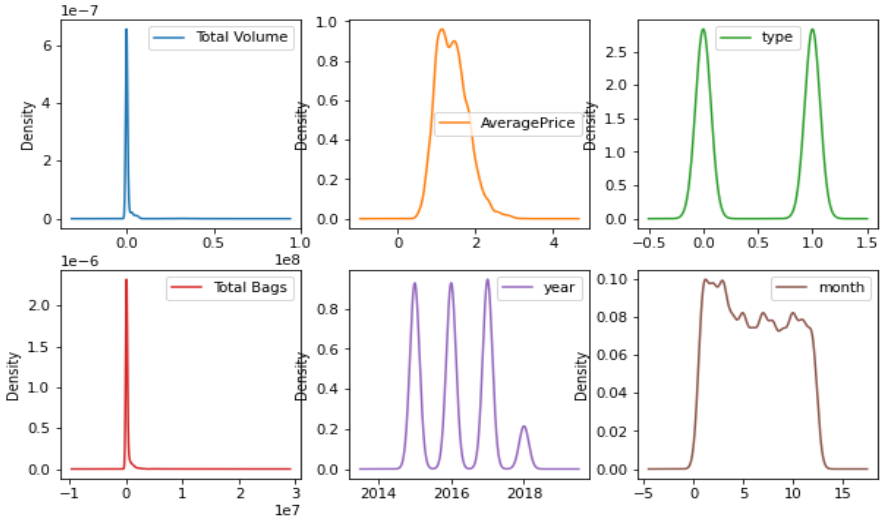
Inicialmente se realiza una previsualización de los datos de origen y se logra evidenciar que no tienen campos vacíos o caracteres especiales que necesiten algún tratamiento especial de limpieza.

Se identifican variables que pueden dificultar el proceso para la construcción de modelo, por lo que es necesario su preprocesamiento, en primer lugar para la variable fecha se trata de la siguiente manera, se separa el dia, el mes en variables separadas dado que el año ya existía como variable en el dataset inicial, posteriormente se elimina la variable fecha, y la variable Unnamed: 0 que actuaba como índice incremental de dataset.

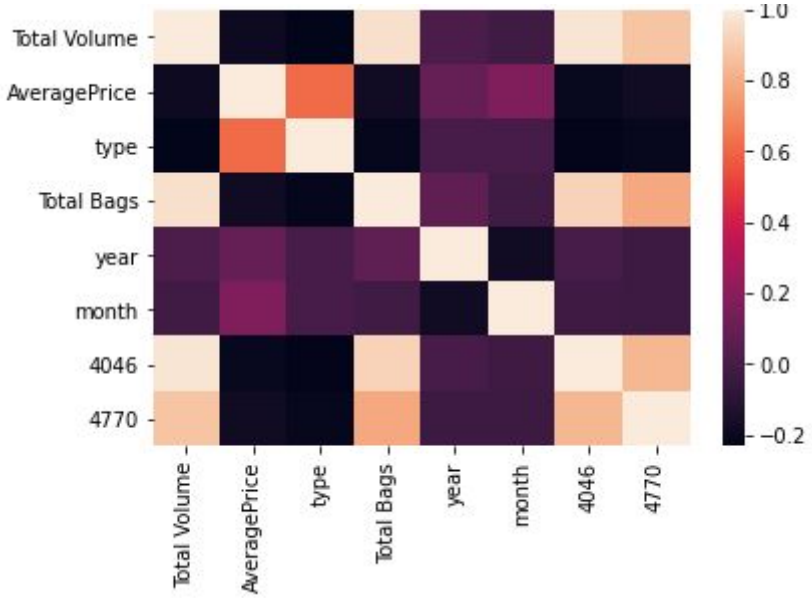
De igual manera se identifican variables categóricas: **type** y **región**, para las que se realiza la conversión a numéricas. Para la variable type se mapea el valor convencional a  0 y el valor orgánico a 1. Para la variable región se utiliza **one-hot-encoding** al ser aplicado el tamaño del dataset resultante es de 18249 filas × 67 columnas dado que nos convierte cada uno de los valores iniciales en nuevas columnas con valores de 0 o 1 en el caso que el registro pertenezca al estado o no.

Se evidencia que para la variable type es multimodal por lo que se espera a futuro contar con muchos más datos, también se cuentan con variables que tienen una distribución no uniforme como el caso de precio promedio y el mes

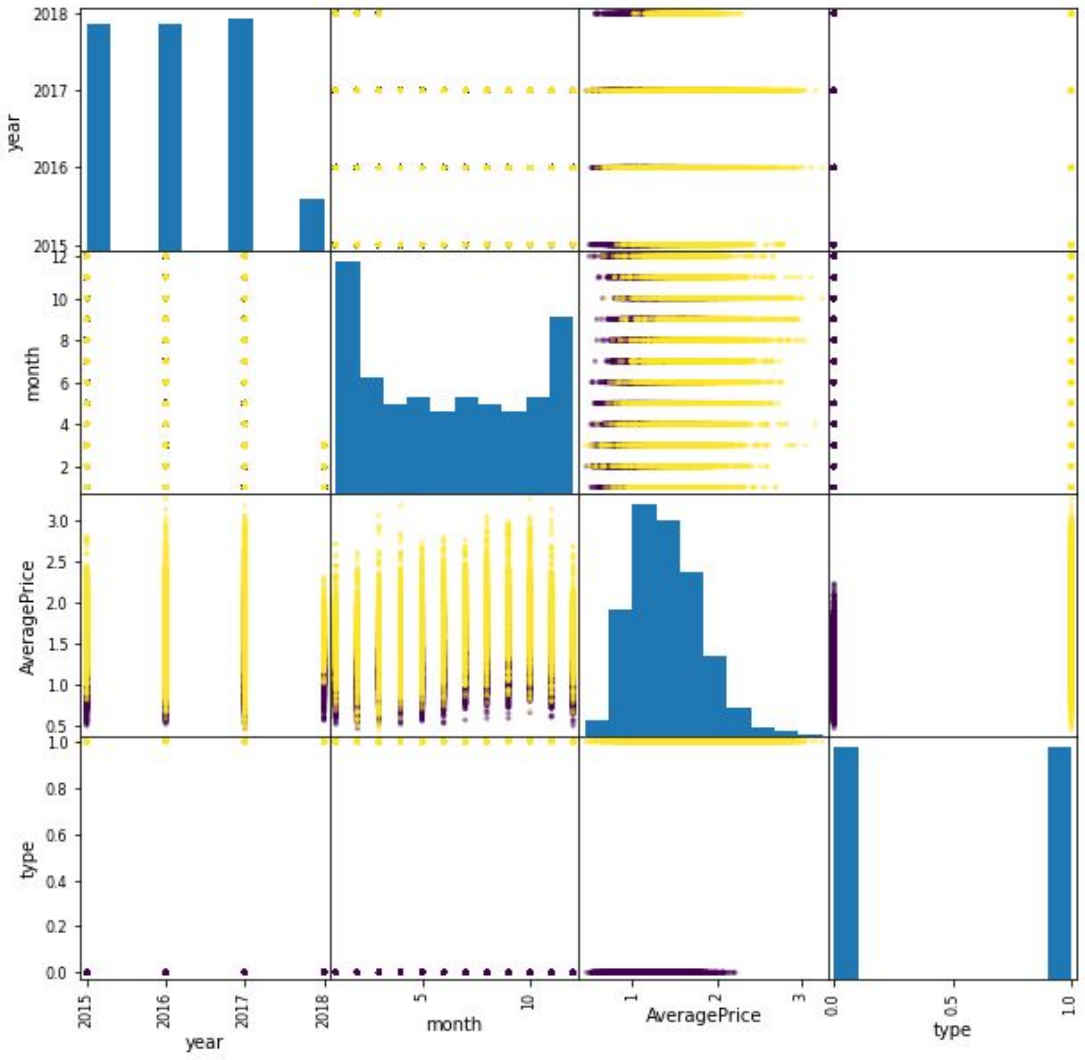





Para el observar la correlación entre las variables se observa que existe una correlación fuerte entre total volumen y el total de canastos como es de esperarse, y también para interés las variables del tipo de aguacate frente al precio promedio, esto lo observamos a través del siguiente mapa de calor




Excluimos las variables que no tienen relación alguna para type y a través de una matriz de gráficos de dispersión que incluye gráficos de dispersión individuales para cualquier combinación de características, observamos la siguiente gráfica




Finalmente extraemos del dataset la característica que servirá de clase la cual es type.

## 5. MODELADO

Para fines del proyecto se implementa el modelo de Random Forest, dado que la cantidad de datos se ajusta perfectamente a este tipo de modelo y de ser uno de los algoritmos por excelencia para resolver problemas de clasificación como es este el caso para predecir la clase 0 o 1, aguacate convencional u orgánico respectivamente, además de las ventajas que posee al no ser sensible a los hiperparametros y de no dar un sobreajuste al incrementar los estimadores

Se utilizan valores aleatorios con RandomizedSearchCV, evitando el uso de GridSearchCV dado que se ha demostrado que resulta ser más eficiente que optimizar los parámetros con GridSearchCV.

Para este caso se utilizan 5 pliegues evitando al máximo los sesgos que se pueden dar al escoger al azar datos de entrenamiento y la evaluación. para la optimización aleatoria con 20 iteraciones.

El tiempo que toma RandomizedSearchCV para entrenarse es de 25 minutos usando las 20 configuraciones. Observamos los datos con mejor desempeño a continuación luego del entrenamiento.

| | param_n_estimators | param_max_features | mean_test_score |
|---|---|---|---|
| **1** | 521 | 0.0647357 | 0.998904 |
| **4** | 180 | 0.0864772 | 0.998904 |
| **0** | 393 | 0.327087 | 0.998826 |
| **15** | 126 | 0.0877184 | 0.998669 |
| **2** | 416 | 0.444503 | 0.998669 |

La mejor configuración sobre todas las configuraciones es

```
random_search.best_params_
```

```
{'max_features': 0.06473566994470903, 'n_estimators': 521}
```

finalmente se reporta el score

```
random_search.best_score_
```

```
0.9989040176416409
```

## 6. EVALUACIÓN

Tras realizar el modelado con los datos, encontramos que el modelo se comporta de manera efectiva usando el mejor estimador que se halló para el conjunto de datos.

Al realizar la matriz de confusión y comparar los datos reales con la predicción realizada, podemos observar que se encontraron solamente 2 falsos, uno negativo y otro positivo, dejando el resto de datos en un acierto de la realidad.

```
y_pred = random_search.best_estimator_.predict(X_test)

print (confusion_matrix(y_test, y_pred))
```

```
[[2737    1]
 [   1 2736]]
```

Esto permite demostrar que el modelo es bastante efectivo para el problema de determinar si el aguacate es convencional u orgánico.

De este modo encontramos que tuvo un gran porcentaje de acierto.

```
random_search.score(X_test, y_test)
```

```
0.9996347031963471
```

También podemos observar el valor de las variables de precisión, exhaustividad y la métrica F1, la cual combina las mediciones de precisión y exhaustividad en una sola.

```
from sklearn import metrics

print('Precision: {}'.format(metrics.precision_score(y_test, y_pred)))
print('Recall: {}'.format(metrics.recall_score(y_test, y_pred)))
print('F_1 score: {}'.format(metrics.f1_score(y_test, y_pred)))
```

```
Precision: 0.999634636463281
Recall: 0.999634636463281
F_1 score: 0.999634636463281
```

## 7. DISTRIBUCIÓN

Dados los resultados obtenidos, el modelo es bastante acertado y tiene el potencial de permitir análisis más detallados para que el gobierno se enfoque en la producción de un tipo de aguacate para exportar.

Luego de estudiar los datos anteriores se evidencia que el aguacate sí es un producto capaz de volver potencia económica a un país como Colombia que lo conoce y maneja de primera mano.

## 8. Video

En el siguiente link se encuentra disponible el video.

https://youtu.be/qx_VE00zS6s

## REFERENCIAS EN LA RED

Aguacate, Persea Americana / Lauraceae - Frutas & Hortalizas:
https://www.frutas-hortalizas.com/Frutas/Presentacion-Aguacate.html

Aguacate - Exotic Fruit Box blog:
https://exoticfruitbox.com/frutas-exoticas/aguacate/