Universidad Nacional de San Agustin

Recuperación de la Información

Modeling Wine Preferences By Data Mining From Physicochemical Properties

Eduardo Diaz Huayhuas
Estefany Pilar Huaman Colque
Angel Yvan Choquehuanca Peraltilla

Maestria en Ciencias de la Computación

Docente:
MSc. Ana Maria Cuadros Valdivia

12 de junio de 2023

Resumen

El proposito de esta investigación, es determinar los parametros optimos para la produccion de vino de un vinicola de Portugal usando tecnicas de Data Mining, Algoritmos de clasificacion y software de analisis de datos, para poder lograr con satisfacción las certificaciones enologicas internacionales. Estas certificaciones permiten a la productora del vino demostrar que el producto es de excelente calidad de importación. Para ello se realizó tomas de pruebas en laboratorio de diferentes propiedades fisicoquimicas de vino para determinar ciertos patrones ocultos que permitan modelar una combinacion adecuada para presentarlo ante los certificadores internacionales. Dando como resultado propiedades fisicoquimicas adecuadas que se necesitan reportar al area de produccion para que se tome en consideración en el sembrío, cosecha y vendimia del vino

Índice general

Índice de figuras

Índice de cuadros

Capítulo 1

Introducción

El Data Mining o Minería de Datos nace como consecuencia de la generación masiva de datos, para dar soluciones al problema que plantea el uso de los mismos con la finalidad de ayudar a las empresas en la correcta toma de decisiones basadas en resultados determinados por estas tecnicas. El objetivo que se busca es determinar patrones ocultos en un conjunto de datos que permita tomar decisiones acertadas para mejorar un producto, presentar resultados optimos, etc.

En esta investigación se permitirá obtener un patrón oculto adecuado que permita a la empresa vinicola preparar una formulación adecuada para el vino elaborado. Esta combinación permitirá a los ingenieros modelar diseños geneticos que permitan producir la mejor uva que genere el vino que será presentada para la certificación internacional.

1.1. Definición del Problema

En el paper "Modeling wine preferences by data mining from physicochemical properties"[1]. El problema principal es proponer un enfoque de minería de datos para predecir las preferencias de sabor del vino que se basa en información generada por pruebas en humanos para su posterior certificación internacional. Un gran conjunto de datos es considerado, con muestras de vino blanco y tinto (de Portugal).

1.2. Objetivos

Los objetivos que buscamos en esta investigación, estan relacionadas al ambito de la Ciencias de la Computación, en base al data mining realizado al conjunto de datos que se tiene.

- Aplicar tecnicas de reduccion de dimensionalidad (PCA, T-SNE, SVD) al conjunto de datos, comparar y analizar los resultados
- Aplicar un algoritmo jerarquico a los datos
- Utilizando software de data mining, desarrollar y visualizar graficas en 2D y 3D.
- Descubrir patrones no visibles en base a los resultados de los analisis antes mencionados.

1.3. Descripcion de la Base de Datos

Este estudio considerará el vino verde, un producto único de la Región de Minho (noroeste) de Portugal. Medio en alcohol, es particularmente apreciado por su frescura. Este vino representa el 15 por ciento del total de la producción de Portugal, y alrededor del 10 por ciento se exporta, en su mayoría vino blanco. En este trabajo analizaremos las dos variantes más comunes, blanco y tinto (también se produce rosado), de la región demarcada del "Vinho Verde" Los datos fueron recolectados de mayo/2004 a febrero/2007 utilizando únicas muestras de denominación de origen protegidas que se analizaron en el entidad oficial de certificación (CVRVV). El CVRVV es una organización interprofesional con el objetivo de mejorar la calidad y comercialización de vino en Portugal. Los datos fueron registrados por un equipo computarizado (iLab), que gestiona automáticamente el proceso de elaboración basado en análisis de muestras de solicitudes de productores a laboratorio y análisis sensoriales. Cada entrada denota una prueba determinada (analítica o sensorial) y la base de datos final se exporta a una sola hoja (.csv).

1.4. Descripcion de cada Atributo

1.4.1. Variables de Entrada

1. Fixed acidity (g(tartaric acid)/dm3 : Gramos de Acido Tartárico en 1 dm3
2. Volatile acidity (g(acetic acid)/dm3 : Gramos de Acido Tartarico volátil en 1 dm3
3. Citric acid (g/dm3) : Gramos de Ácido Cítrico en 1 dm3 Residual sugar (g/dm3): Gramos de Azúcar Residual
4. Chlorides (g(sodium chloride)/dm3) : Gramos de Cloruro de Sodio en 1 dm3
5. Free sulfur dioxide (mg/dm3) : Mg de Dioxido de Azufre en 1 dm3 libre
6. Total sulfur dioxide (mg/dm3) : mg de Dioxido de Azufre en 1 dm3 total
7. Density (g/cm3) : Densidad (Gramos en 1cm3)
8. pH (Valor de acidez o alcalinidad)
9. Sulphates (g(potassium sulphate)/dm3) : gramos de Sulfato de Potasio en 1dm3
10. Alcohol (vol.porc) Volumen de Alcohol en porc (etanol)

1.4.2. Variables de Salida

1. Tipo de Vino: Tinto (rojo) o Blanco
2. Calidad: de 0 a 10

1.5. Rango de Valores

Se presentan los rango de valores encontrados en el Vino Rojo

Cuadro 1.1: Rango de Valores - Vino Rojo

Variable	Minimo	Maximo
fixed acidity	4.6	15.9
volatile acidity	0.1	1.6
citric acid	0.0	1.0
residual sugar	0.9	15.5
chlorides	0.01	0.61
free sulfur dioxide	1	72
total sulfur dioxide	6	289
density	0.990	1.004
pH	2.7	4.0
sulphates	0.3	2.0
alcohol	8.4	14.9
quality	0	10

Cuadro 1.2: Rango de Valores - Vino Blanco

Variable	Minimo	Maximo
fixed acidity	3.8	14.2
volatile acidity	0.1	1.1
citric acid	0.0	1.7
residual sugar	0.6	65.8
chlorides	0.01	0.35
free sulfur dioxide	2	289
total sulfur dioxide	9	440
density	0.987	1.039
pH	2.7	3.8
sulphates	0.2	1.1
alcohol	8.0	14.2
quality	0	10

1.6. Clasificacion de los Datos, Valores Null o No Null, y tipo de Datos

En la siguiente tabla se encuentra la cantidad de datos, tanto para el Vino blanco como para el Vino Rojo En el dataset recolectado, no se han encontrado valores nulos. Por tanto se considera la dataset como limpia.

Cuadro 1.3: Clasificacion de Datos - Vino Rojo

Num	Columna	Num de datos	¿Se encontró valores null?	Tipo de Dato
1	fixed acidity	4898	non-null	float64
2	volatile acidity	4898	non-null	float64
3	citric acid	4898	non-null	float64
4	residual sugar	4898	non-null	float64
5	chlorides	4898	non-null	float64
6	free sulfur dioxide	4898	non-null	float64
7	total sulfur dioxide	4898	non-null	float64
8	density	4898	non-null	float64
9	pH	4898	non-null	float64
10	sulphates	4898	non-null	float64
11	alcohol	4898	non-null	float64
12	quality	4898	non-null	int64
13	color	4898	non-null	object

Cuadro 1.4: Clasificacion de Datos - Vino Blanco

Num	Columna	Num de datos	¿Se encontró valores null?	Tipo de Dato
1	fixed acidity	1599	non-null	float64
2	volatile acidity	1599	non-null	float64
3	citric acid	1599	non-null	float64
4	residual sugar	1599	non-null	float64
5	chlorides	1599	non-null	float64
6	free sulfur dioxide	1599	non-null	float64
7	total sulfur dioxide	1599	non-null	float64
8	density	1599	non-null	float64
9	pH	1599	non-null	float64
10	sulphates	1599	non-null	float64
11	alcohol	1599	non-null	float64
12	quality	1599	non-null	int64
13	color	1599	non-null	object

Capítulo 2

Materiales a Utilizar

2.1. Software y datos

2.1.1. Dataset

Para realizar la siguiente investigación se utilizaron dos datasets conteniendo los valores fisicoquimicos del vino rojo y blanco. Estos archivos están disponibles bajo dominio publico en las siguientes direcciones:

- Vino Blanco: https://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/winequality-white.csv
- Vino Rojo: https://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/winequality-red.csv

2.1.2. Google Colab

Para desplegar nuestro cuaderno de Python que nos permitirá ejecutar algoritmos, procedimientos y analisis en base a los objetivos mencionados en el Capitulo 1.

2.1.3. Tableau

Se utilizará la aplicacion Tableau para el analisis de graficas autogeneradas por analisis de datos. Tableau es una aplicación de bussiness intelligence.

2.1.4. Libreria Seaborn, de Python

Se utilizará las herramientas de data mining que provee seaborn. Esta libreria nos permitirá generar datos de interes basados en algoritmos precargados. Los resultados basados en sus algoritmos nos permitirán tomar decisiones acertadas en la produccion de vino.

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH
count	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000
unique	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
top	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
freq	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
mean	6.854788	0.278241	0.334192	6.391415	0.045772	35.308085	138.360657	0.994027	3.188267
std	0.843868	0.100795	0.121020	5.072058	0.021848	17.007137	42.498065	0.002991	0.151001
min	3.800000	0.080000	0.000000	0.600000	0.009000	2.000000	9.000000	0.987110	2.720000
25%	6.300000	0.210000	0.270000	1.700000	0.036000	23.000000	108.000000	0.991723	3.090000
50%	6.800000	0.260000	0.320000	5.200000	0.043000	34.000000	134.000000	0.993740	3.180000
75%	7.300000	0.320000	0.390000	9.900000	0.050000	46.000000	167.000000	0.996100	3.280000
max	14.200000	1.100000	1.660000	65.800000	0.346000	289.000000	440.000000	1.038980	3.820000

Figura 2.1: Descripcion de la database

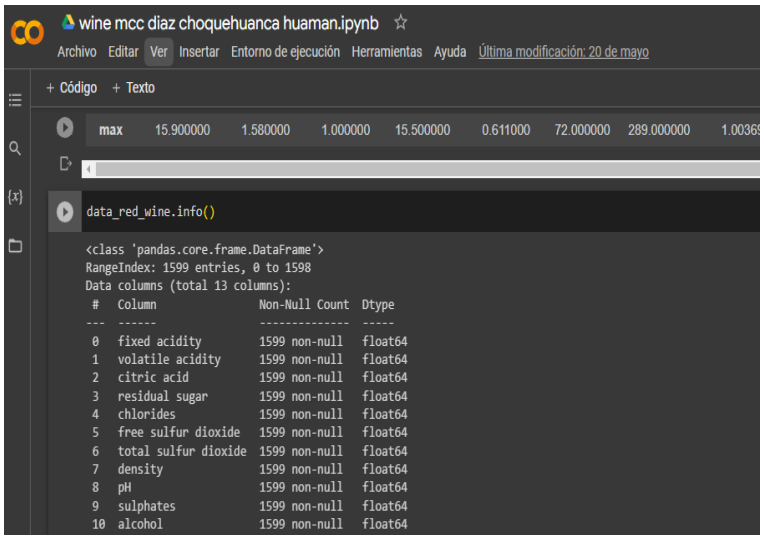


Figura 2.2: Google Colab del proyecto

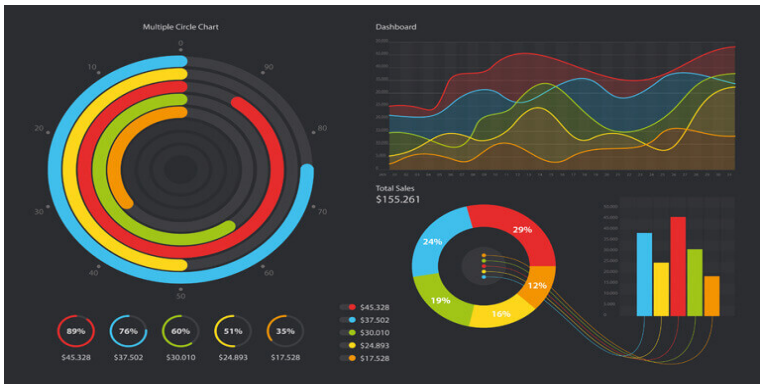


Figura 2.3: Tableau - Dashboard

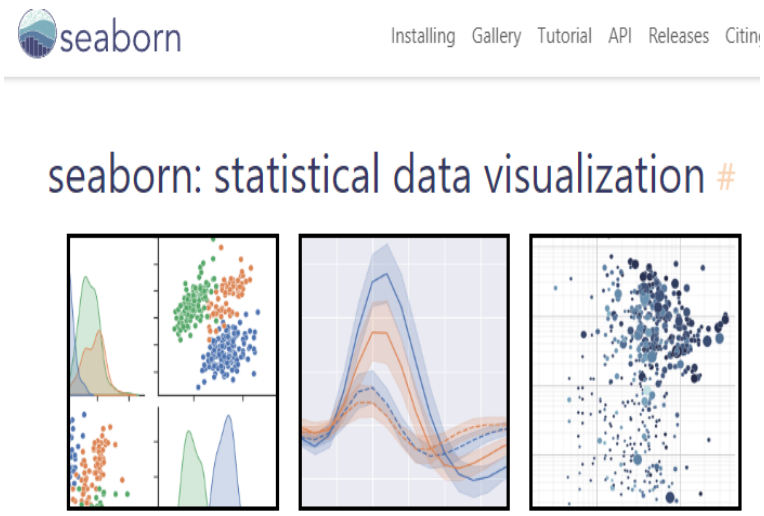


Figura 2.4: Proyecto Seaborn para Python

2.2. Algoritmos, tecnicas de reduccion de dimensionalidad y procesamiento de datos

2.2.1. Tecnicas de Reduccion de Dimensionalidad

PCA

Principal Component Analysis (PCA) es un método estadístico que permite simplificar la complejidad de espacios muestrales con muchas dimensiones a la vez que conserva su información. El método de PCA permite por lo tanto "condensar" la información aportada por múltiples variables en solo unas pocas componentes. Esto lo convierte en un método muy útil de aplicar previa utilización de otras técnicas estadísticas tales como regresión, clustering... Principal Component Analysis pertenece a la familia de técnicas conocida como unsupervised learning. [2]

SVD

La descomposición del valor singular (SVD) ayuda a reducir los conjuntos de datos que contienen un gran número de valores. Además, este método también es útil para generar soluciones significativas para menos valores. Sin embargo, este menor número de valores también comprende la inmensa variabilidad disponible en los datos originales.

Tambien se puede utilizar esta técnica para interpolar mediciones dispersas o para un algoritmo de aprendizaje automático. Esta técnica ayuda a la regresión y clasificación del conjunto de datos.[3]

tSNE

Incrustación Stochastic Neighbor-t distribuida (tSNE) es una técnica no lineal no supervisada utilizada principalmente para la exploración de datos y la visualización de datos de alta dimensión.

En términos más simples, tSNE le da una sensación o intuición de cómo se organizan los datos en un espacio de alta dimensión. Fue desarrollado por Laurens van der Maatens y Geoffrey Hinton en 2008.

El algoritmo tSNE calcula una medida de similitud entre pares de instancias en el espacio de alta dimensión y en el espacio de baja dimensión. Además, tSNE podría usarse para investigar, aprender o evaluar la segmentación. Muchas veces seleccionamos la cantidad de segmentos antes del modelado o iteramos después de los resultados. tSNE a menudo puede mostrar una separación clara en los datos.

Esto se puede usar antes de usar su modelo de segmentación para seleccionar un número de clúster o después para evaluar si sus segmentos realmente se mantienen. tSNE, sin embargo, no es un enfoque de agrupamiento, ya que no conserva las entradas como PCA y los valores a menudo pueden cambiar entre ejecuciones, por lo que es pura exploración.

Algoritmo KNN

El algoritmo de k vecinos más cercanos, también conocido como KNN o k-NN, es un clasificador de aprendizaje supervisado no paramétrico, que utiliza la proximidad para hacer clasificaciones o predicciones sobre la agrupación de un punto de datos individual. Si bien se puede usar para problemas de regresión o clasificación, generalmente se usa como un algoritmo de clasificación, partiendo de la suposición de que se pueden encontrar puntos similares cerca uno del otro.

Para determinar qué puntos de datos están más cerca de un punto de consulta determinado, será necesario calcular la distancia entre el punto de consulta y los otros puntos de datos. Estas métricas de distancia ayudan a formar límites de decisión, que dividen los puntos de consulta en diferentes regiones. [4]

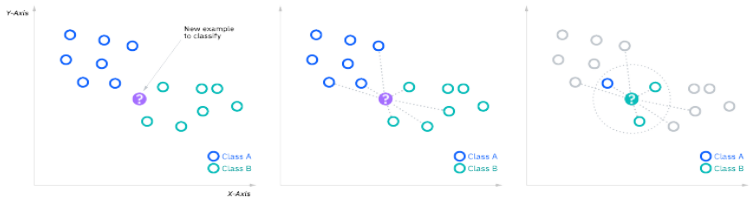


Figura 2.5: Algoritmo KNN

Capítulo 3

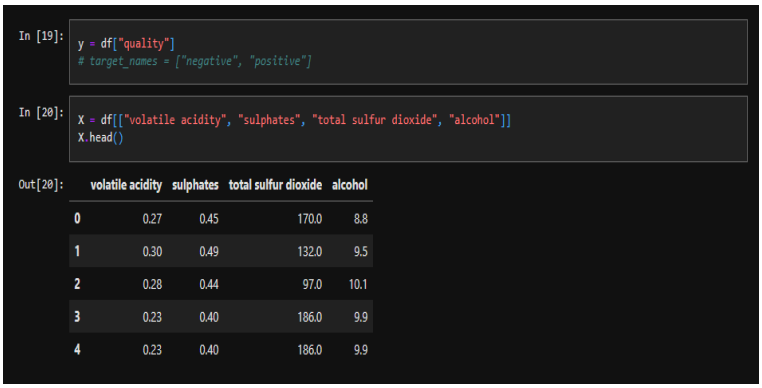
Metodologia a aplicar

3.1. Reduccion de Dimensionalidad

3.2. Algoritmo KNN en la calidad del vino

Para el algorimo KNN se utilizó la libreria **Sklearn**, que incluye el modulo *sklearn.neighbors. .KNeighborsClassifier*. Esta libreria nos permitirá utilizar el algoritmo KNN aplicado.

Primero se aplicará el uso del modulo sklearn.modelselection.traintestsplit para desarrollar un entrenamiento de los datos. Luego se aplicará el algoritmo KNN para determinar los valores de K Neighbors comparados con la puntuación obtenida del entrenamiento con el train test split. Todos estos datos seran llevados en una grafica.



```
In [19]:
y = df["quality"]
# target_names = ["negative", "positive"]

In [20]:
X = df[["volatile acidity", "sulphates", "total sulfur dioxide", "alcohol"]]
X.head()
```

Out[20]:

	volatile acidity	sulphates	total sulfur dioxide	alcohol
0	0.27	0.45	170.0	8.8
1	0.30	0.49	132.0	9.5
2	0.28	0.44	97.0	10.1
3	0.23	0.40	186.0	9.9
4	0.23	0.40	186.0	9.9

Figura 3.1: Preparación del modelo entrenado

Como se aprecia en la figura anterior. Se va a utilizar como dataset x los datos que almacena las variables: volatile acidity, sulphates, total sulfur dioxide, alcohol, y se compara con la **calidad** (quality).

En la siguiente figura se comparará los datos obtenidos por los datos entrenados y el de KNN. Con ello tendremos los datos para cada k-vecino. Luego se mostrará una grafica.

```python
train_scores = []
test_scores = []
for k in range(1, 20, 2):
    knn = KNeighborsClassifier(n_neighbors=k)
    knn.fit(X_train, y_train)
    train_score = knn.score(X_train, y_train)
    test_score = knn.score(X_test, y_test)
    train_scores.append(train_score)
    test_scores.append(test_score)
    print(f"k: {k}, Train/Test Score: {train_score:.3f}/{test_score:.3f}")
```

Figura 3.2: Preparando el analisis comparativo

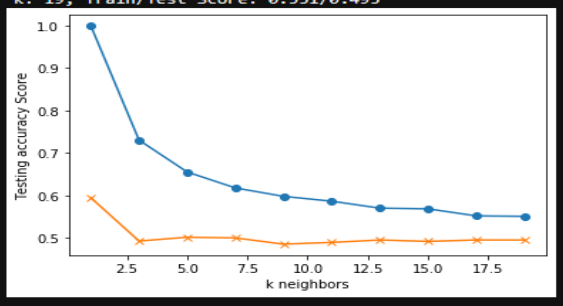
3.3. Matriz de Correlación

Una matriz de correlación es una tabla que indica los coeficientes de conexión entre los factores. Cada celda de la tabla muestra la conexión entre los dos factores. En el caso del analisis de datos del vino. Se pretende demostrar la correlacion de diferentes variables fisicoquimicas que permita determinar los factores que mas influyen en la calidad.

3.3.1. KNN: Vino Blanco

3.3.2. KNN: Vino Rojo

```
k: 1, Train/Test Score: 0.999/0.594
k: 3, Train/Test Score: 0.730/0.492
k: 5, Train/Test Score: 0.655/0.501
k: 7, Train/Test Score: 0.617/0.500
k: 9, Train/Test Score: 0.598/0.485
k: 11, Train/Test Score: 0.586/0.489
k: 13, Train/Test Score: 0.570/0.495
k: 15, Train/Test Score: 0.568/0.491
k: 17, Train/Test Score: 0.552/0.495
k: 19, Train/Test Score: 0.551/0.495
```

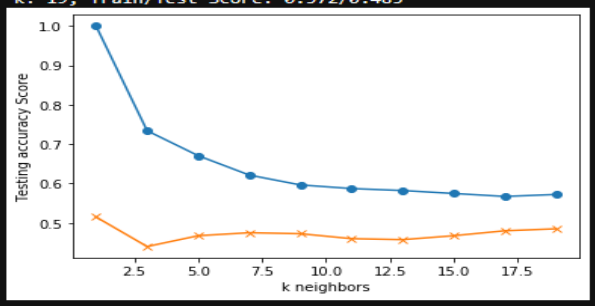



```
In [25]: knn = KNeighborsClassifier(n_neighbors=9)
         knn.fit(X_train, y_train)
         print('k=9 Test Acc: %.3f' % knn.score(X_test, y_test))

k=9 Test Acc: 0.485
```

Figura 3.3: KNN para Vino Blanco

```
k: 1, Train/Test Score: 1.000/0.515
k: 3, Train/Test Score: 0.734/0.440
k: 5, Train/Test Score: 0.671/0.468
k: 7, Train/Test Score: 0.621/0.475
k: 9, Train/Test Score: 0.596/0.472
k: 11, Train/Test Score: 0.587/0.460
k: 13, Train/Test Score: 0.582/0.458
k: 15, Train/Test Score: 0.575/0.468
k: 17, Train/Test Score: 0.567/0.480
k: 19, Train/Test Score: 0.572/0.485
```




```
[8]: knn = KNeighborsClassifier(n_neighbors=9)
     knn.fit(X_train, y_train)
     print('k=9 Test Acc: %.3f' % knn.score(X_test, y_test))

k=9 Test Acc: 0.472
```

Figura 3.4: KNN para Vino Rojo

Capítulo 4

Analisis de Resultados

4.1. Comparativa de las propiedades mas importantes

Segun los parametros tradicionales: los parametros Alcohol, Acido Volatil y la presencia de Sulfatos son los valores mas importantes en la produccion de un vino. Se analizará esta comparativa despues de realizar las reducciones de datos respectivos.

Alcohol y Calidad

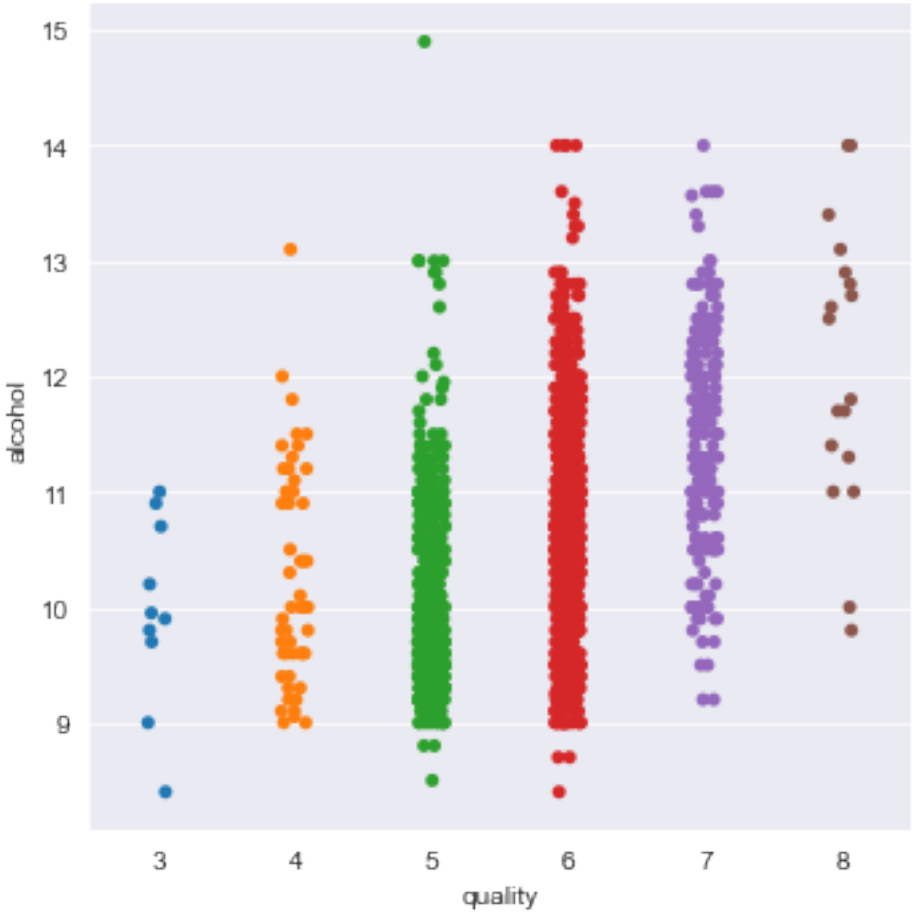


Figura 4.1: Alcohol vs Calidad

La imagen representa la presencia de Alcohol y la Calidad. En un principio se considera que si tiene la presencia del Alcohol en medias cantidades, se obtendrá un excelente vino. Sin embargo si se tiene una cantidad (En grados) de Alcohol entre 9 y 13, se obtendrá un vino de calidad media.

Sin embargo, si se tiene pocas trazas de alcohol entre 9 y 11 grdaos. Se obtendrá un vino de baja calidad. Para una calidad de alcohol esperada aceptable de 7. Se necesitara la presencia de alcohol entre 10 y 13.

Acido Volatil y Calidad

La importancia de tener acidos volatiles en el vino determina su calidad. Para obtener la mejor calidad se necesita tener la cantidad de acido volatil entre 0.3 y 0.5 g/litro. Segun la grafica anterior,

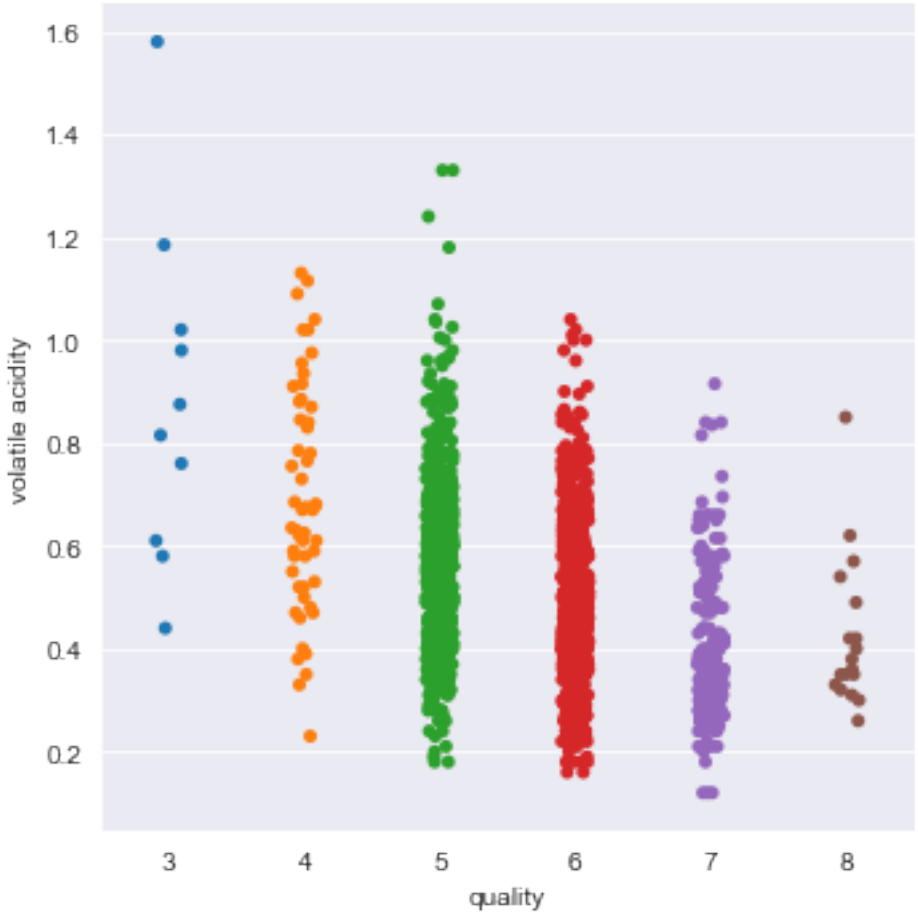


Figura 4.2: Acido Volatil vs Calidad

se aprecia que ese rango de valor asegura entre calidad 6, 7 y 8.

Presencia de Sulfatos y Calidad

La presencia de sulfatos influye en la cantidad de Dioxido de Azufre libre. Estos sulfatos producirán a futuro el SO2 en gas para darle mejor aroma al vino. Siendo el aroma un punto importante en la calidad de los evaluadores.

Es por ello que los valores entre 0.70 y 0.75 de presencia de sulfatos, determinaran una mejor calidad de vino. Ya que estos valores otorgan de calidad 6 a 8.

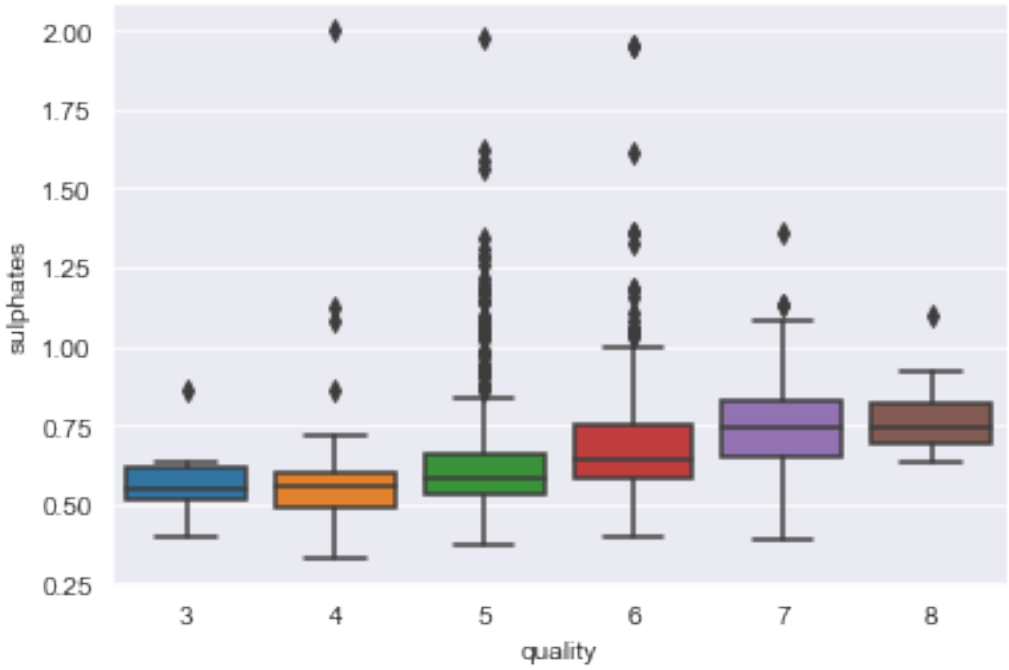


Figura 4.3: Presencia de Sulfatos vs Calidad

4.2. Algoritmo KNN

4.2.1. KNN Vino Blanco

Considerando el numero de vecinos igual a 9. Se obtiene un Test Accuracy de un *47*. Esto significa que: la precision en k=9 los datos generados son aceptables para el uso del algoritmo.

4.2.2. KNN Vino Rojo

Considerando el numero de vecinos igual a 9. Se obtiene un Test Accuracy de un *47*. Esto significa que la precision en k=9 los datos generados son aceptables para el uso del algoritmo.

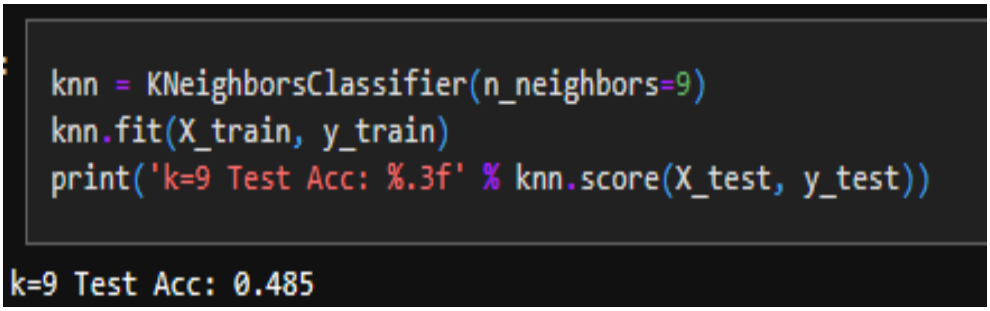


Figura 4.4: Resultados de KNN para el Vino Blanco

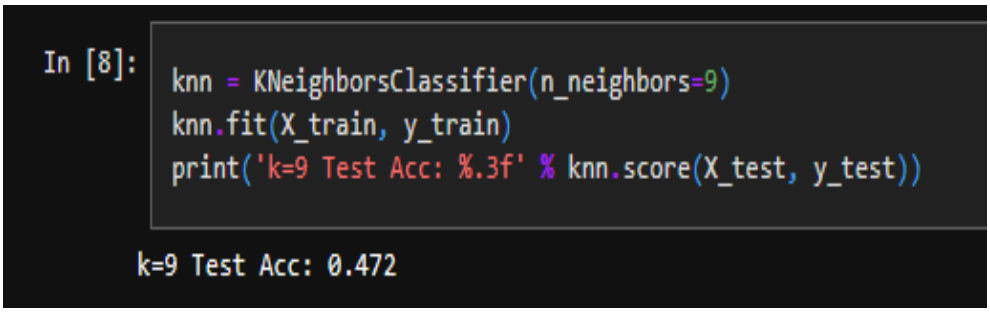


Figura 4.5: Resultados de KNN para el Vino Rojo

4.3. Matriz de Correlacion

Los resultados determinan que "free sulfur dioxidez "total sulfur dioxide"(Dioxido de Azufre liberado y Total de Dioxido de Sulfuro), son muy importantes en la calidad del vino. O sea que en su parametro de correlación permite obtener una calidad mayor de 7.

En esta grafica se observa la importancia utilizando el .ºrbol de Decisiones". Corroborando que efectivamente "Total Sulfure Dioxide.ºs la variable MAS IMPORTANTE a considerar en la calidad del vino.

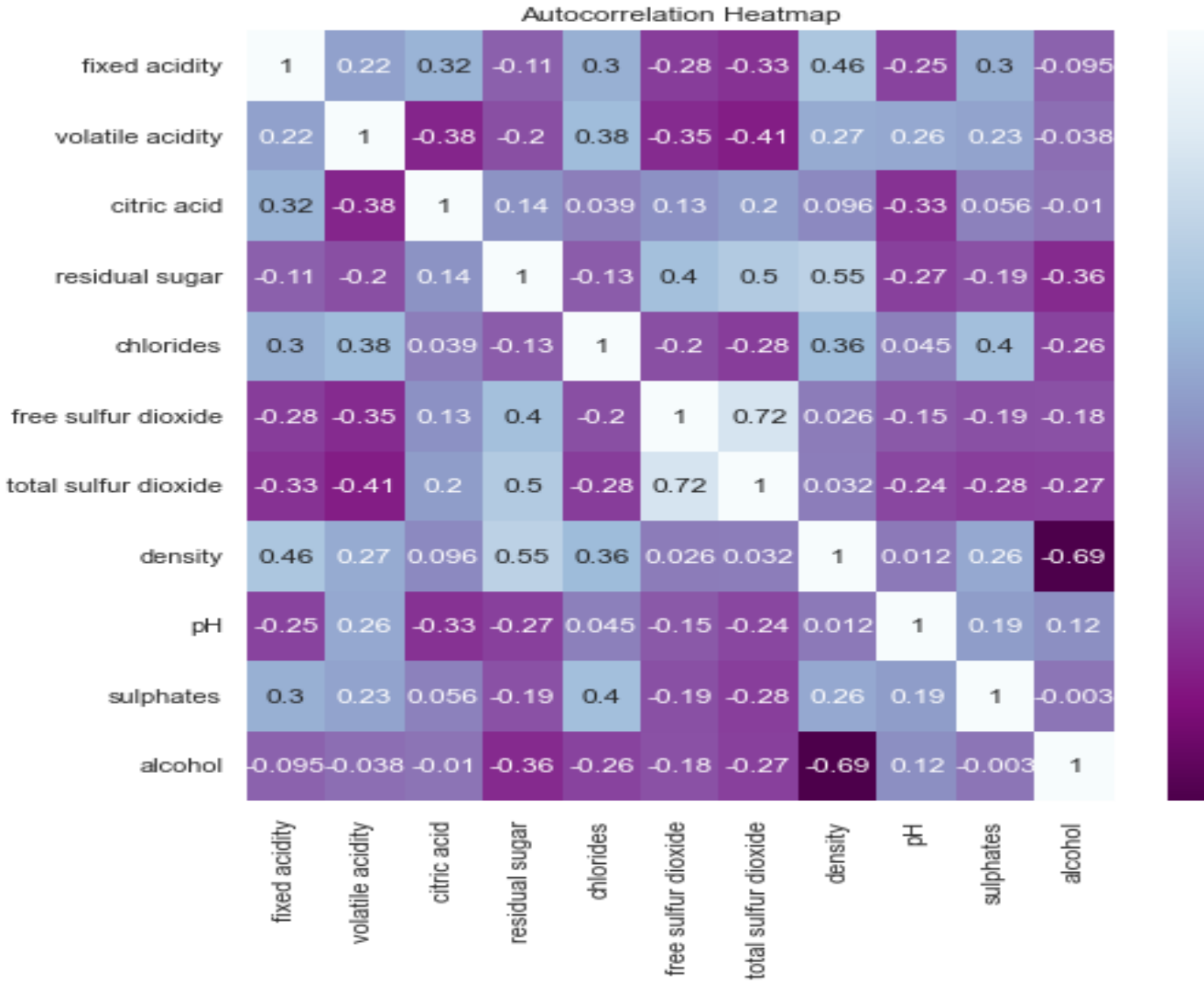


Figura 4.6: Matriz de Correlación

In [35]:

```
sorted_idx = DTC.feature_importances_.argsort()
plt.barh(X.columns[sorted_idx], DTC.feature_importances_[sorted_idx])
plt.title('Feature Importance : DecisionTreeClassifier');
```

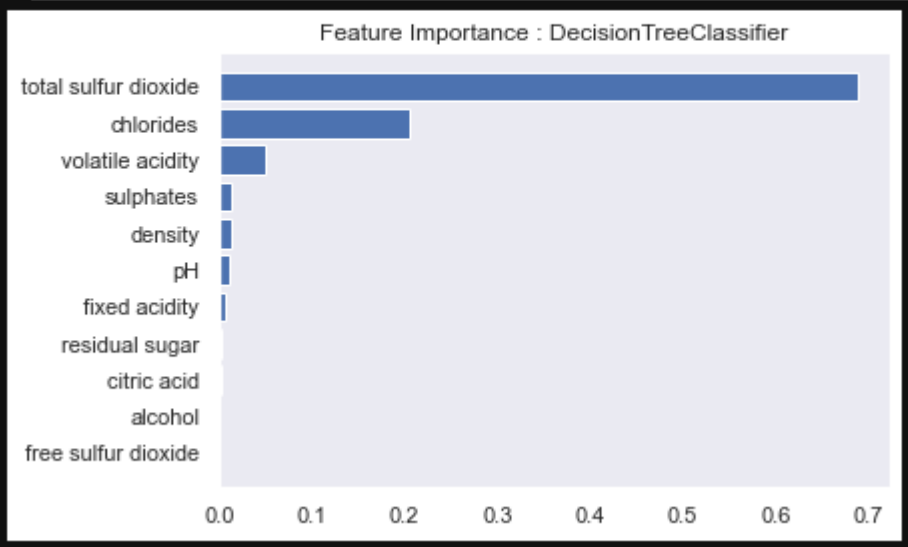


Figura 4.7: Analisis de Arbol de Decisiones

Capítulo 5

Conclusiones

En base a los estudios de datos realizados al vino rojo y blanco. Podremos determinar:

- Para ambos vinos, la cantidad de "free dioxide sulfur.[es] importante para la calidad. Ya que, este le otorga un aroma adecuado al vino. Obteniendo entre 6 y 8 de calidad.
- La presencia de sulfatos tambien es importante para la formacion de reaccion del dioxido de azufre gas. Parte de ello incluye tambien el sabor del vino.
- Los analisis de reduccion nos permiten analizar mejor los valores en base a las propiedades mas comunes entre si.
- La matriz de correlacion tambien determina que el "dioxido de azufre libre.[es] una variable que está relacionada con las demas propiedades fisicoquimicas.

Bibliografía

[1] Paulo Cortez, António Cerdeira, Fernando Almeida, Telmo Matos, and José Reis. Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, 47(4):547–553, 2009. Smart Business Networks: Concepts and Empirical Evidence.

[2] Kamila Zdybał, Elizabeth Armstrong, Alessandro Parente, and James C Sutherland. Pcafold: Python software to generate, analyze and improve pca-derived low-dimensional manifolds. *SoftwareX*, 12:100630, 2020.

[3] Michael L. Waskom. seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60):3021, 2021.

[4] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.