Entrenando a un sommelier

Martín Quijano - Martina Coletto

Introducción

Variables

- fixed acidity – Acidez fija
- volatile acidity – Acidez volátil
- citric acid – Ácido cítrico
- residual sugar – Azúcar residual
- chlorides – Cloruros
- free sulfur dioxide – Dióxido de azufre libre
- total sulfur dioxide – Dióxido de azufre total
- density – Densidad
- pH – Nivel de pH
- sulphates – Sulfatos
- alcohol – Contenido de alcohol
- quality – Calificación
- sensorial del vino (target)

Modelos

KNN y Random Forest

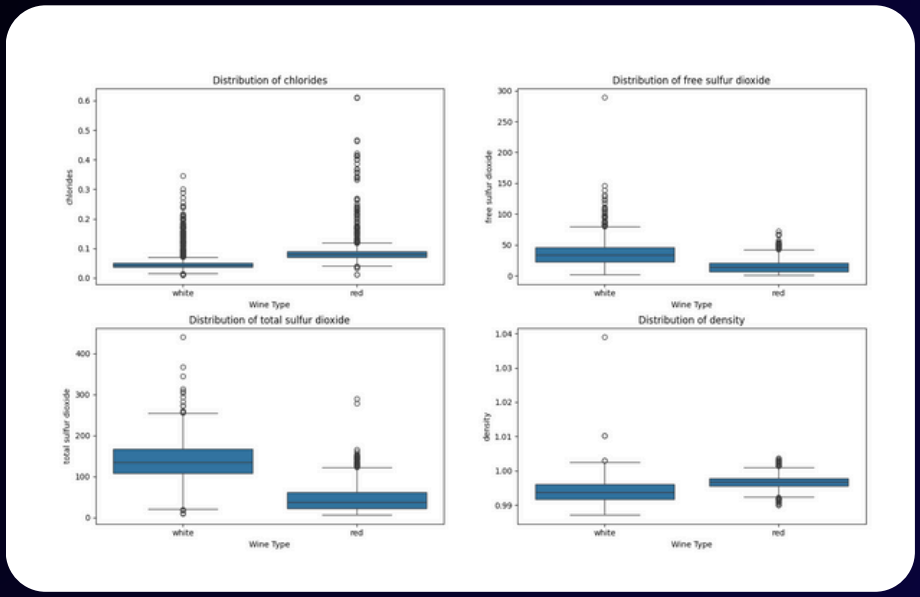
Objetivo

Realizar un análisis exhaustivo de variables (gráfico y analítico) para determinar la calidad y el color del vino

Vemos que las variables que mas influyen son credit-history, saving-account-amount, la tasa de interes, la duracion del credito y la antiguedad en el trabajo

Análisis exploratorio

Boxplot

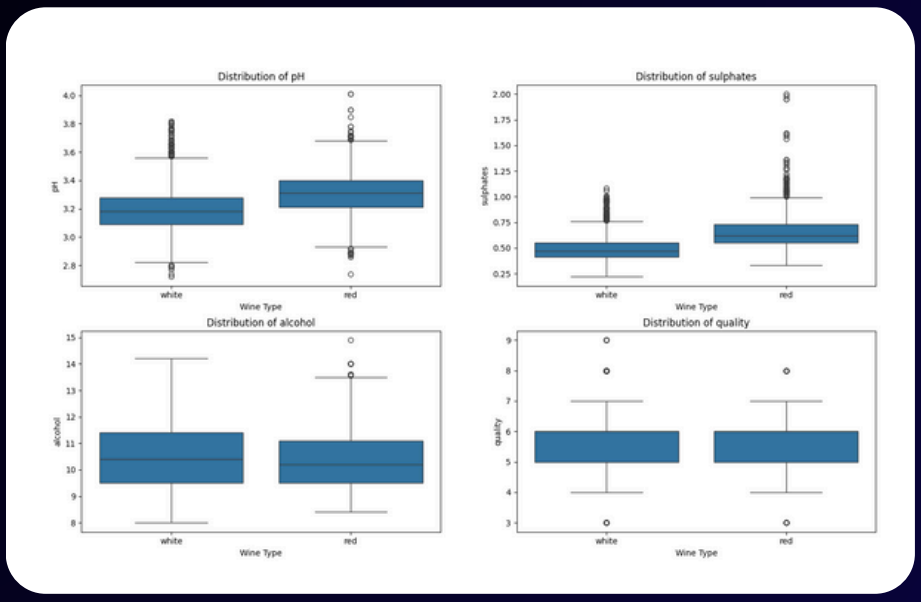


usamos boxplots para analizar la distribución de cada variable por color. Nuestro objetivo en este paso es evaluar si hay diferencias significativas en los valores de las variables dependiendo de su color.

Vemos que total sulfur dioxide es de estas cuatro en la que mas diferencia hay por tipo de vino

Análisis exploratorio

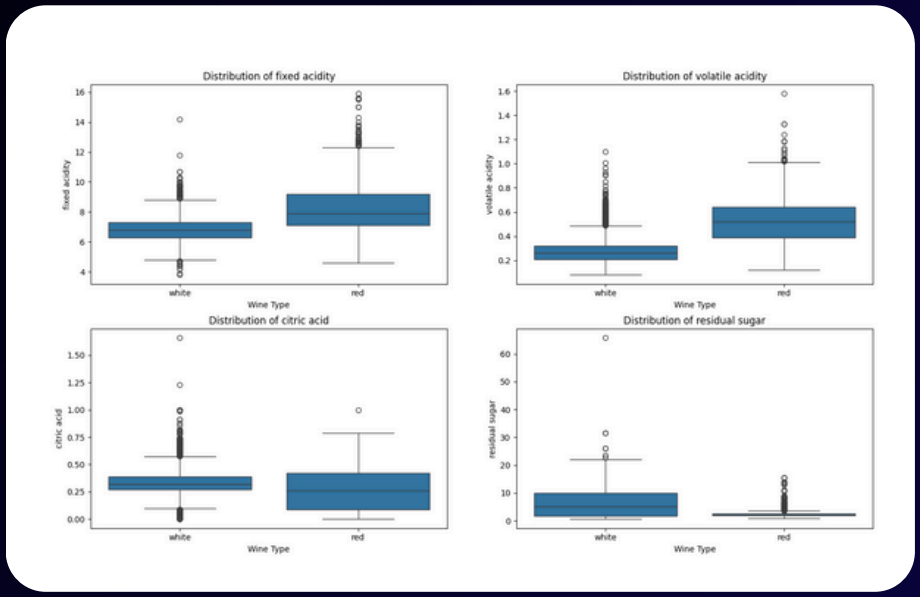
Boxplot



aca nos parecio interesante ver como la calidad de los vinos no varia tanto entre tipo de vino, con la excepcion de que el vino blanco si llega a tocar el 9, mientras que el valor maximo de calidad de vino rojo es de 8

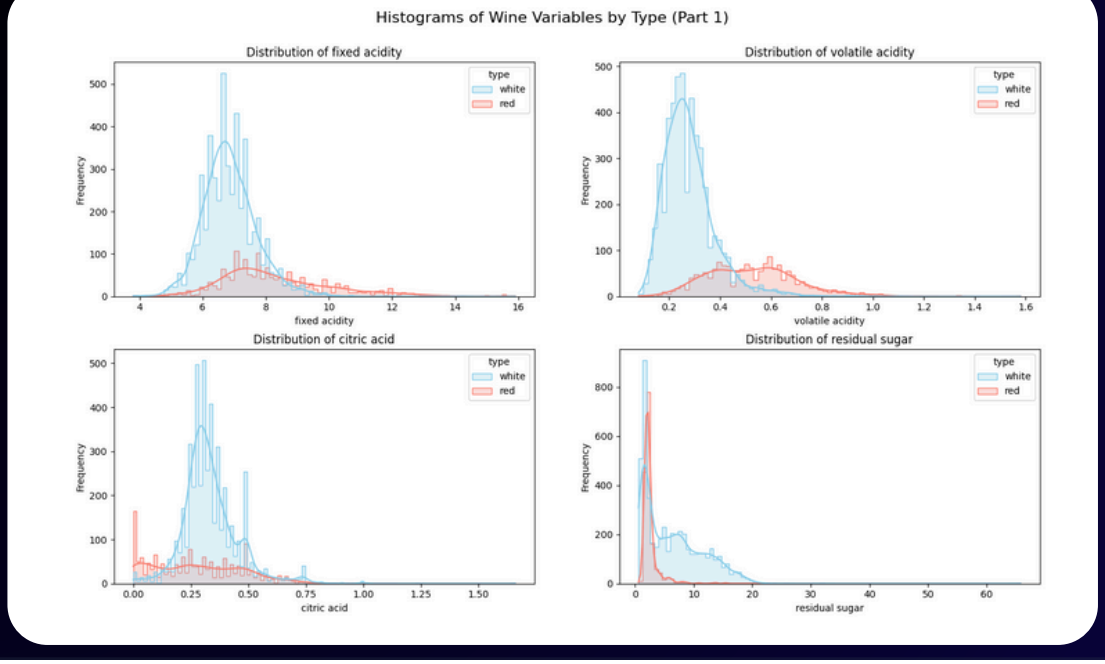
Análisis exploratorio

Boxplot



aca podemos ver como fixed acidityм volatile acidity y residual sugar pueden ayudarnos a diferenciar el tipo de vino

Análisis exploratorio

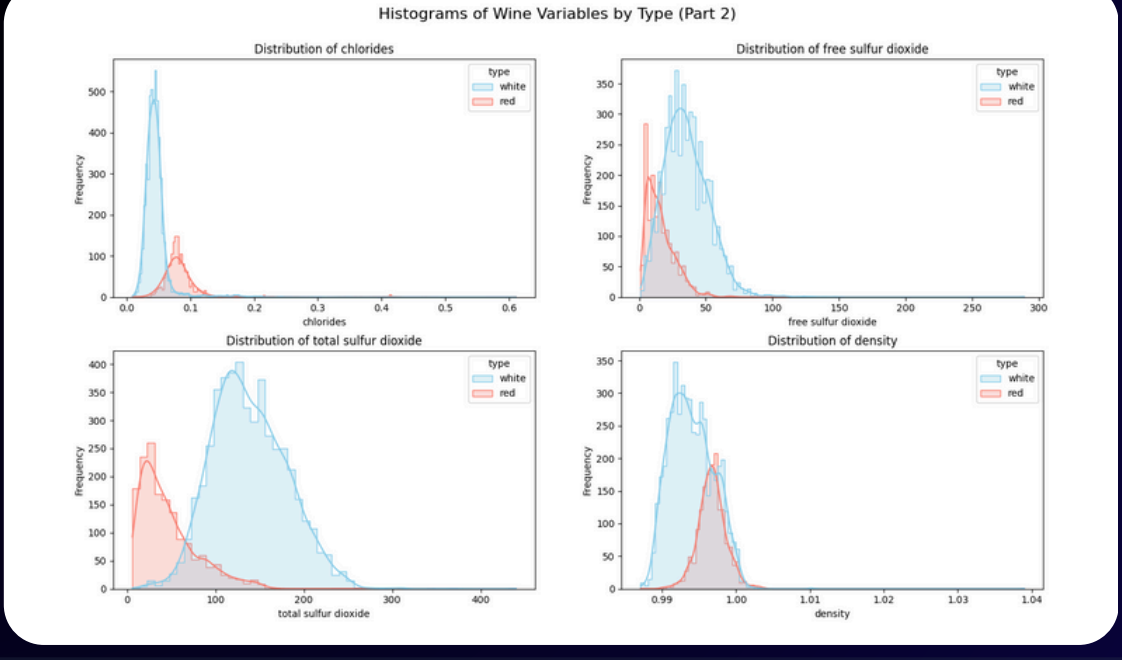


Ver las distribuciones nos ayuda a ver como afecta el tipo de vino a la distribucion de los valores. A diferencia de los boxplots, nos ayudan a ver con mas claridad la frecuencia de los datos.

Hay que destacar que hay 1599 muestras de vino rojo contra las 4898 de vino blanco. Esto es importante ya que va a hacer que la curva de vino blanco siempre este por arriba en frecuencia que la de rojo. Aun asi, lo que nos importa es ver la forma de las curvas, por lo que la diferencia en las escalas de las frecuencias no es algo que nos afecte.

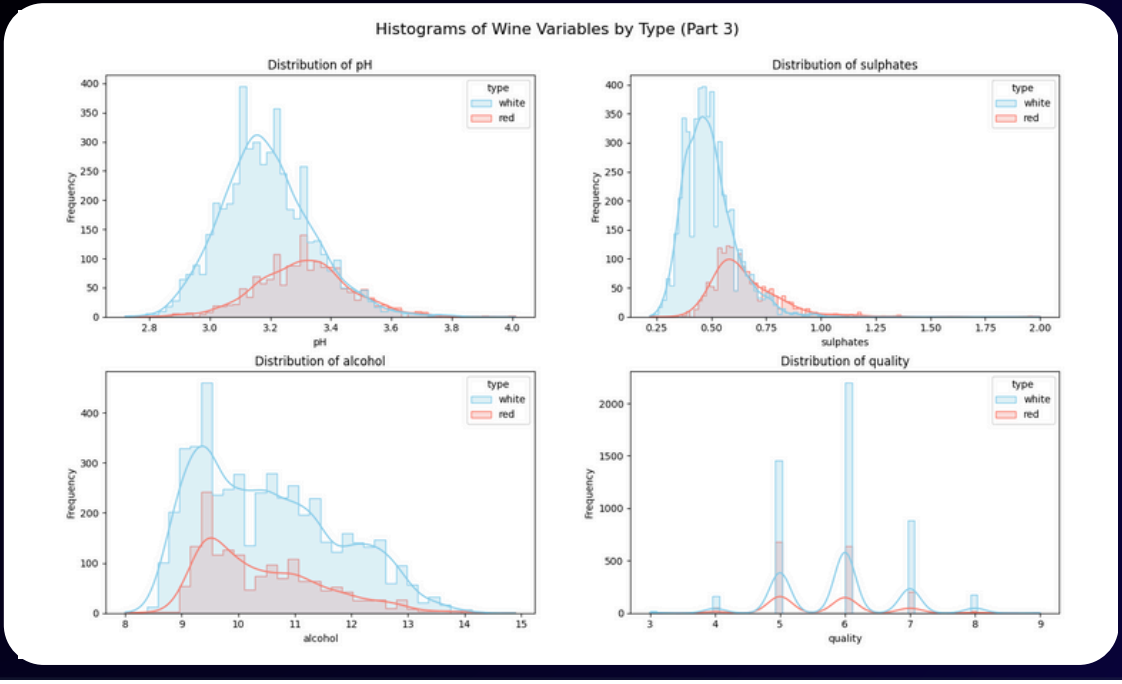
En estos graficos nos sorprendio como el vino rojo tiene una distribucion mucho mas acotada de residual sugar, mientras que las muestras de vino blanco ofrecen mayor variedad en su contenido de azucar residual.

Análisis exploratorio



En estos graficos no hay tantas diferencias en las distribuciones, pero si en los valores. Vemos como claramente en chlorides y total sulfur dioxide las muestras se separan por tipos de vinos, mientras que en free sulfur dioxide y density no hay tanta separacion entre las curvas

Análisis exploratorio

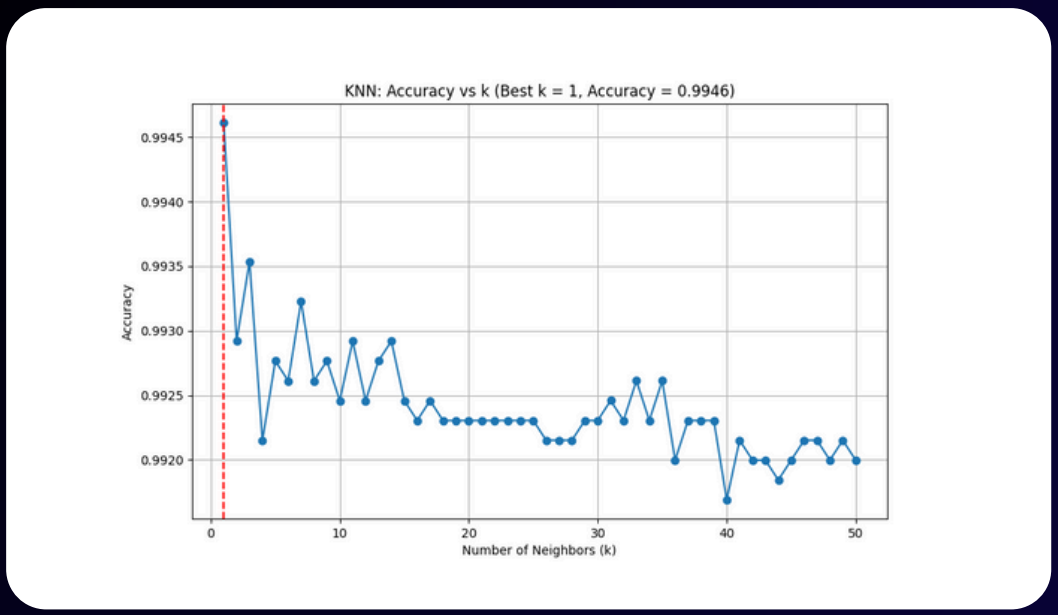


Aca demos que tanto para ph como sulphates, el vino rojo tiene una media mayor a vino blanco, mientras que en alcohol la distribucion es sorprendentemente parecida, aunque esto puede deberse a que estos vinos provinenen todos de una misma empresa.

En cuanto a calidad, vemos distribuciones similares, con el vino blanco teniendo una media un poco superior

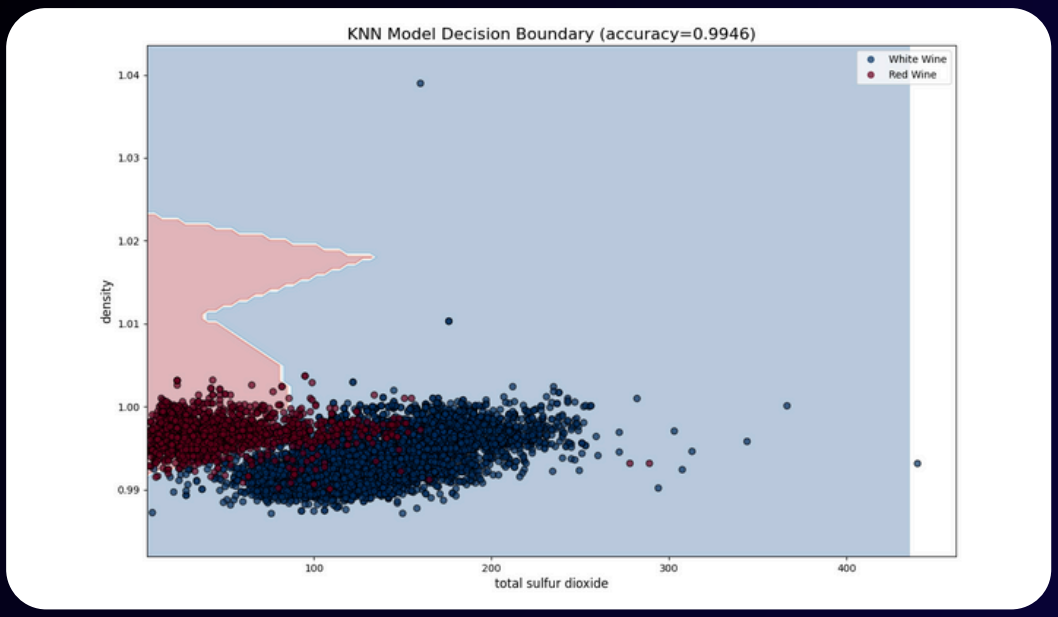
Vemos que las variables que mas influyen son credit-history, saving-account-amount, la tasa de interes, la duracion del credito y la antiguedad en el trabajo

Predecir color: KNN



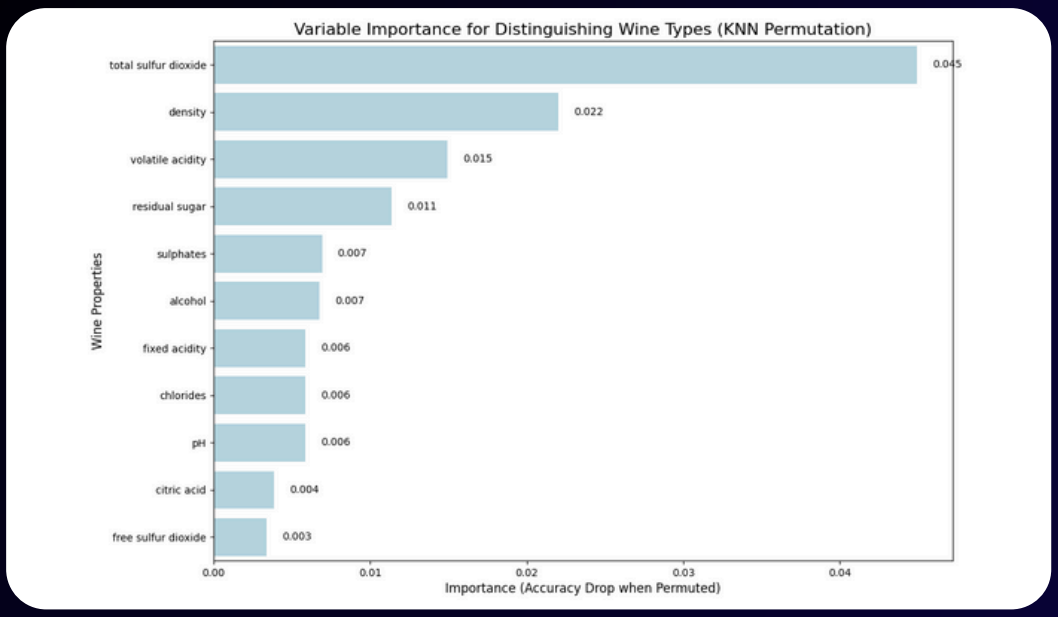
para predecir el color hicimos un knn. La mejor accuracy en base a las k fue de 0.9946

Predecir color: KNN



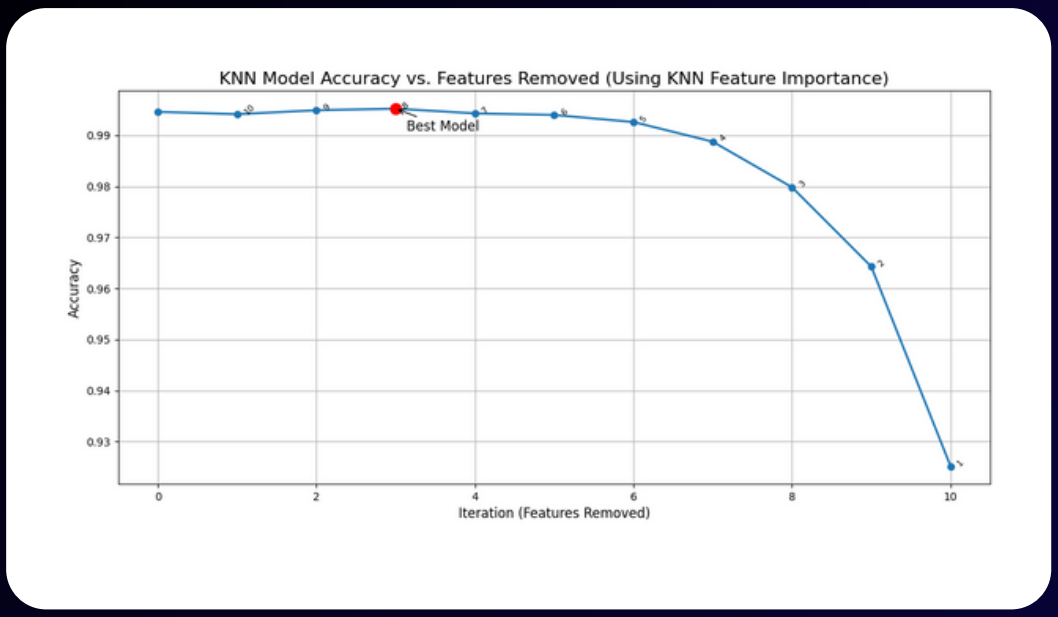
ploteamos las fronteras de decision para ver como predecia el modelo. Vemos que en este caso afeta mas el valor de x (total sulfur dioxide) que la y. Usamos esas variables por ser las mas significativas (siguiente slide)

Predecir color: KNN



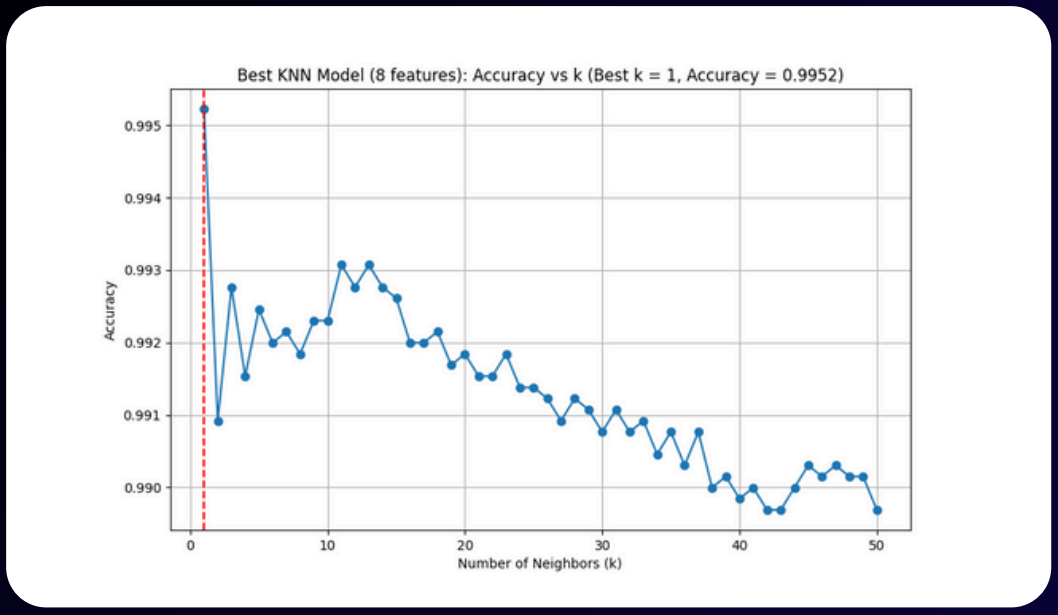
medimos las variables mas significativas para el modelo

Predecir color: KNN



analisis de sensibilidad en base a las variables mas importantes segun el modelo. Vamos eliminando de a una variable y evaluando el modelo. El orden elegido para borrar es de menos significativas a mas significativas

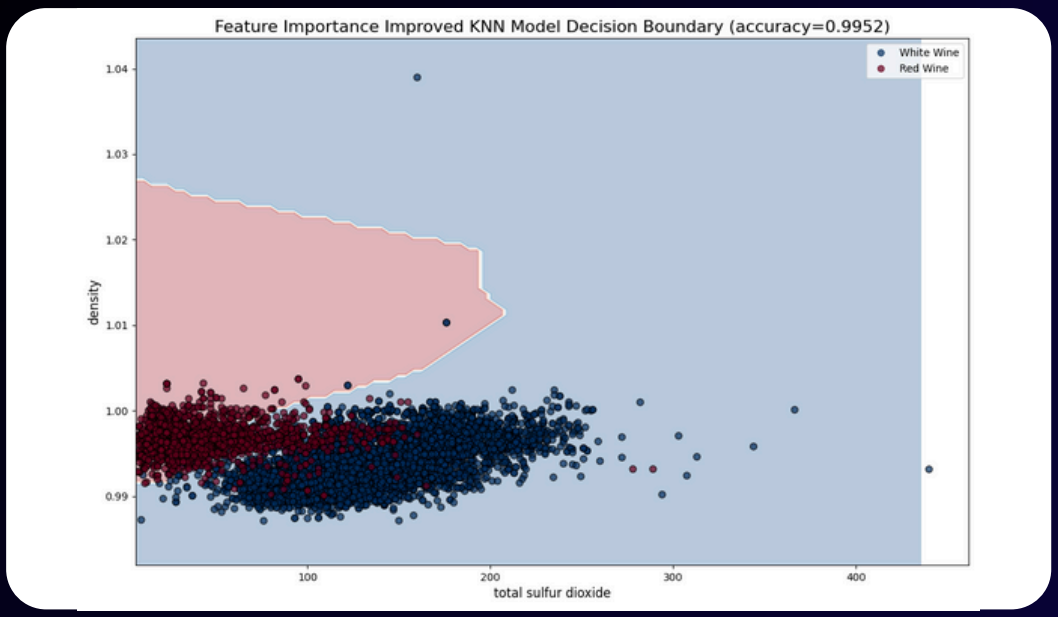
Predecir color: KNN



medimos el accuracy vs k del modelo con la cantidad optima de variables.

el accuracy mejoro con respecto al modelo anterior

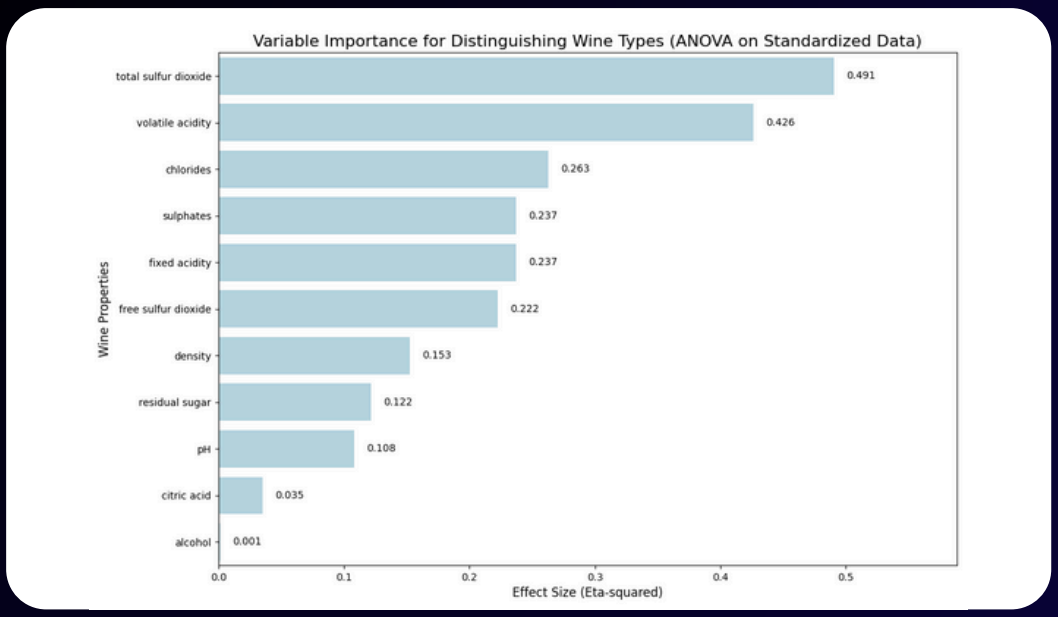
Predecir color: KNN



ploteamos las fronteras de decision para el modelo optimo por cantidad de variables segun el modelo.

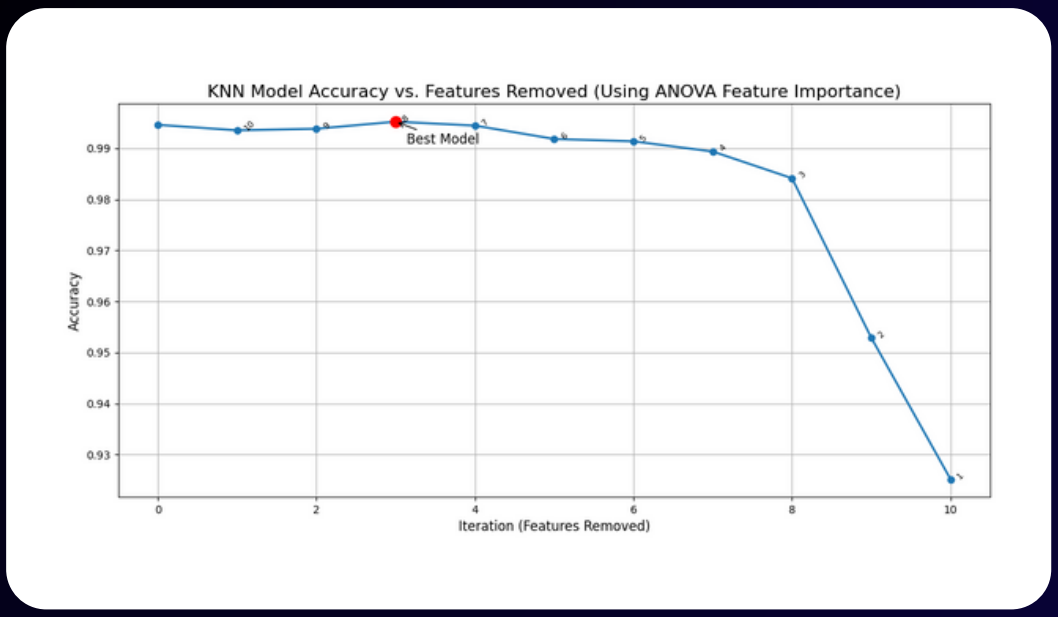
Vemos pocos cambios con respecto a las fronteras de decision del modelo completo, sigue influyendo mucho el total sulfur dioxide

Predecir color: KNN



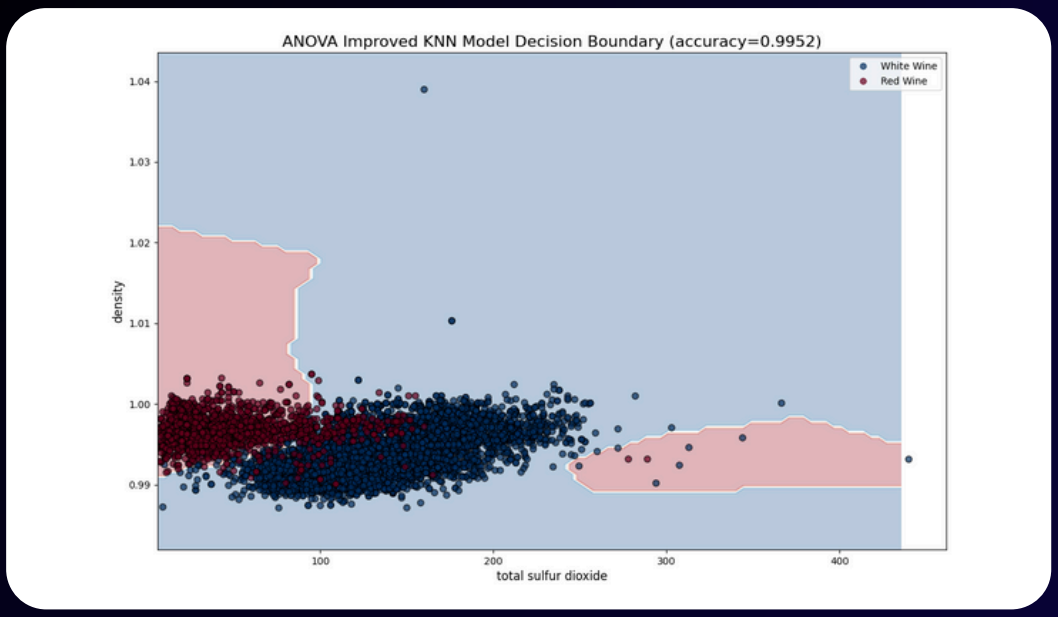
hacemos ANOVA para evaluar la importancia de las variables a la hora de predecir el color

Predecir color: KNN



con el orden de las variables de anova, hacemos analisis de sensibilidad, sacando de a una las variables para encontrar el punto con mas accuracy

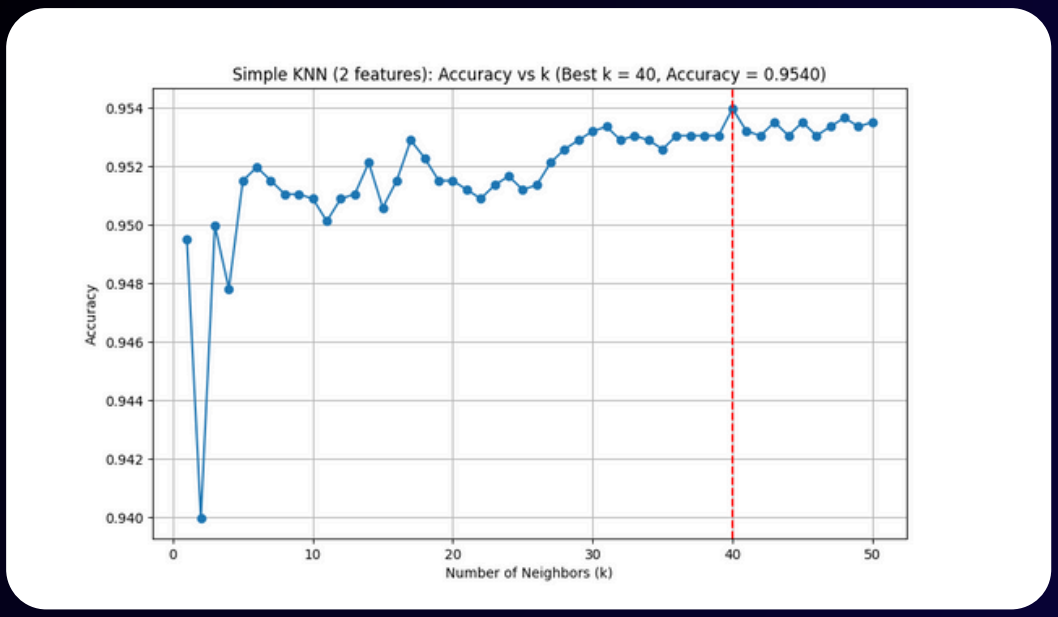
Predecir color: KNN



ploteamos las fronteras de decision. La realidad es que este grafico no nos es de mucha utilidad ya que al haber muchas variables este grafico no es de mucha utilidad.

no hubo mejora con respecto al anterior

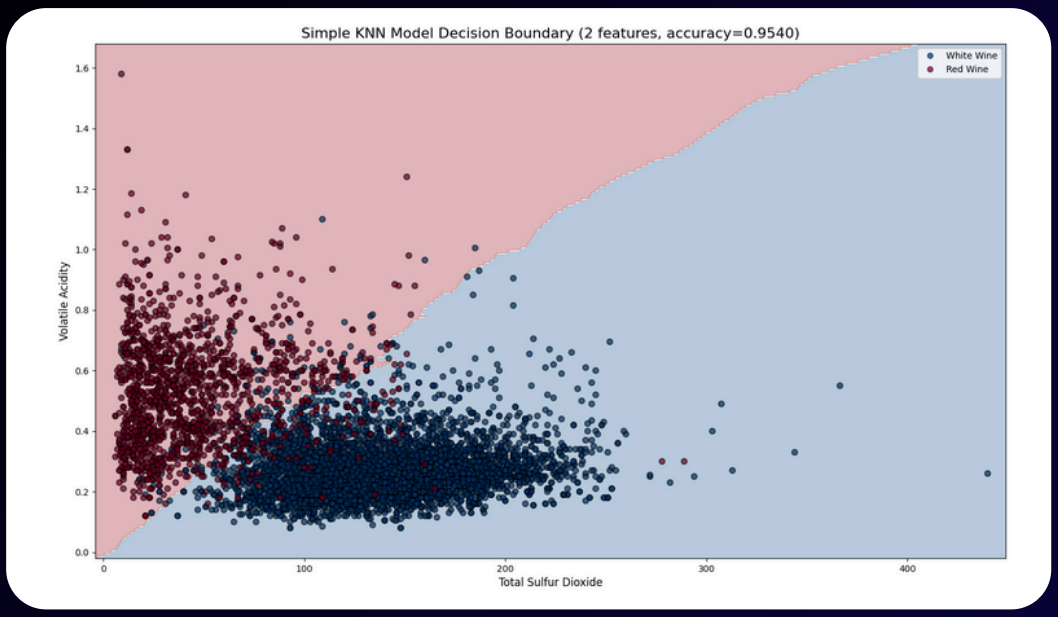
Predecir color: KNN



decidimos hacer un modelosimple con dos variables (total sulfur dioxide y density) para poder hacer un grafico de fronteras de decision que sea mas interpretable

el accuracy empeoro en comparacion a todos los anteriores

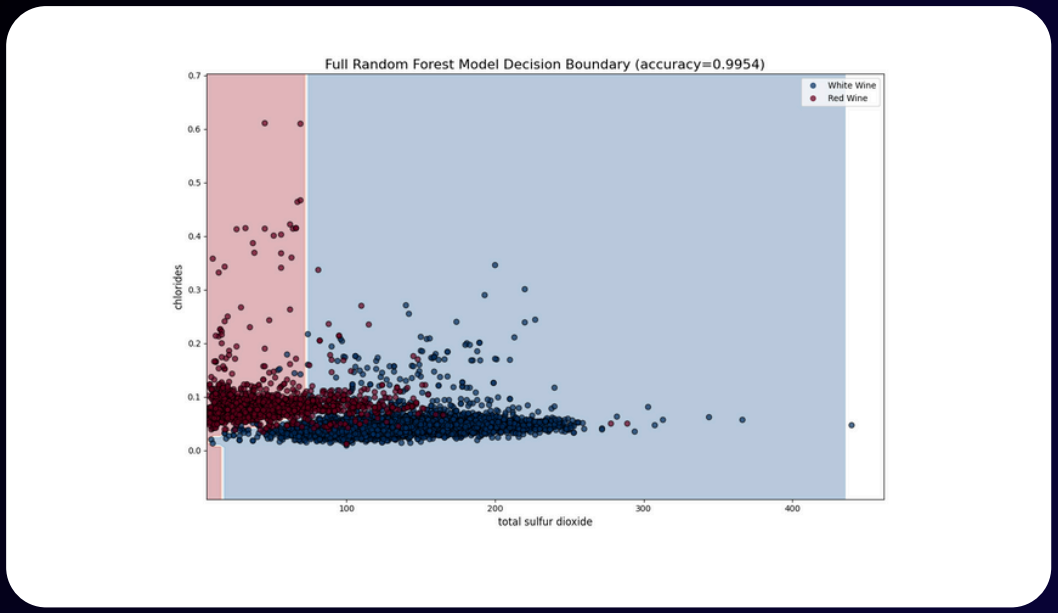
Predecir color: KNN



ahora si tenemos un grafico de fronteras de decision 100% fiel. Lo que vemos es que la frontera de decision se asemeja a una linea x = y (linea diagonal que arranca en 0,0 y crece con un m=1). Esto nos da la pauta que en general los vinos blancos son los que tienen mayor sulfur dioxide y menor volatile acidity, y que los vinos rojos tienen menos total sulfur dioxide y mas volatile acidity

decidimos hacer random forest tambien para determinar el color

Predecir color: Random Forest

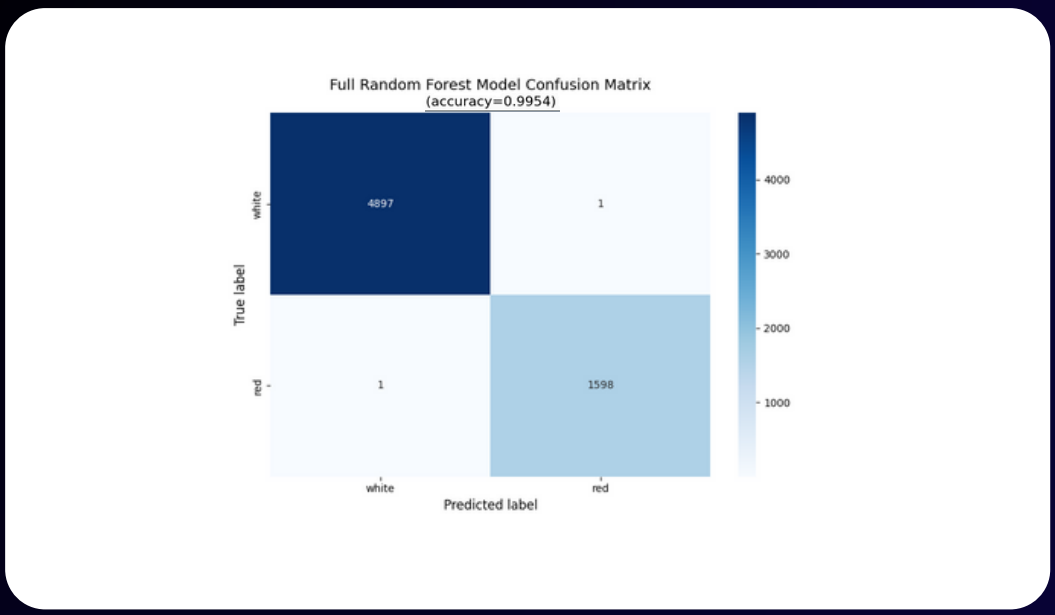


hicimos un grafico de frontera de decision, pero no es muy interpretable. Las variables en los ejes son las mismas que antes.

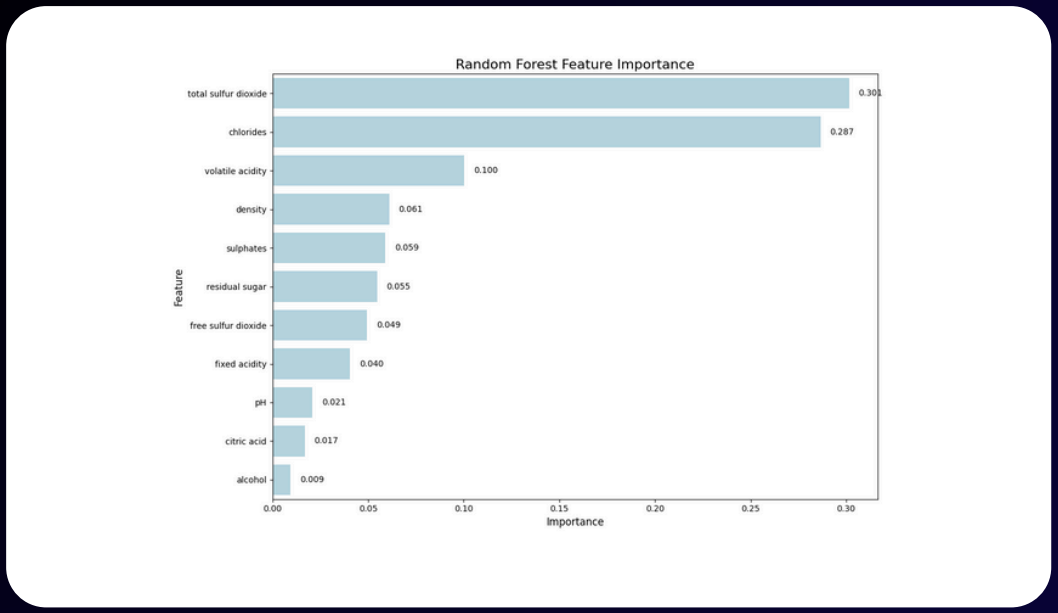
El accuracy esta en linea con los modelos de KNN

Nuestra forma de trabajar fue igual a knn, solo que no hicimos el modelo "simple" de dos variables

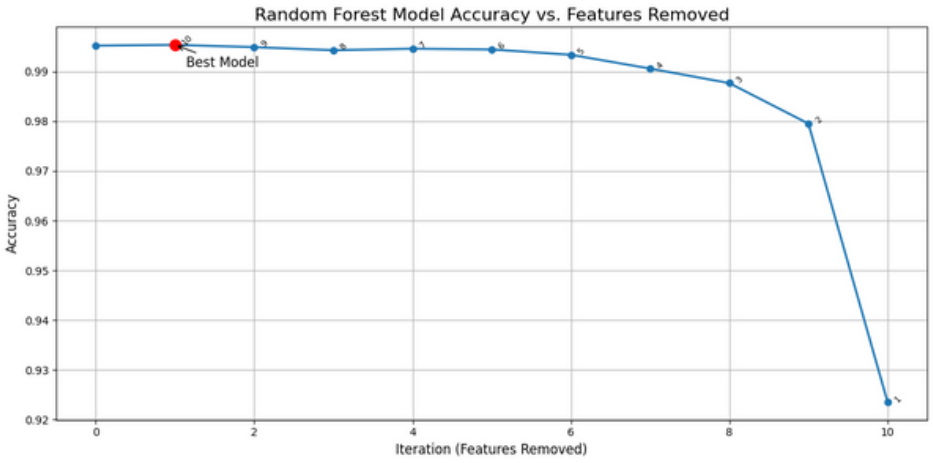
Predecir color: Random Forest



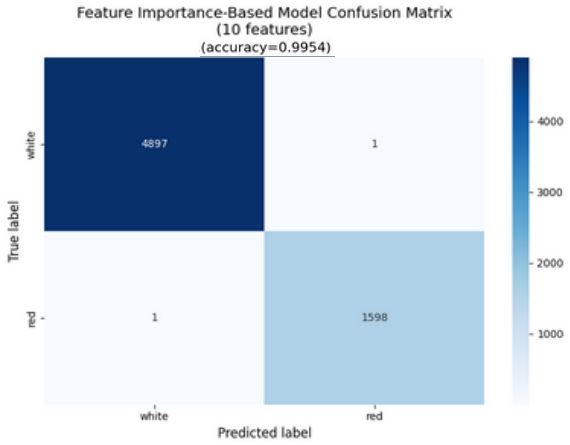
Predecir color: Random Forest



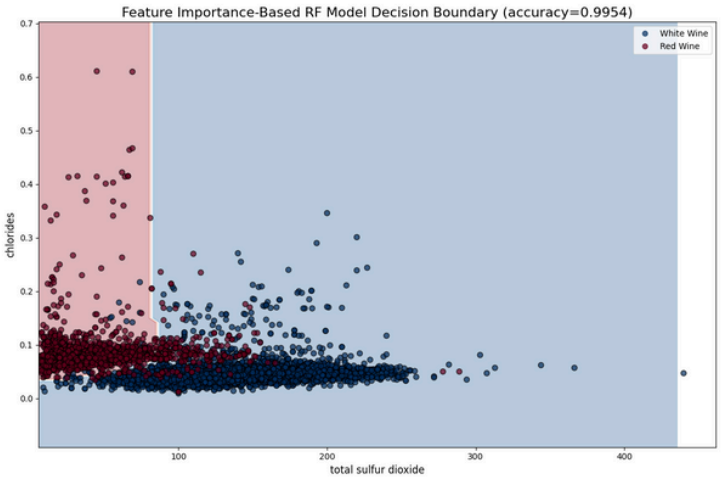
Predecir color: Random Forest



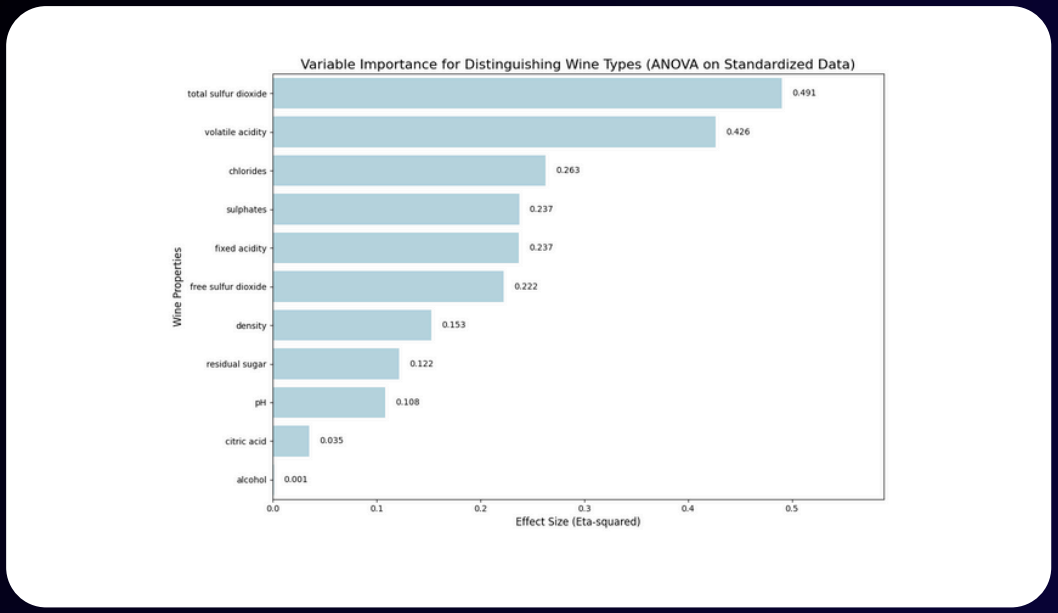
Predecir color: Random Forest



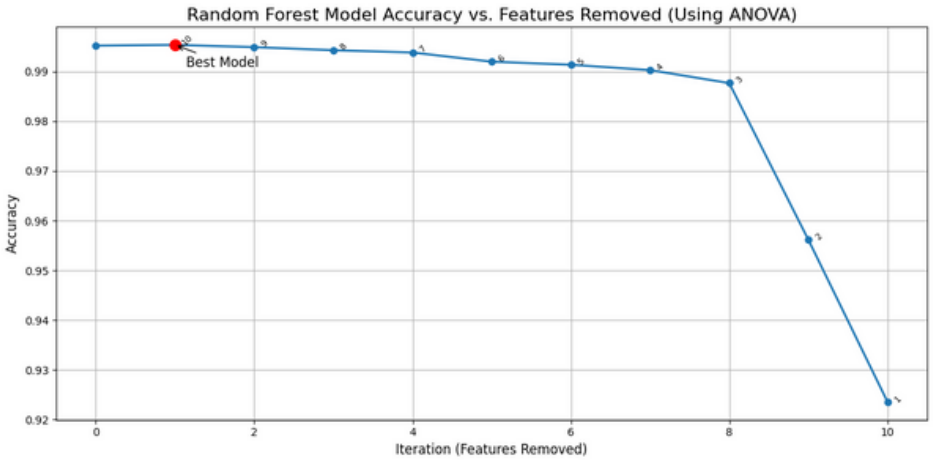
Predecir color: Random Forest



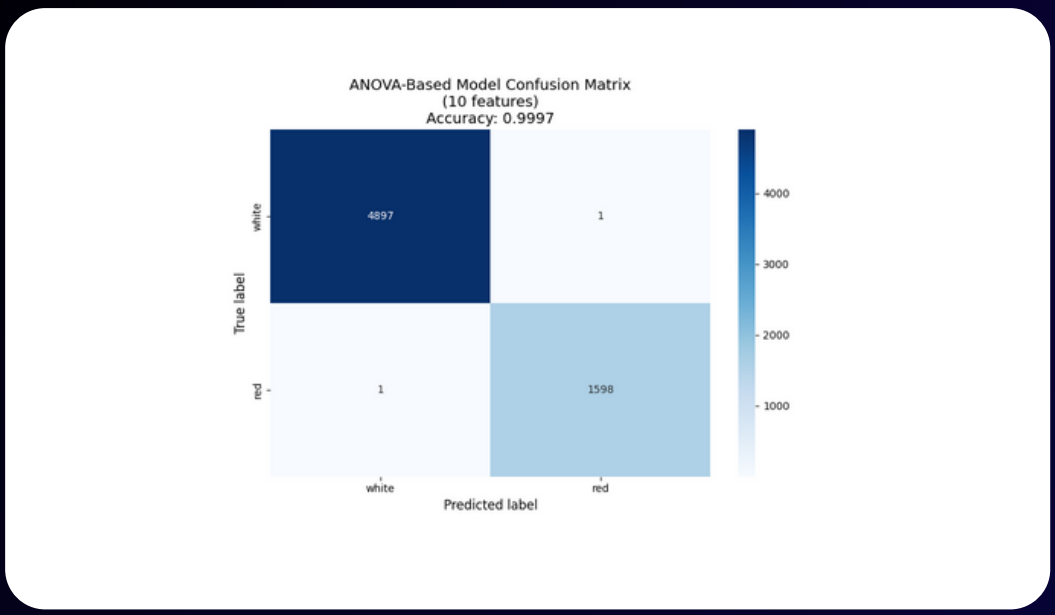
Predecir color: Random Forest



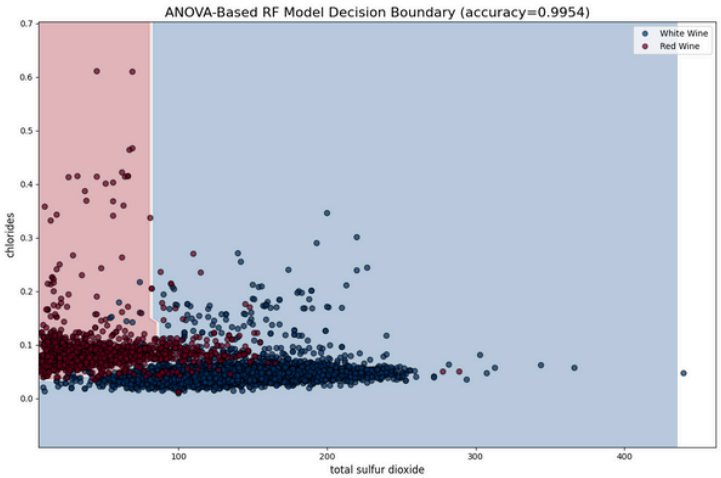
Predecir color: Random Forest

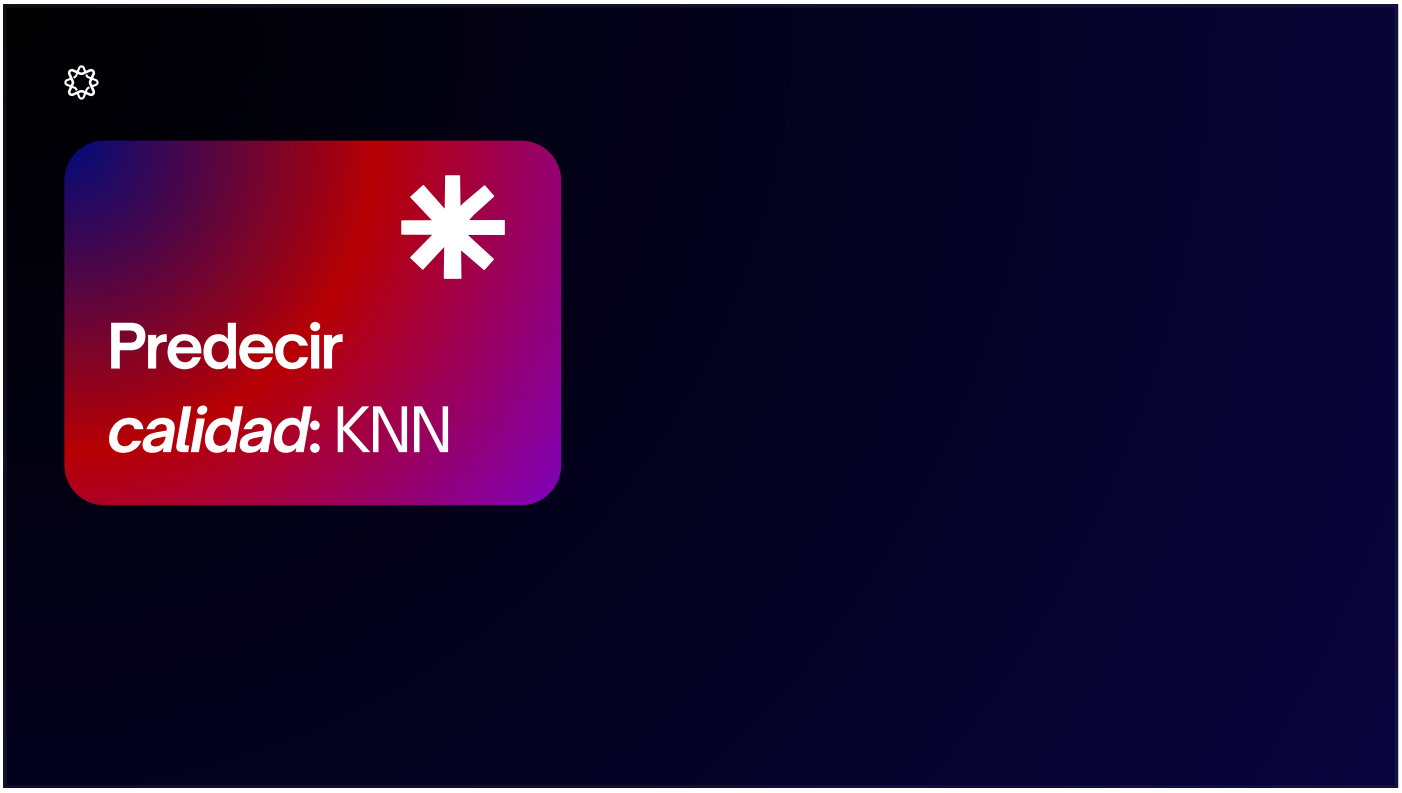


Predecir color: Random Forest



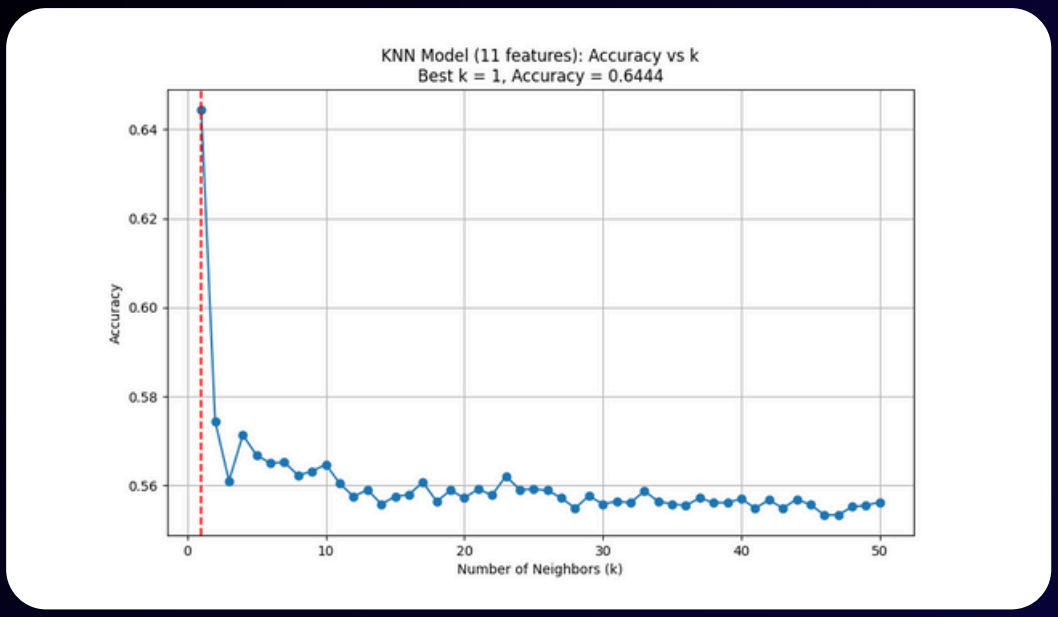
Predecir color: Random Forest



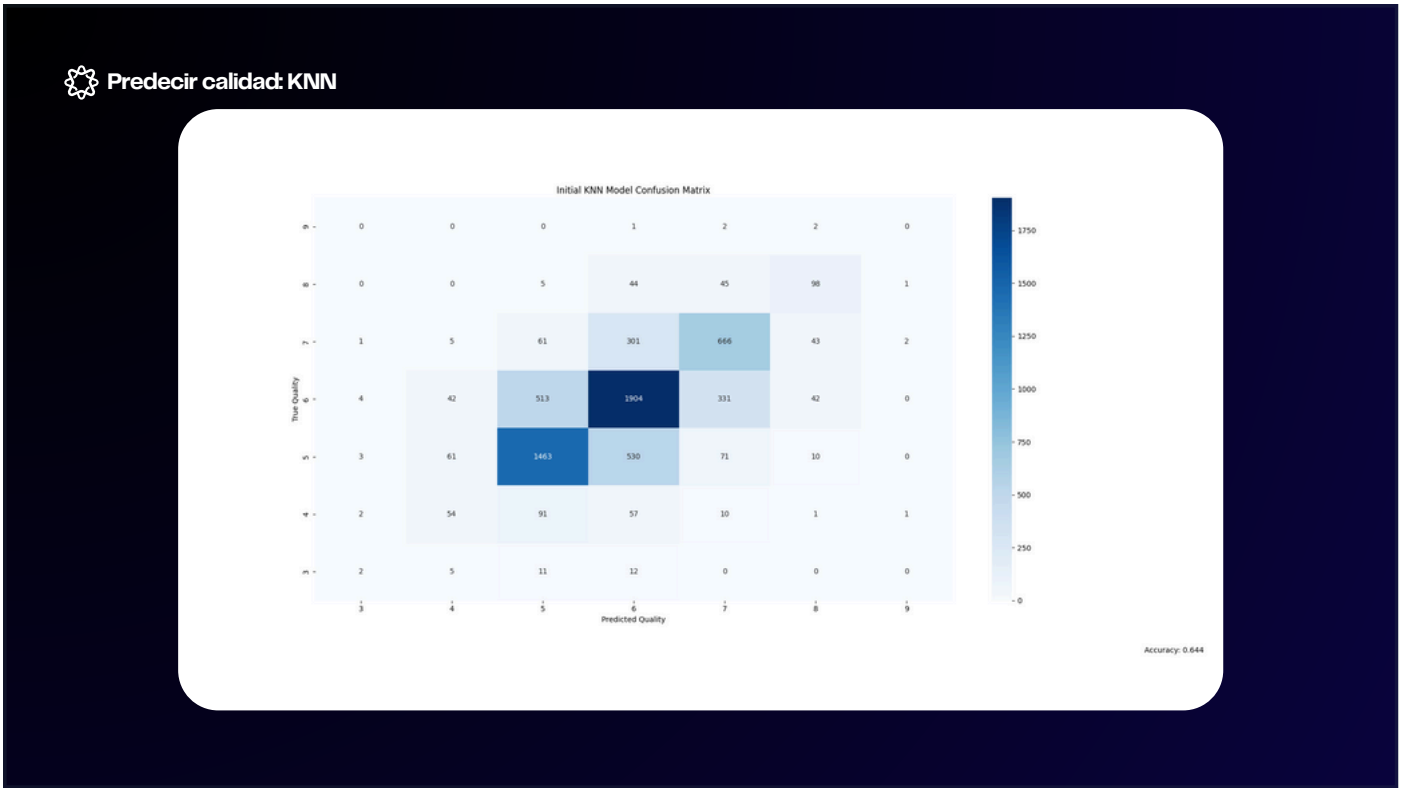


Para predecir calidad, empezamos con un KNN

Predecir calidad: KNN

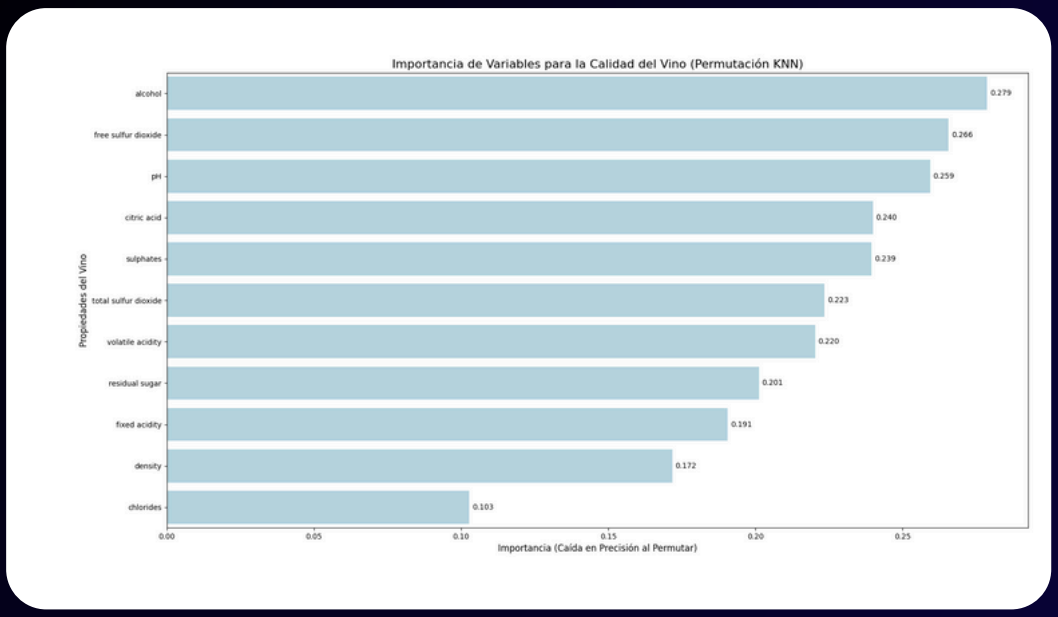


Vemos que la accuracy disminuye significativamente con respecto al analisis de color. Esto se debe a que pasamos de una variable binaria a una variable con mas resultados posibles, por lo que es logico que el accuracy del modelo baje



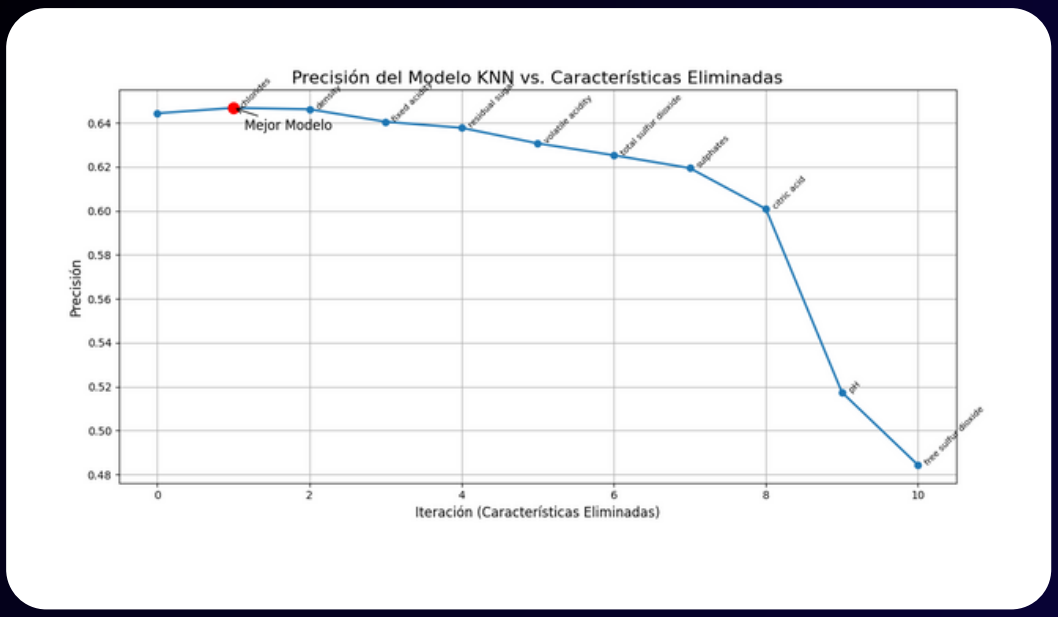
matriz de confusion del modelo knn normal

Predecir calidad: KNN



evaluamos la importancia segun el modelo para hacer analisis de sensibilidad

Predecir calidad: KNN

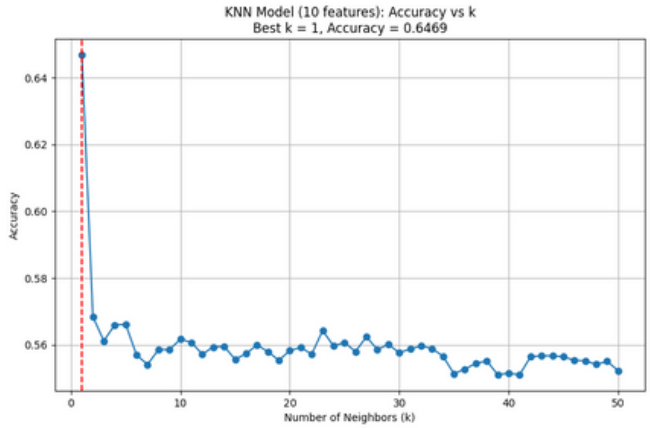


vemos que el mejor modelo implica sacar chlorides

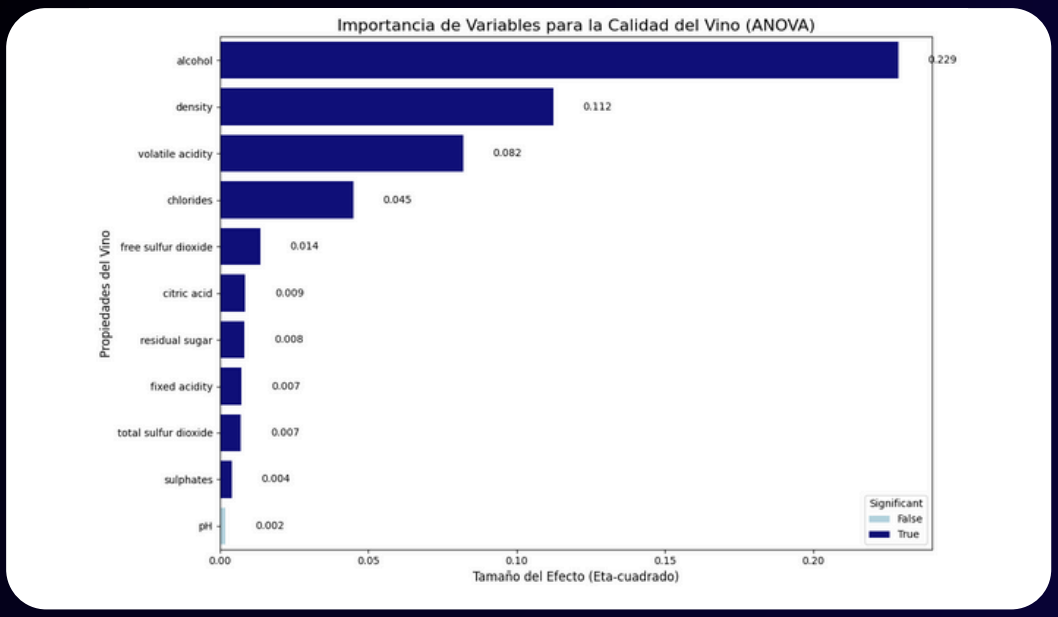
Predecir calidad: KNN

la accuracy mejora de 0.644 a 0.647, un cambio poco significativo

Predecir calidad: KNN

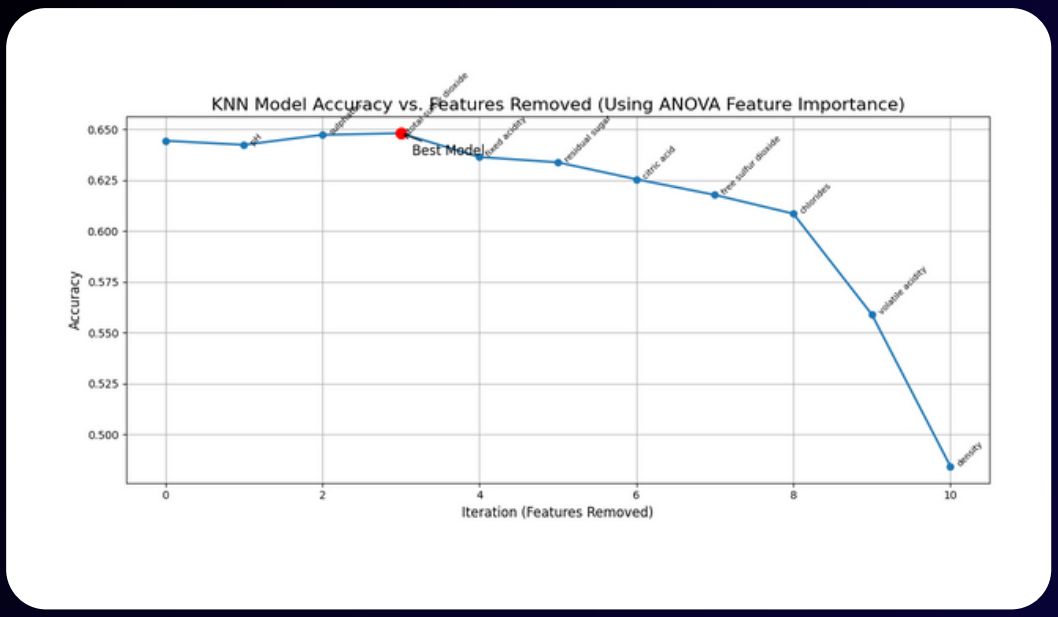


Predecir calidad: KNN



hacemos analisis de importancia en predecir calidad por anova para hacer analisis de sensibilidad

Predecir calidad: KNN

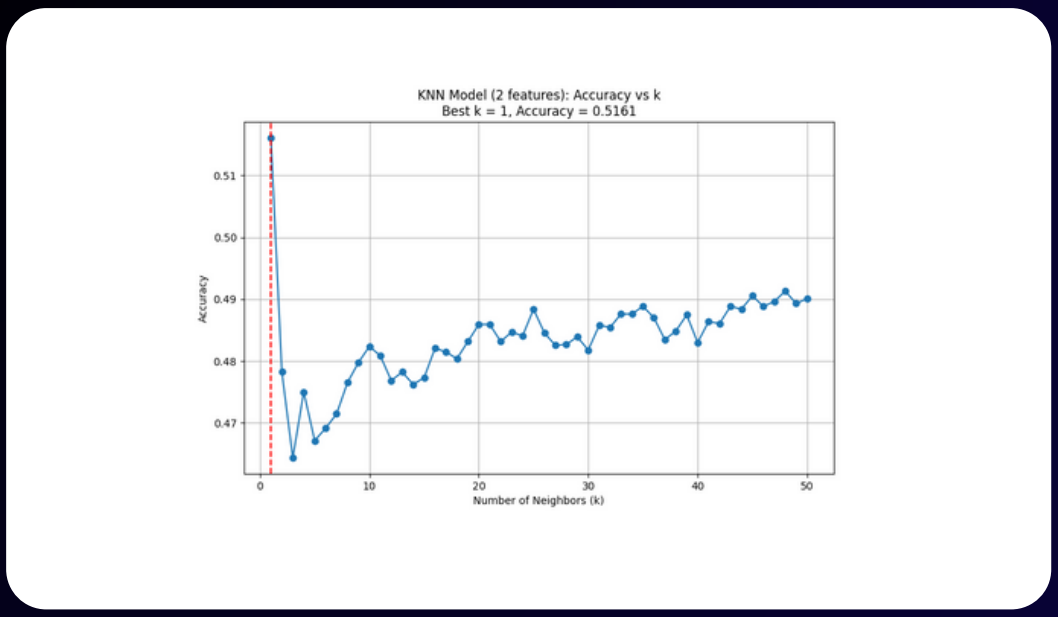


el punto optimo fue sacando ph, sulphates y total sulfur dioxide

Predecir calidad: KNN

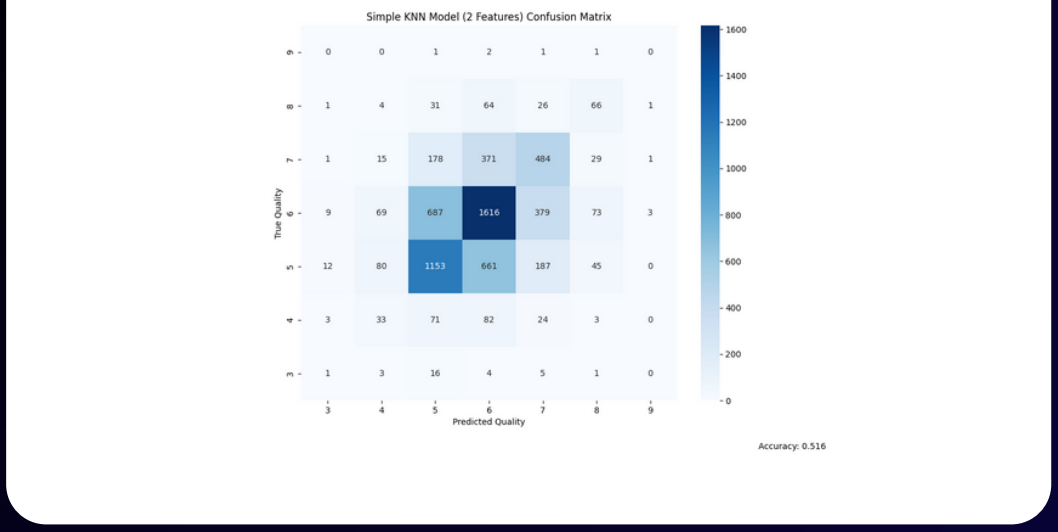
La accuracy paso de 0.647 a 0.648

Predecir calidad: KNN



modelo basico con 2 variables. Accuracy muy mala de 0.5161

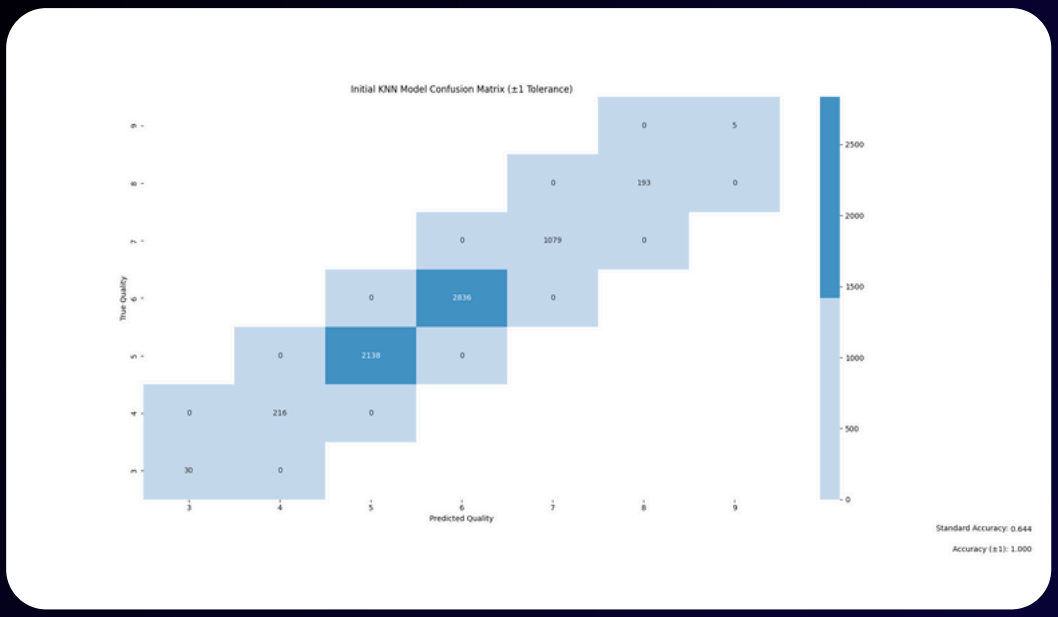
Predecir calidad: KNN



matriz de confusion del modelo simple

T=1 implica poner una "bonanza" de 1 en el modelo. Es decir vamos a tomar el resultado como valido si es mayor o menor a uno del real. Es decir si el dato real es 4, los valores aceptados como correctos con 3, 4 y 5

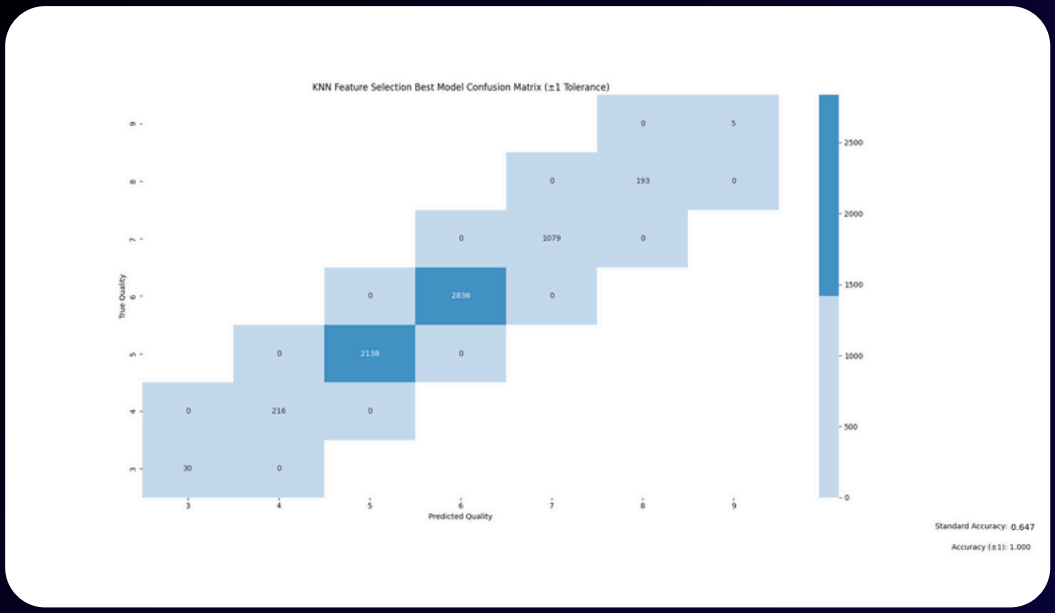
Predecir calidad: KNN con T=1



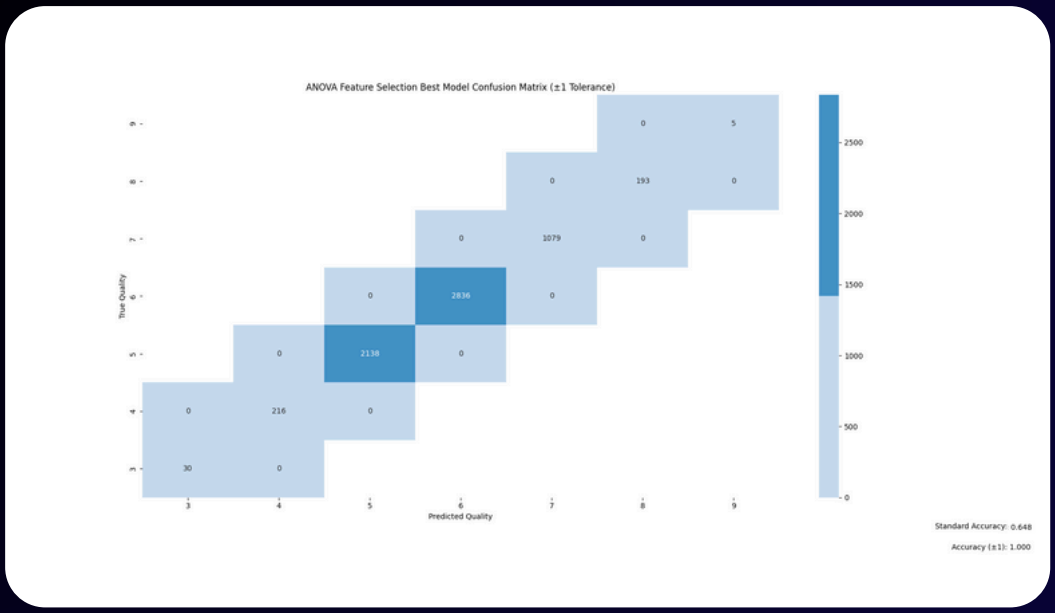
Vemos que la accuracy sube a uno al incluir esta bonanza en el modelo.

Nos sorprende que la accuracy sea de 1, ya que viendo la matriz de confusion original vemos que hay errores con mas de 1 de desvio

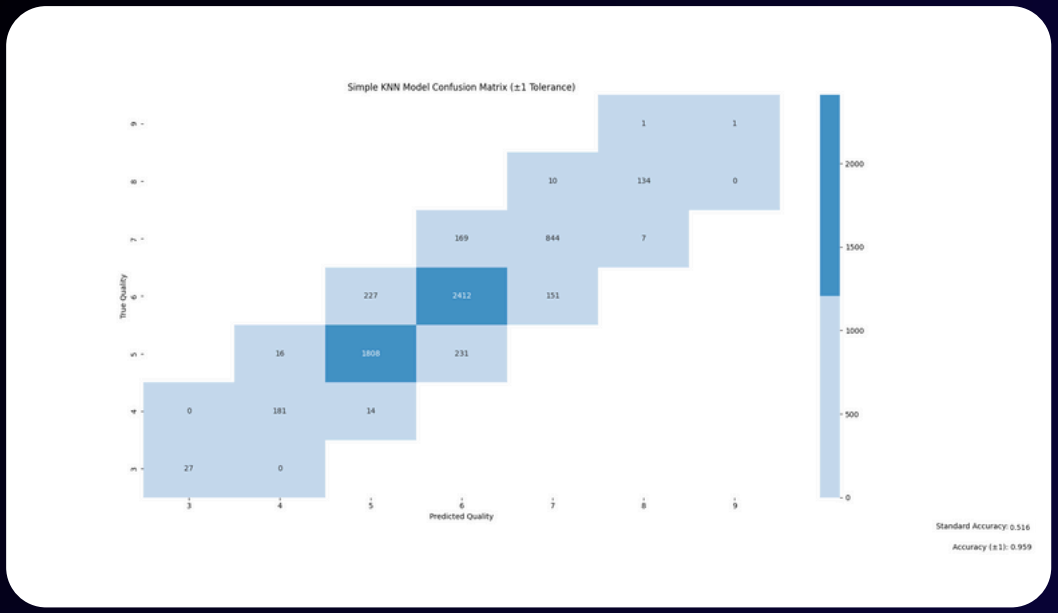
Predecir calidad: KNN con T=1

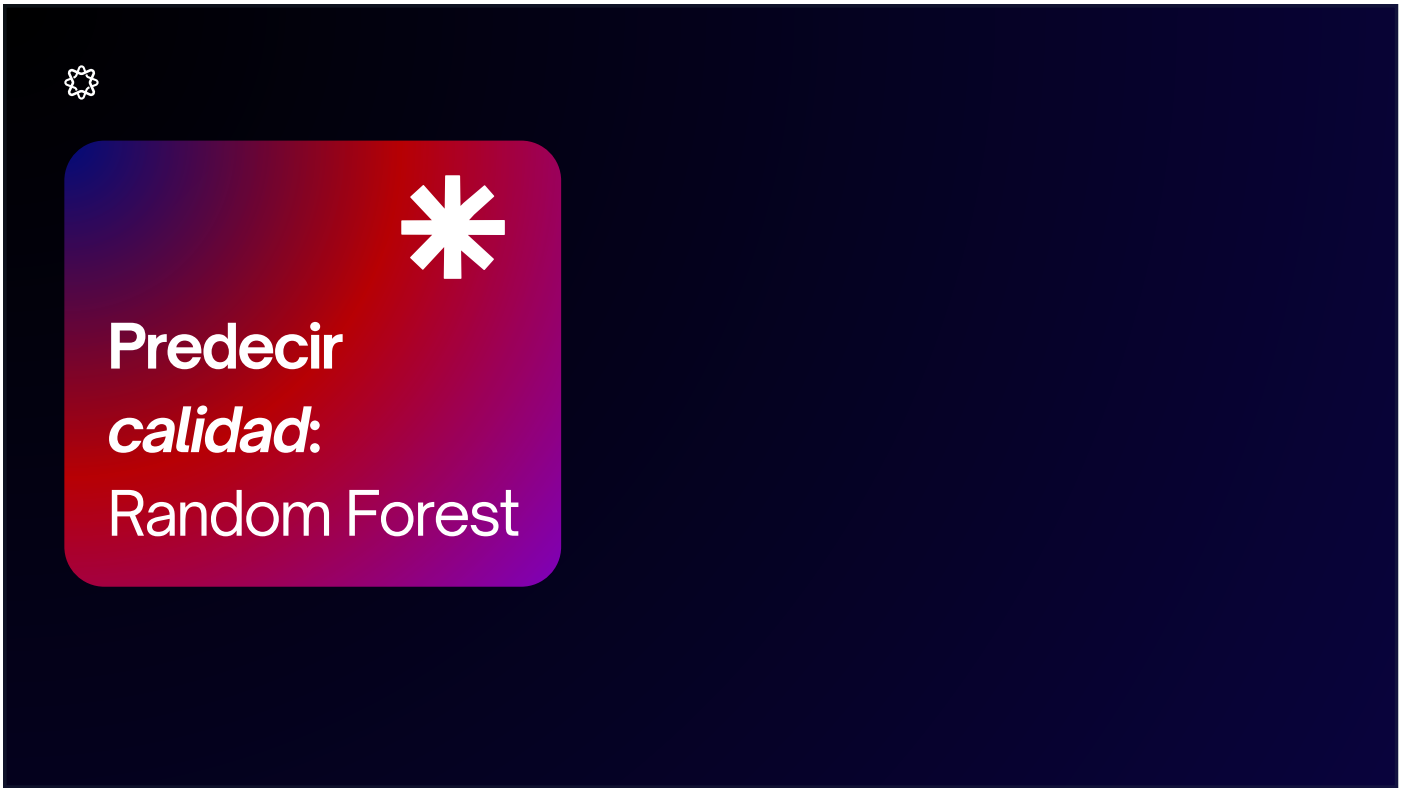


Predecir calidad: KNN con T=1

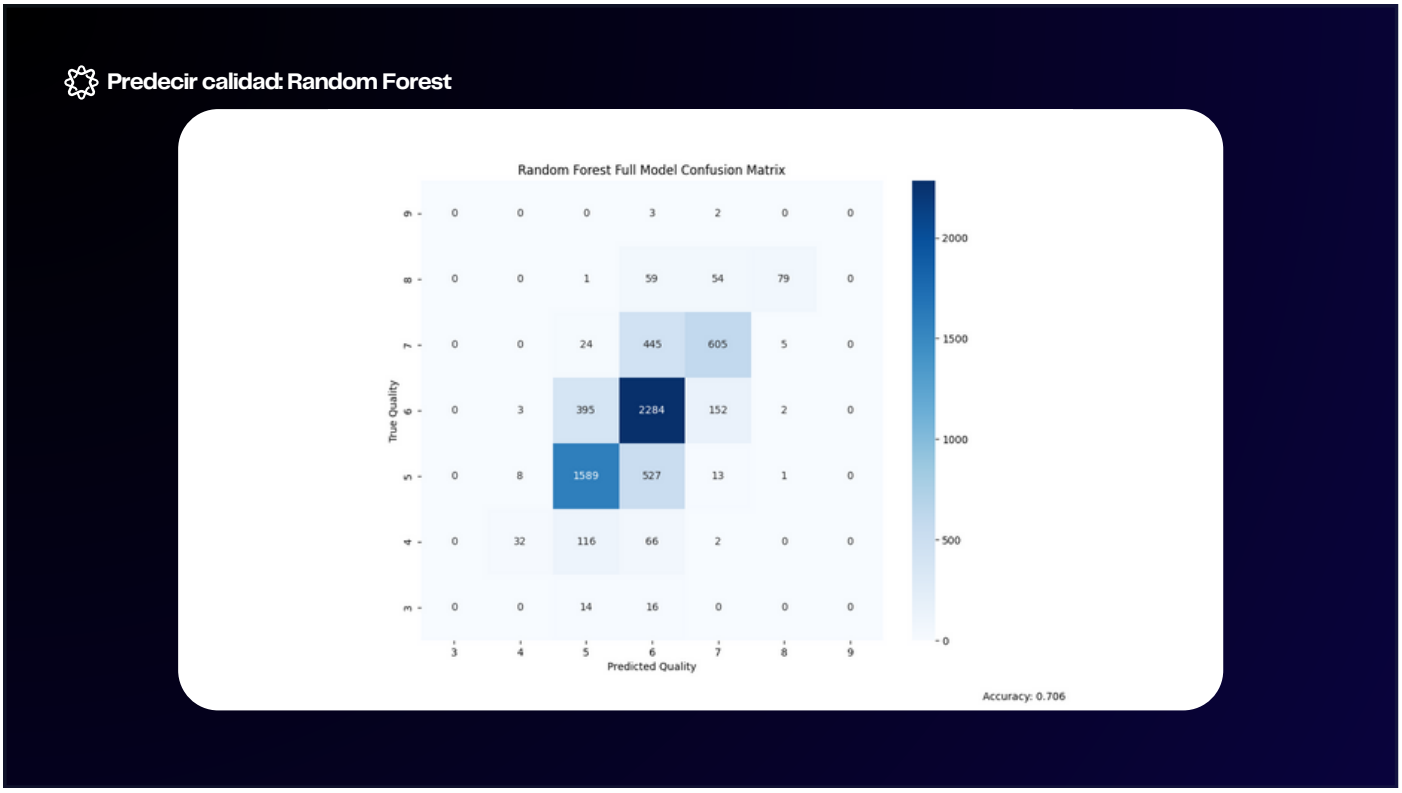


Predecir calidad: KNN con T=1



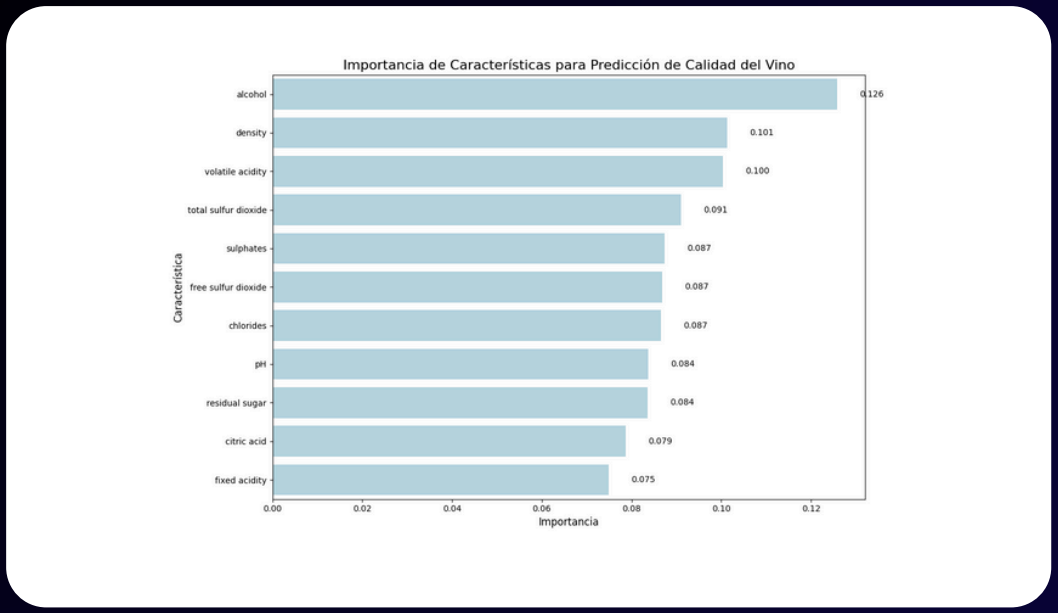


Tambien usamos random forest para predecir la calidad



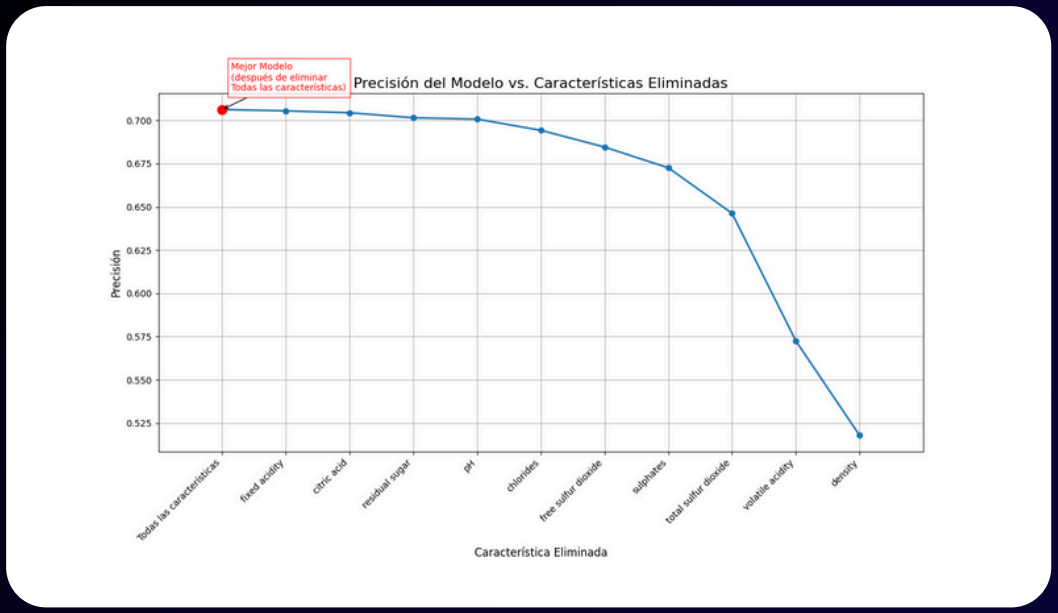
matriz de confusion al usar random forest, vemos que la accuracy es superior comparada con knn

Predecir calidad: Random Forest

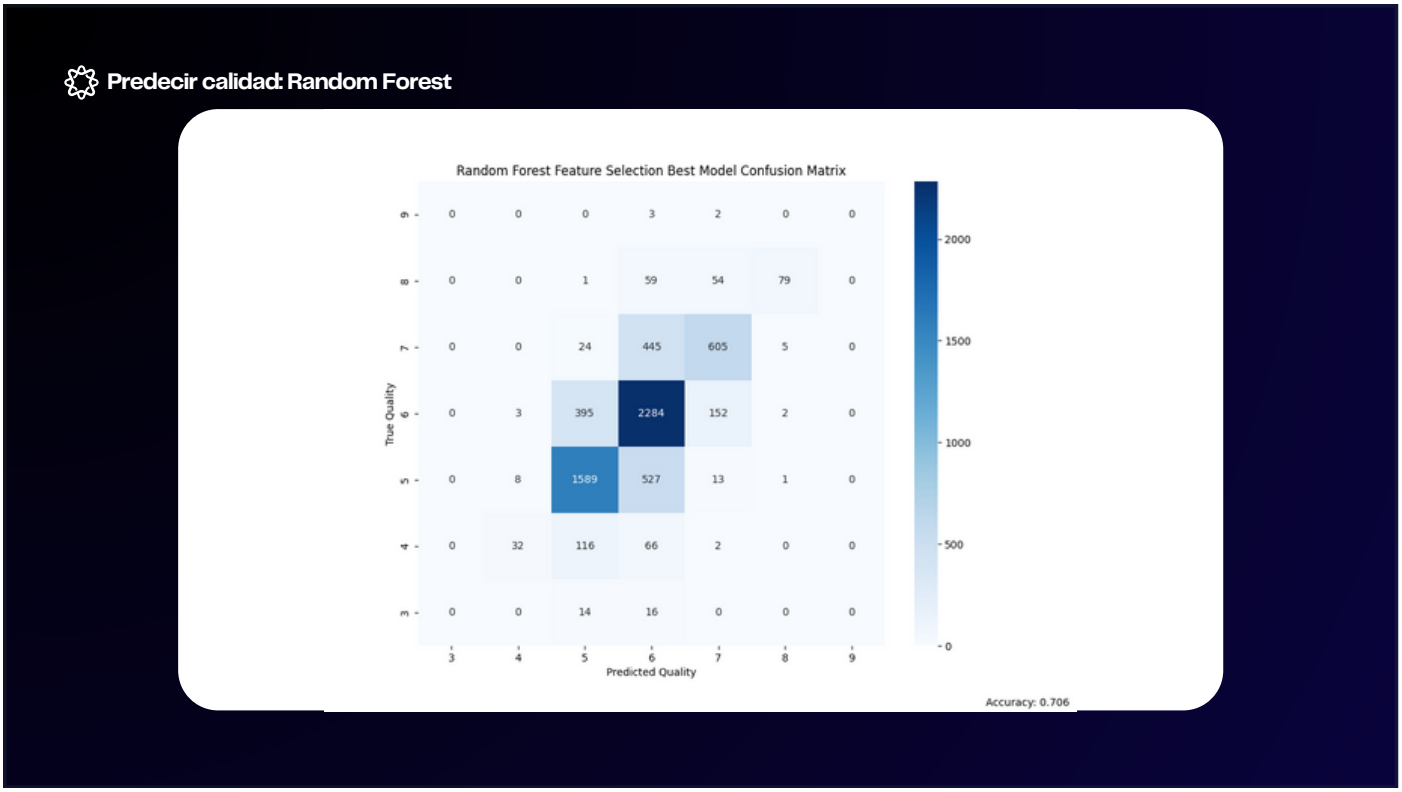


ordenamos por importancia segun el modelo para analisis de sensibilidad

Predecir calidad: Random Forest

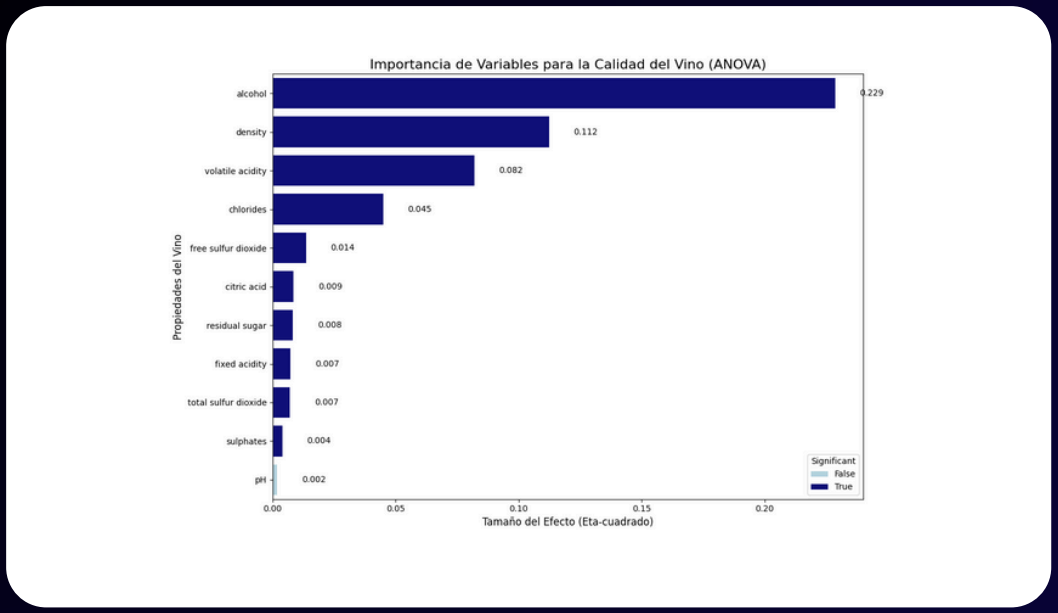


la mayor accuracy se da sin borrar



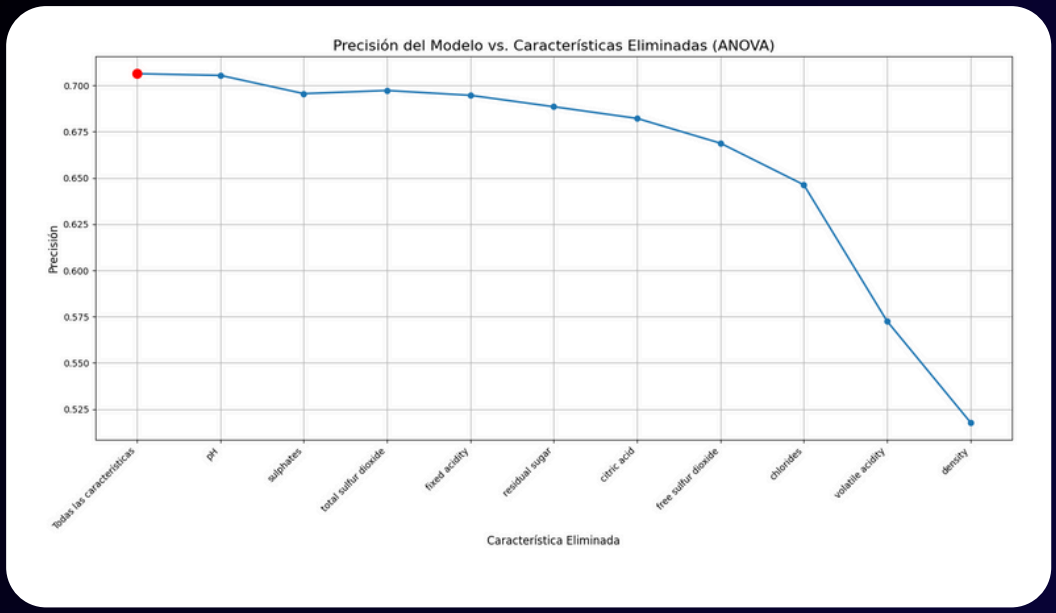
como no borramos ninguna variable, sigue igual

Predecir calidad: Random Forest

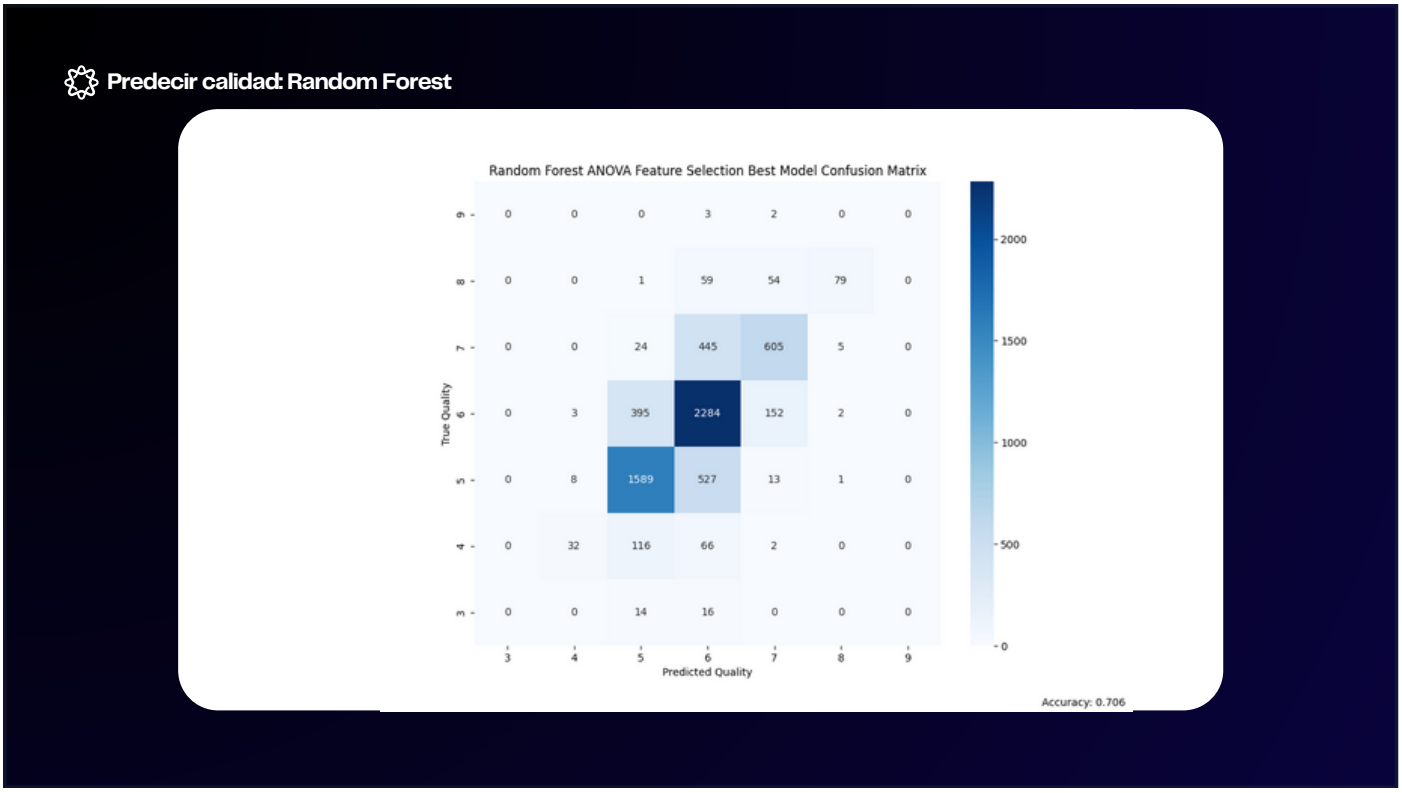


ordenamos por ANOVA para analisis de sensibilidad

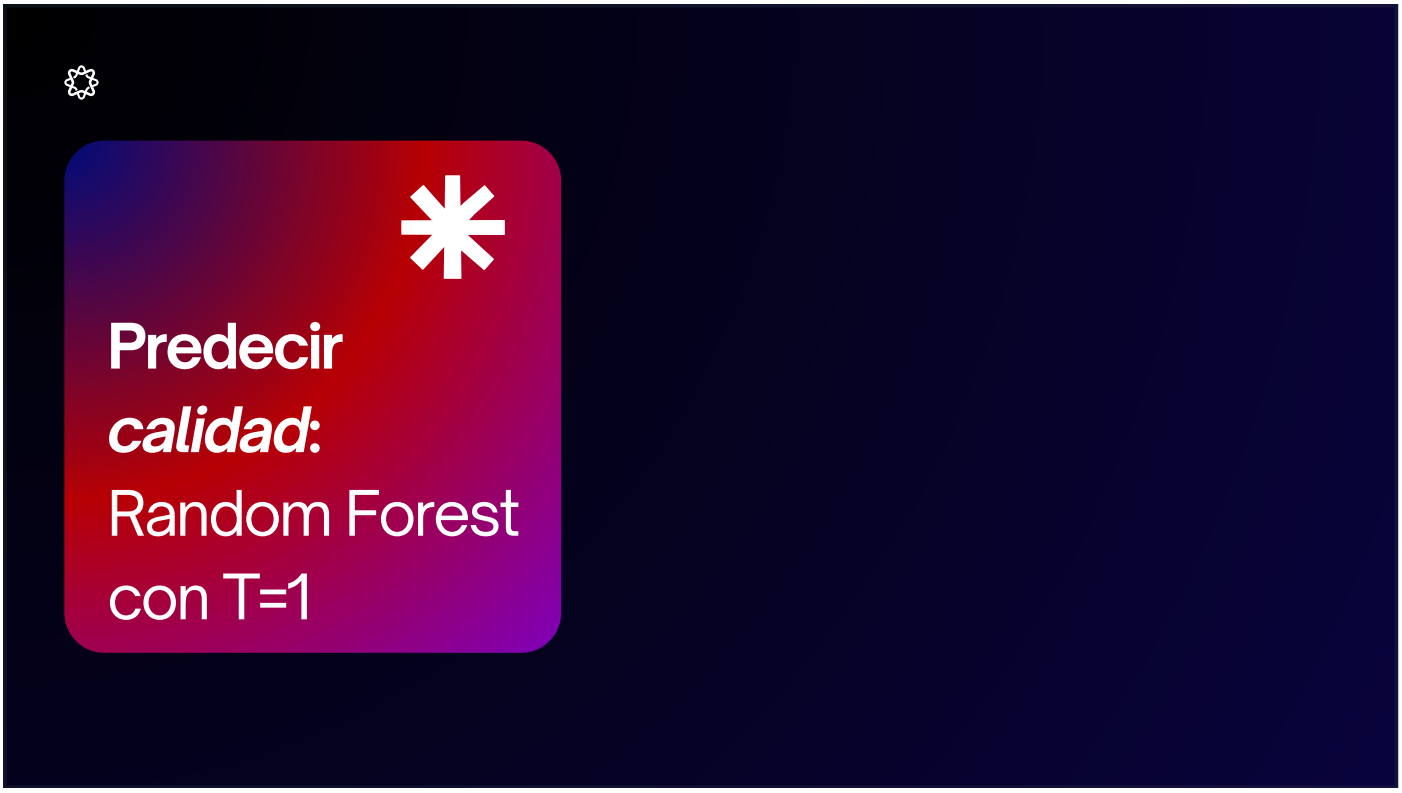
Predecir calidad: Random Forest



vemos que el mejor punto es no sacar nada

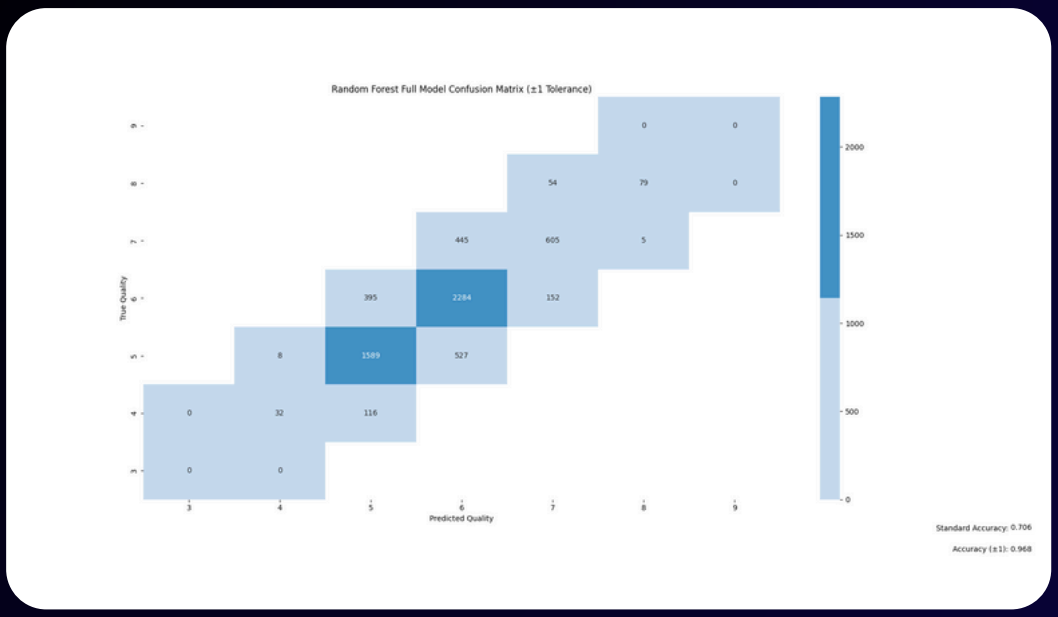


sigue igual. Lo que nos parece interesante es ver como el modelo se reduce a predecir 5,6 o 7, ya que no hay casi predicciones con 3, 4, 8 y 9



Vemos que las variables que mas influyen son credit-history, saving-account-amount, la tasa de interes, la duracion del credito y la antiguedad en el trabajo

Predecir calidad: Random Forest con T=1



Vemos que mejora significativamente al incluir una bonanza de +-1.

Nos sorprende que la accuracy sea de casi 1, ya que viendo la matriz de confusion original vemos que hay errores con mas de 1 de desvio

Gracias

Martín Quijano - Martina Coletto