UNIVERSIDAD DE GRANADA

MODELADO Y PREDICCIÓN DE LA CALIDAD DE DISTINTAS VARIEDADES DE CAFÉ UTILIZANDO TÉCNICAS DE APRENDIZAJE SUPERVISADO Y ANÁLISIS MULTIVARIANTE

Trabajo de Fin de Máter
Máster Oficial en Estadística Aplicada

Alumno: Víctor Manuel López Mendoza
Tutor: Prof. Dr. Ramón Gutiérrez Sánchez

Curso Académico 2020/2021

INDICE

INDICE DE TABLAS

INDICE DE FIGURAS

RESUMEN

El café como rubro de consumo es una de las materias primas más comercializadas en todo el mundo cuya característica distintiva es su aroma, siendo fácilmente reconocible a no pocas personas en el mundo y que va muy acompañada de otras como el cuerpo, sabor, uniformidad y dulzor. Existe literatura relevante que sostiene que estas propiedades dependen en gran medida del tipo de grano (arábica y robusta) y sus variedades (caturra, bourbon, typica, java, etc.). así pues, el presente proyecto tiene como propósito analizar la puntuación que han otorgado los catadores profesionales a distintas muestras de café y como esta se relaciona con las propiedades organolépticas, tipo y variedad del grano. Este análisis implicaría también analizar el papel de la procedencia de dicho grano, tanto por país como por continente. Para realizar este análisis se ha recurrido al uso de técnicas de aprendizaje supervisado y análisis multivariante partiendo de los registros de propiedades organolépticas, origen, tipo, variedad y puntuaciones de catadores internacionales publicados por el Instituto de Calidad del Café (IQC, siglas en inglés) en el presente año. Con lo anterior, se han construido 4 modelos predictivos (3 clasificación y 1 de regresión) como son: K-vecino más cercano (KNN), Bosque aleatorio (RF), Regresión logística (RLog) y Regresión linear múltiple (RLM). Se evidenció que tanto la procedencia del grano (país y continente de origen) y sus propiedades de cata, juegan un papel importante al momento de determinar la variedad del grano. Se encontró que mayores puntuaciones de café correspondían a las variedades Caturra y Bourbon, del tipo arábico, y que en su mayoría provenían de del continente americano. Los países protagonistas fueron México, Guatemala, Colombia y Brasil.

Palabras clave: Café, variedad de grano, tipo de grano, análisis multivariante, aprendizaje supervisado, modelos predictivos, KNN, Bosque aleatorio, regresión logística, regresión lineal múltiple.

ABSTRACT

Coffee as a consumer item is one of the most commercialized raw materials in the world whose distinctive characteristic is its aroma, which is easily recognizable to many people around the world and which is accompanied by others such as body, flavor, uniformity, and sweetness. There is relevant literature that maintains that these properties depend to a great level on the type of grain (arabica and robusta) and its varieties (caturra, bourbon, typica, java, etc.). Thus, the present project aims to analyze the score given by professional tasters to different coffee samples and how this is related to the organoleptic properties, type and variety of the bean. This analysis would also imply analyzing the role of the origin of said grain, both by country and by continent. To carry out this analysis, we have resorted to the use of supervised learning techniques and multivariate analysis based on the records of organoleptic properties, origin, type, variety and scores of international tasters published by the Coffee Quality Institute (CQI) in the present year. With the above, 4 predictive models have been built (3 of classification and 1 of regression) such as: K-nearest neighbor (KNN), Random forest (RF), Logistic regression (RLog) and Multiple linear regression (RLM). It was evidenced that both origin of the grain (country and continent of origin) and its tasting properties play an important role when determining the variety of the grain. It was found that higher coffee scores corresponded to the Caturra and Bourbon varieties, of the Arabic type, and that most of them came from the American continent. The leading countries were Mexico, Guatemala, Colombia, and Brazil.

Keywords: Coffee, bean variety, bean type, multivariate analysis, supervised learning, predictive models, KNN, Random forest, logistic regression, multiple linear regression.

Introducción

El café como rubro de consumo es una de las materias primas más comercializadas en todo el mundo (Samoggia y Riedel, 2018). Una característica distintiva y llamativa es su aroma, el cual es fácilmente reconocible por la mayoría de las personas en el mundo (Seninde, 2020). Dicha propiedad organoléptica sirve como estimulante al momento de tomar el desayuno, especialmente en países de América Latina y Europa.

En las últimas décadas, el café ha experimentado una transformación de lo cotidiano a un producto de especialidad, evolución que comúnmente se divide en las llamadas “tres oleadas de consumo de café” (Bookman, 2014; Manzo, 2014). La primera ola de consumo de café surgió en la década de 1960, que se caracterizó por ser un mercado masivo con un crecimiento exponencial del consumo y una amplia disponibilidad a nivel mundial. La segunda ola de consumo de café comenzó en la década de 1990 con la formación de cadenas de cafeterías, como Starbucks. En este punto, las cafeterías introducen cafés especiales para responder al nuevo interés de los consumidores por la calidad de este rubro. Así, para autores como Carvalho et al. (2015) el café se convierte en un producto de lujo más que en una mercancía más que se compra en el supermercado. Por su parte, la tercera ola tuvo su génesis con pequeñas empresas tostadoras, que promovieron regiones específicas y nuevas técnicas de elaboración; otorgando de manera indirecta, más importancia al origen del café. Como consecuencia, el café ahora se considera un alimento artesanal de alta calidad, a menudo comparado con el vino.

Por consiguiente, el acto de tomar un café implica mucho más que simplemente consumir una bebida; es una experiencia multisensorial. Se trata vivir una experiencia, formar un estilo de vida, disfrutar de un momento placentero, y, en algunos casos, hasta de mostrar un estatus social. Este cambio en el comportamiento del consumidor a lo largo de los años ha sido posible gracias a tres orientaciones que definen en este momento al café: placer, salud y sostenibilidad (Mitchell, 2018; Sänger, 2018).

El café en números podemos resumirlo de la siguiente forma: Entre 2015 y 2016 se consumieron alrededor de 151,3 millones de sacos de café de 60 kg en todo el mundo (Voora et al., 2019). Estados Unidos fue el mayor consumidor de café a escala de países con 25 millones de sacos. En segundo lugar, se encuentra Brasil con 20 millones y a su vez, es el mayor productor del mundo con 55 millones sacos en ese mismo tiempo.

La Unión Europea como institución supranacional esgrime un consumo de 42 millones de sacos de 60 kg, mientras que las tasas de crecimiento más fuertes del consumo de café se han encontrado en Asia y Oceanía (Nevins et al., 2019). Mientras que los

escandinavos tienen el mayor consumo de café per cápita, por ejemplo, Finlandia ostenta un consumo de 12,2 kg por habitante; cifra considerable si tomamos en cuenta que Italia, conocido por su fuerte cultura cafetera, tiene un consumo de café per cápita de 5,6 kg. En comparación, el Reino Unido y Japón, como países con una larga tradición de consumo de té, tienen un consumo per cápita más bajo de 3,6 kg (Voora et al., 2019).

A parte de ser un producto de consumo social, el café también es utilizado por muchas otras industrias como la alimentaria (Çelik y Gökmen, 2018), la cosmética (Palmieri et al., 2018), la farmacia o la medicina (O'Keefe et al., 2018). Una ingesta diaria regular de café ha demostrado ser beneficiosa para los humanos, ya que reduce el riesgo de desarrollar algunos trastornos específicos como cirrosis, mal de Parkinson o cáncer intestinal (Dranoff, 2018).

A tenor de lo anterior resulta evidente que el café es un bien preciado por muchos. No obstante, detrás de este subyace un mundo repleto de diferentes variedades, tipos, sabores, y formas de procesamiento. Es por ello elegir el "el mejor" café pasa a ser una decisión importante; especialmente si se tiene como objetivo comercializarlo u ofrecer servicios relacionados con el mismo (cafeterías, bares, latte art, etc.) y bajo este contexto surge el presente estudio.

Por ejemplo, se sabe que existen aproximadamente 500 variedades distintas de café, pero solo dos de ellas son los más demandadas: Arábica y Canephora (también conocida como Robusta). La variedad arábica es una planta que tiene su origen en los países de América Latina y el Caribe debido a su clima cálido favorable (entre 15 y 24 ºC). Las plantas de esta variedad suelen crecer entre tres y cinco metros de altura y pueden cultivarse en zonas en torno a los 800 y 2000 msnm. En líneas generales, la variedad arábica ha demostrado ser extremadamente resistente a las inclemencias del clima (Anthony et al., 2001; Wintgens, 2004). Por su parte, la variedad Robusta es una planta que suele crecer entre siete y 13 metros de altura y se cultivan en zonas entre 200 y 900 msnm a temperaturas que oscilan entre los 24 ◦C y los 30 ◦C. A diferencia de la arábica, estas no son tan resistentes a los efectos del clima (Voora et al., 2019; Wintgens, 2004).

Como es de esperarse, ambas especies difieren en sus propiedades físicas y en algunas características químicas de sus granos (Caporaso et al., 2018; Toledo et al., 2016). La literatura actual muestra que los granos de Arábica contienen más trigonelina y lípidos, mientras que los granos de Robusta se caracterizan por un mayor contenido de cafeína y ácidos clorogénicos.

Estas diferencias en sus composiciones se ven reflejadas en su aroma y sabor. Por ejemplo, mientras que el sabor del café Arábica se considera suave y agradable, el café Robusta presenta un sabor más amargo con un olor casi a lodo (Flament, 2002; Sunarharum et al., 2014). Según la Organización Internacional del Café 2020 (ICO, 2020), el café Arábica representa el 61% de todo el café que se produce actualmente a nivel mundial, mientras que el café Robusta solo alcanza el 39%. Además, el precio del café Arábica es mucho más alto que el precio del café Robusta.

Por consiguiente, desarrollar una metodología concreta y confiable para saber cual es "el mejor café" se convierte en uno de los retos que se presentan en este trabajo. Partiendo de que existen catadores y personas expertas en la selección e identificación de las propiedades del café y su calidad, podemos utilizar dichos registros y construir un modelo estadístico que permita predecir o identificar la variedad del café y conocer si su calidad es aceptable o no para ser comercializado (Korhoňová et al., 2019). Desde el punto de vista económico esto puede resultar interesante para aquellas empresas que estén iniciando en el sector y estén desarrollando métodos de decisión optimizados con las que puedan obtener grandes rendimientos; o incluso, podría tener usos asociados a la detección de fraude alimentario.

El desarrollo de una metodología basadas en modelos estadísticos abre un abanico de posibilidades al momento de elegir el mejor camino para la construcción de los modelos predictivos. Una forma de hacerlo es aplicando técnicas de segmentación (clustering) y que cuentan con algoritmos muy conocidos y fiables que veremos más adelante (Suslick et al., 2010).

Así, encontramos trabajos como el Condliffe et al. (2008) que aplicaron técnicas de segmentación en ciertas variedades de café en Kenia. La segmentación o *clustering* de datos suele ser útil en este tipo de problemas ya que podemos conocer en cuantos grupos se divide la muestra de estudio en función de sus propiedades y puntuaciones. En 2019 Maciejewski et al. aplicaron un estudio análogo al del Condliffe et al. (2008) pero en el mercado polaco. En propósito de este ultimo era estudiar no solamente las variedades de café ofertadas en Polonia, sino además predecir la preferencia de los consumidores basados en propiedades como el aroma, sabor y tipo de recolección (arábiga o robusta)

Si se parten de registros cuantitativos y con muestras representativas sobre la calidad de café, se podrían construir modelos de regresión adaptados a la naturaleza de los registros analizados (de Morais et al., 2019). La regresión podría ser lineal o polinómica,

y dependerá en gran medida de la correlación vista entre las variables independientes y la dependiente.

Dentro del modelo de los análisis multivariantes se pueden considerar aquellos algoritmos de clasificación que se enmarcan dentro del aprendizaje supervisado. Uno de ellos es la técnica de la regresión logística (Irmeilyana et al., 2021; Park et al., 2019). Este método permitiría calcular la probabilidad de que una variedad de café estudiada pertenezca a arábica o robusta simplemente utilizando puntuaciones de aroma y sabor; entre otras propiedades.

Siguiendo con el estudio de probabilidades, el método de Naive-Bayes ingenuo se muestra como una alternativa aceptable si se quiere tomar en cuenta la influencia de la probabilidad de ocurrencia de las variables independientes, y como estas afectan la variable dependiente (Barbosa et al., 2014).

Llegados a este punto, se entiende que existen varias alternativas dentro de las ciencias actuariales que son aplicables en este trabajo. Así pues, para determinar la variedad de un tipo de café y/o predecir la calidad de estas basadas en sus propiedades, surge la importancia de proponer e implementar algoritmos robustos de cálculo que sean ampliamente conocidos por su flexibilidad, bajo coste computacional y alta precisión.

En el presente trabajo analizaremos dos datasets que corresponden a dos variedades de café (arábica y robusta). Cada una de las tablas están compuestas por más de 1300 observaciones diferentes de café de las dos principales variedades. Algunas de los atributos que se tomarán en cuenta son: Especie, País de Origen, altitud, aroma, sabor, acidez, amargo, puntos de copa y color. En el apartado de “metodología” se explicará con más detalle la característica de ambas tablas.

Las bases de datos, tanto de la variedad arábica como robusta fueron tomadas del Instituto de la Calidad del Café (Coffee Quality Institute) disponibles a través del siguiente enlace: https://github.com/jldbc/coffee-quality-database/tree/master/data. Para el procesamiento completo de los datos y la implementación de la solución algorítmica utilizaremos el lenguaje de programación R que será ejecutado desde su entorno de desarrollo más conocido “RStudio”.

Objetivos

A tenor de lo mencionado en el apartado anterior, en este estudio vamos a analizar dos bases de datos sobre la calificación que han otorgado catadores profesionales del café a distintas muestras de café a fin de averiguar cuáles son las características y los factores que hace que un café sea el mejor entre una variedad y otra.

Por consiguiente, podríamos formalizar el objetivo general de este estudio de la siguiente forma:

- Determinar la calidad del café a través de uso de técnicas de aprendizaje supervisado y análisis multivariante a partir de registros de propiedades organolépticas, origen, tipo, variedad y puntuaciones de catadores internacionales

Para dar cumplimiento a este objetivo se proponen los siguientes objetivos específicos:

- Determinar la variedad predilecta de los catadores de café con base a su puntuación
- Segmentar las bases de datos en función del aroma, sabor, puntos de taza, y demás atributos de catado
- Realizar un análisis exploratorio de las bases de datos analizadas para conocer factores como frecuencia de origen y tipo, altitud, país con mayor producción, origen de la mejor semilla de café, entre otras.
- Implementar al menos 3 modelos predictivos basados en aprendizaje supervisado y multivariante para determinar la calidad del café con base a variables como el aroma, variedad, tipo, origen, puntos taza, sabor y color.

Al finalizar este estudio seremos capaces de responder a preguntas como las que se formulan a continuación:

- ¿Qué país tiene el café con mejor puntuación?
- ¿Qué país o continente produce el café con el mejor sabor?
- ¿En qué continente se obtiene la semilla de café de mejor calidad?
- ¿Qué variedad de café es la favorita de los catadores? ¿Cuáles variables responden a eso?

Metodología

Una vez se cuenta con un estado del arte apropiado sobre el tema en cuestión y con los objetivos en mente, se procede a plantear el "cómo" se va a llevar a cabo esta investigación. Para ello, se describe el tipo de estudio, las bases de datos utilizadas, las técnicas de procesamiento elegidas y las herramientas informáticas estadísticas necesarias para llevar a cabo el análisis multivariante sobre la calidad del café.

Tipo de estudio

El presente estudio se enmarca como tipo cuantitativo, descriptivo y empírico. Será de tipo cuantitativo ya que nos basaremos en puntuaciones numéricas ponderadas para determinar la relación entre la calidad del café y factores como origen, variedad y tipo de grano respectivamente. Será descriptivo también porque analizaremos el contexto de una variable concreta, en este caso será lo referido a la calidad del café. Un análisis descriptivo permite, entre otras cosas, detallar como se manifiesta un fenómeno o evento y buscan especificar las propiedades y/o características del mismo mediante técnicas de análisis científico (Yannis y Nikolaos, 2018). Esto ultimo va estrechamente relacionado con los objetivos propuestos, ya que se pretende encontrar la posible relación empírica entre las propiedades del café (Puntos, aroma, cuerpo, balance, dulzor, etc.) con atributos como su origen por país o continente, el tipo (robusta o arabica) y variedad (Caturra, Bourbon, Typica, etc.)

Bases de datos

El origen de los datos con los que se trabajó en esta investigación corresponde a los registros de variedades de café del Instituto de la Calidad del Café (CQI, siglas en inglés), y que pueden ser encontradas de manera pública a través del siguiente enlace[1]:

- https://database.coffeeinstitute.org/

El CQI es una institución fundada en 1996 y que al presente goza de reconocimiento internacional. Su misión es la de trabajar constantemente en la mejora de la calidad del café y de la vida de las personas que lo producen. Para el CQI la calidad es una de las variables más importantes que influyen en el valor de un café. Sin embargo, muchos productores no tienen acceso a las herramientas y el apoyo que necesitan para comprender la calidad de su café, mejorar esa calidad, acceder a mercados que recompensan esa calidad y, en última instancia, les permiten tomar decisiones comerciales más informadas. Ahí es donde radica el protagonismo de esta institución.

[1] Estos registros también están disponibles en el siguiente repositorio GitHub. https://github.com/jldbc/coffee-quality-database/tree/master/data

Debido a la trayectoria de esta institución, su compromiso con la calidad del café, y su reconocimiento internacional debido a su rigurosidad en los criterios de evaluación, es que hemos determinado que los registros obtenidos de esta serían apropiados para el presente trabajo. De esta forma también nos aseguramos que se trabajan con datos reales, lo cual se convierte en un valor agregado al momento de presentar los resultados y las conclusiones.

Características y atributos de los datos

Los datos contienen registros de 1312 granos de café arábica y 28 robusta de revisores capacitados del CQI. Las características incluyen: medidas de calidad, atributos de grano y datos referidos al método de cultivo y cosecha.

Para las medidas de calidad se han considerado las siguientes variables:

- Aroma
- Sabor
- Retrosabor
- Acidez
- Cuerpo
- Equilibrio
- Uniformidad
- Taza Limpia
- Dulzor

Con respecto al grano se tiene los siguientes atributos:

- Método de procesamiento
- Color
- Tipo (Robusta o Arábica)
- Variedad (Typica, Bourbon, Caturra, etc.)

Para las técnicas de cultivo y cosecha se tiene:

- Dueño
- País de Origen
- Nombre de la granja/finca
- Numero de lote
- Ingenio
- Compañía
- Altura

- Región

Para el presente estudio se utilizaron los datos en bruto (Raw Data) y que posteriormente fueron depurados a través de RStudio (Ver apartado de resultados)

Variables

Por iniciativa del investigador se ha decidido trabajar con los nombres originales de las variables, por lo que se mantendrán las etiquetas de los atributos del dataset. Así, tenemos:

Tabla 1. Correspondencia de nombre de variable.

Variable	Nombre Original
Aroma	Aroma
Sabor	Flavor
Retrosabor	Aftertaste
Acidez	Acidity
Cuerpo	Body
Uniformidad	Balance
Taza Limpia	Clean Cup
Dulzor	Sweetness
Puntos de taza	Cupper points
País	Country
Continente	Continent
Variedad	Variety
Tipo	Type
Puntos	Points

Técnicas de procesamiento

A fin de mantener la mayor rigurosidad posible a la altura de un trabajo de investigación se han seleccionado cuatro técnicas de análisis multivariante cuyos resultados individuales podrán ser comparados y determinar su "*performance*". Con ello se busca identificar cual método es más adecuado para este fenómeno en concreto.

Los métodos elegidos corresponden a herramientas del aprendizaje supervisado enmarcados dentro de la disciplina del aprendizaje automático o "*Machine Learning*" y son: Regresión lineal múltiple (RLM), Vecino más cercano o "*K-nearest neighbourg*" (KNN), Random Forest (RF) y Regresión logística (RLog). Métodos ampliamente utilizados en el campo de la ciencia de datos, big data y ciencias actuariales en general.

Importante señalar en este punto que los métodos RF, KNN y RLog son de clasificación, mientras que el RLM es de modelado predictivo. Pese a estas diferencias, podemos utilizar los primeros como herramientas de regresión gracias a librerías especializadas

como “caret” de R. A continuación, describiremos brevemente los métodos seleccionados.

Regresión Lineal Múltiple

Este método nos permitirá estimar un modelo matemático múltiple entre los puntos de café y sus propiedades organolépticas. Para ello, la variable dependiente *“y”* serán los puntos (Points), y las variables independientes “*β*” serán Aroma, cuerpo, sabor, retrosabor, acidez, cuerpo, balance, uniformidad, taza limpia, dulzor y puntos taza. Es decir, todas aquellas variables derivadas de la evaluación del catado.

La ecuación propuesta queda de la siguiente forma:

$$Pt = \beta_0 + A\beta_1 + F\beta_2 + Af\beta_3 + Ac\beta_4 + Bd\beta_5 + Bl\beta_6 + U\beta_7 + Cc\beta_8 + Sw\beta_9 + Cp\beta_{10}$$

Donde:

Pt: Points

β_0: Intercepto

β_n: coeficiente de regresión en parámetro n

A: Aroma

F: Flavor

Af: Aftertaste

Ac: Acidity

Bd: Body

Bl: Balance

U: Uniformity

Cc: Clean cup

Sw: Sweetness

Cp: Cupper points

Vecino más cercano KNN

El algoritmo KNN es ampliamente utilizado como una herramienta de aprendizaje supervisado de clasificación. Este algoritmo se caracteriza por almacenar todos los casos disponibles y predice el objetivo numérico en función de una medida de similitud, como pueden ser las funciones de distancia (euclidiana, de Coseno, Jaccard, etc.). KNN es ampliamente utilizado en la estimación estadística y el reconocimiento de patrones ya a principios de la década de 1970 como una técnica no paramétrica (Zhang et al., 2018).

Para que el algoritmo de KNN pueda arrojar resultados óptimos será necesario estimar un valor de clúster (K) tal que tenga una distancia lo más cercana posible a las características de la variable en estudio (Sayad, 2021).

La mejor forma de elegir el valor óptimo para K es inspeccionando primero los datos. En general, un valor de K grande es más preciso ya que reduce el ruido general. Sin embargo, para evitar solapamiento entre fronteras de vecinos se puede recurrir a la validación cruzada ("*cross-validation*", cv) como una forma de determinar retrospectivamente un buen valor de K. La K óptima para la mayoría de los conjuntos de datos es 10 o más (Deng et al., 2016). De manera gráfica también se puede identificar el K óptimo mediante la técnica del codo ("Elbow method").

Random Forest

El algoritmo de Random Forest (RF) es una extensión del método de árboles de decisión ("*Decision tree regression"*) el cual es también ampliamente utilizado dentro de los métodos supervisados de ML; con la diferencia de que en Random Forest utiliza las cualidades y características de múltiples "arboles" para tomar decisiones (Schonlau y Zou, 2020).

Es por ello que RF se refiere a un "bosque" de árboles aleatorios (de allí su nombre). El término "aleatorio" se debe al hecho de que este algoritmo es un bosque de árboles de decisión creados aleatoriamente (Gurucharan, 2020).

Debido a que los "arboles de decisión" son muy propensos a caer en el sobreajuste ("overfitting"), este problema se puede resolver implementando la regresión aleatoria de bosque en lugar de la regresión del árbol de decisión. Además, el algoritmo Random Forest también es muy rápido y robusto que otros modelos de regresión. Estos aspectos hacen del algoritmo una alternativa a tomar en cuenta y es la razón por la que se ha decidido incluir dentro de este estudio.

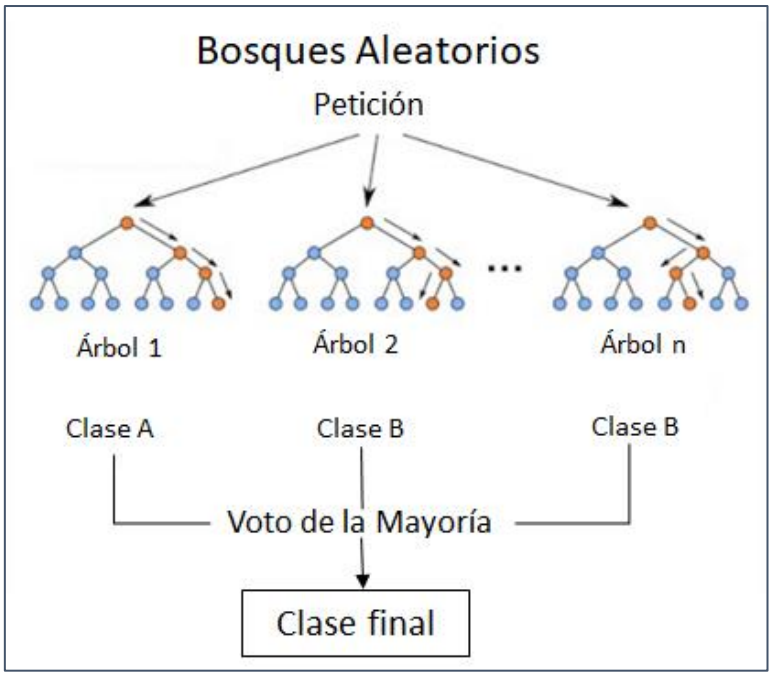


Figura 1. Representación esquemática del funcionamiento del algoritmo Random Forest. Fuente: (Huacasi, 2020)

En este estudio se ha construido un modelo de regresión RF que analiza la correlación entre la variedad del grano y todas las demás variables expuestas en la tabla 1 de esta memoria. El propósito es que este algoritmo prediga las variedades del grano de café en función de su origen y sus propiedades organolépticas evaluadas por los catadores.

Regresión logística

La regresión logística (RLog) es uno de los algoritmos de ML más populares y utilizados dentro de las ciencias actuariales, siendo también una técnica de aprendizaje supervisado. Este algoritmo es útil si se quiere predecir la variable dependiente categórica mediante un conjunto dado de variables independientes (Dreiseitl y Ohno-Machado, 2002). Por consiguiente, el resultado de dicha predicción será un valor categórico o discreto. Por ejemplo, pueden ser Sí o No, 0 o 1, verdadero o falso, etc., y en lugar de arrojar el valor exacto como 0 y 1, arroja valores probabilísticos que se encuentran entre 0 y 1 (Subasi y Erçelebi, 2005).

El algoritmo RLog es tomado en cuenta en este estudio debido a que tiene la capacidad de proporcionar probabilidades y clasificar nuevos datos utilizando conjuntos de datos continuos y discretos al mismo tiempo. Es decir, puede usarse para clasificar las observaciones usando diferentes tipos de datos determinando fácilmente las variables más efectivas usadas para la clasificación.

Cabe señalar que, si bien la RLog utiliza el concepto de modelado predictivo como regresión (de ahí su nombre), su utilidad es para clasificar muestras; Por lo tanto, se incluye dentro de los algoritmos de clasificación, como el KNN y RF (Paliwal y Kumar, 2009).

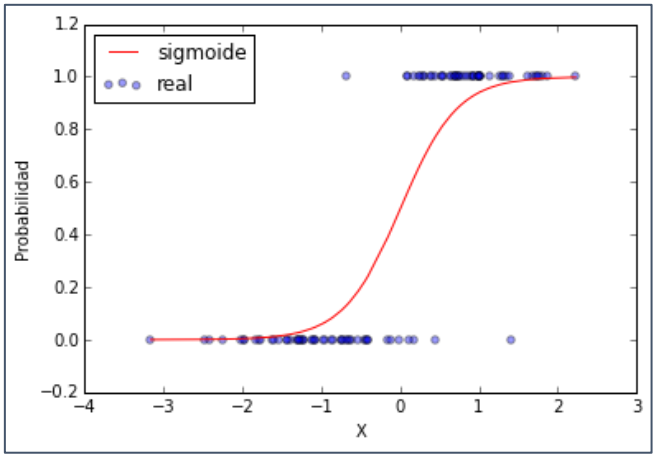


Figura 2. Modelo sigmoidal de la curva de la regresión logística. Fuente:(Fernández, 2018)

Para efectos del presente trabajo se ha considerado como variable binaria (*Dummy*) los valores “Caturra” y “Bourbon” de la variable “Variety”, esto debido a que son las variedades más frecuentes dentro del dataset como veremos más adelante. Con el método de RLog queremos determinar la probabilidad de que un grano de tipo arábico sea de la variedad “Caturra” o “Bourbon” en función de sus valores de catado y su origen.

La función logística sigmoide que define las probabilidades de variedad de grano en el presente estudio viene dada por el siguiente modelo:

$$Variety = \frac{1}{1 + e^{(A\beta_1 + F\beta_2 + Af\beta_3 + Ac\beta_4 + Bd\beta_5 + Bl\beta_6 + U\beta_7 + Cc\beta_8 + Sw\beta_9 + Cp\beta_{10} + Ctry\beta_{11} + Ctnt\beta_{12})}}$$

Donde:

Ctry: Country

Ctnt: Continent

Herramientas estadísticas

El almacenamiento de los datos se hará a través de documentos en CSV (valores separados por coma) y XLSX. Para su acceso y visualización rápida utilizaremos Microsoft Excel 2019.

El motor de procesamiento estadístico, carga de datos, construcción de modelos de aprendizaje y análisis exploratorio de los datos se utilizará RStudio 1.4.1717

Las librerías necesarias que deberán estar instaladas y cargadas en el script final son:

- caret,
- tydverse
- forecast
- janitor
- skimr
- tinytext
- rvest
- dplyr
- xlsx
- pROC
- MLmetrics

Para la presentación de resultados y elaboración de informe final se ha utilizado la herramienta RMarkdown de RStudio.

Fases del proyecto

El presente estudio se realizará en las siguientes fases:

- *Fase 1 - ETL (Extract, Transform and Load)*

En esta fase se realiza el análisis exploratorio de las propiedades de los datos (EDA) y así estimar el número de registros, número de variables, tipo de variables, detectar posible presencia de outliers, visualizar los datos, realizar operaciones de estadística descriptiva, inferencial y pruebas de normalidad según sea el caso.

EL propósito de esta fase es tener un acercamiento a los datos con los que se van a trabajar. De esta manera se podría dilucidar qué tan relacionados están los atributos entre si y cuales son los componentes principales incipientes dentro del dataset.

- *Fase 2 - Elección de métodos de análisis y construcción de los modelos*

En esta fase se aplican los modelos de regresión y clasificación seleccionados. A saber: Regresión lineal múltiple, KNN, RF y regresión logística binomial. En esta fase también se contempla la preparación de los data sets correspondientes a entrenamiento “Train” y pruebas “Test” adecuados para cada modelo y según las aplicaciones que cada algoritmo necesita. También contempla la configuración de los parámetros de control de cada modelo, especialmente en lo referido a los algoritmos de clasificación (KNN, RF y RLog); dando lugar a la consideración de las métricas de evaluación *(“Accuracy o Cross-validation”*), numero de iteraciones, repeticiones y probabilidad de clases.

- *Fase 3 - Presentación y análisis de resultados obtenidos*:

En esta fase procedemos a comparar los resultados obtenidos de cada método elegido en la fase de modelado. Evaluando el rendimiento y precisión de cada modelo se estimaría la elección del mejor método o algoritmo para este estudio en particular.

Esta fase contempla la presentación de tablas de resultados, filtrados, diagramas de barras, histogramas de frecuencia, curvas ROC y detección de verdaderos positivos a través de matrices de confusión y generación del *markdown* final.

- *Fase 4 - Discusión y conclusiones*:

En esta fase se concluye el trabajo luego de la presentación de resultados y se elige el método multivariante con mayor capacidad de predicción para este caso en concreto.

Resultados

En el presente apartado se exponen los resultados finales obtenidos. Tal como se explicó en el apartado de la metodología, el problema fue abordado siguiendo 4 fases específicas y en dicho orden de ejecución se presentan los resultados en la memoria.

Análisis de los Datos

El dataset completo está conformado por 1328 observaciones y 48 atributos. A fin de realizar un estudio más extenso sobre la localización del mejor café, se ha añadido una columna con los nombres de los cinco Continentes en función del país de origen, y de esta forma utilizar estos datos para una posterior clasificación. Por consiguiente, en total tendremos 1328 observaciones y 49 atributos (Ver tabla 2).

Siendo un conjunto de datos grande y con permeable heterogeniedad en los atributos, se ha seleccionado un subconjunto de variables que a priori pudieran ser más relevantes para dar respuesta a los objetivos del estudio. El subconjunto seleccionado final cuenta con un total 15 variables. Estas variables corresponden a las puntuaciones otorgadas por los catadores de café y el lugar de procedencia de las semillas, incluyendo el país como el continente (Ver apartado Metodología).

Los atributos asociados a los catadores son los siguientes:

1. Aroma
2. Flavor
3. Aftertaste
4. Acidity
5. Body
6. Balance
7. Uniformity
8. Clean cup
9. Sweetness
10. Cupper points

Estas 10 variables son cuantitativas con calificaciones que oscilan entre 7 y 10. Hay que destacar que las variables Uniformity, Clean cup y Sweetness tienen de media de puntuación un 10 para casi todos los países.

Como variables cualitativas tendríamos 4:

1. Country,
2. Continent,

3. Typo
4. Variety.

De manera preliminar podemos señalar que los países que lideran el ranking son: **Etiopía, Guatemala y Brasil**.

Tabla 2. Primeros 10 registros del conjunto de datos estudiado. Elaboración propia.

Points	Type	Variety	Country	Cont.	Aroma	Flavor	AT	Acidity	Body	Balance	Un.	Clean cup	Sweetness	Cupper points
90.58	Arabica	NA	Ethiopia	Africa	8.67	8.83	8.67	8.75	8.50	8.42	10.00	10	10.00	8.75
89.92	Arabica	Other	Ethiopia	Africa	8.75	8.67	8.50	8.58	8.42	8.42	10.00	10	10.00	8.58
89.75	Arabica	Bourbon	Guatemala	Americas	8.42	8.50	8.42	8.42	8.33	8.42	10.00	10	10.00	9.25
89.00	Arabica	NA	Ethiopia	Africa	8.17	8.58	8.42	8.42	8.50	8.25	10.00	10	10.00	8.67
88.83	Arabica	Other	Ethiopia	Africa	8.25	8.50	8.25	8.50	8.42	8.33	10.00	10	10.00	8.58
88.83	Arabica	NA	Brazil	Americas	8.58	8.42	8.42	8.50	8.25	8.33	10.00	10	10.00	8.33
88.75	Arabica	Other	Peru	Americas	8.42	8.50	8.33	8.50	8.25	8.25	10.00	10	10.00	8.50
88.67	Arabica	NA	Ethiopia	Africa	8.25	8.33	8.50	8.42	8.33	8.50	10.00	10	9.33	9.00
88.42	Arabica	NA	Ethiopia	Africa	8.67	8.67	8.58	8.42	8.33	8.42	9.33	10	9.33	8.67
88.25	Arabica	Other	Ethiopia	Africa	8.08	8.58	8.50	8.50	7.67	8.42	10.00	10	10.00	8.50

Cont. Continente,
Un. Uniformity
AT. Aftertaste

Valores Ausentes (NAN)

Como parte del primer acercamiento de los datos nos interesaría conocer el número de valores perdidos por cada atributo, ya que la presencia de estos puede afectar la precisión de los modelos predictivos (ver tabla 3).

Tabla 3. Numero de valores ausentes o perdidos dentro del dataset. Elaboración propia

Variable	NAN's
Aroma	140
Flavor	112
Aftertaste	112
Acidity	140
Body	140
Balance	112
Uniformity	140
Clean Cup	112
Sweetness	140
Cupper points	112
Country	0
Continent	0
Variety	306
Type	112
Points	112

Dentro de las variables *Country* y *Continent* no encontramos ningún valor ausente. Sin embargo, se han detectado 112 valores ausentes repartidos en las siguientes variables: *Points, Type, Flavor, Aftertaste, Balance, Clean_cup y Cupper_points*. Y en el resto de atributos se han identificado hasta 140 valores ausentes por cada variable. La variable *Variety* es la que mayor número tiene, con 306.

Análisis exploratorio y visualización

Dentro del esquema ETL, el análisis exploratorio se define como un proceso iterativo mediante el cual analizamos la información potencial que puede ser extraída de los datos (Runtuwene et al., 2018). Esto permite formular preguntas sobre el conjunto de datos, además, de ayudar a responderlas a través de gráficos, transformaciones y técnicas de modelización.

Variación de los datos

Estudiar el cómo varían los datos permite comprender en profundidad las principales características de una variable. Se entiende como *variación* a la tendencia de los valores de una variable a cambiar entre *medida* y *medida* (Graus, 2018). Cada variable puede tener su propio patrón de variación que podría revelar información relevante. Tomando en cuenta lo anterior mencionado, una forma de evidenciar estas variaciones es a través de diagramas de barras e histogramas. Las primeras fueron aplicadas a las variables cualitativas (Type, Variety, Continent) y los histogramas a todas las variables cuantitativas.

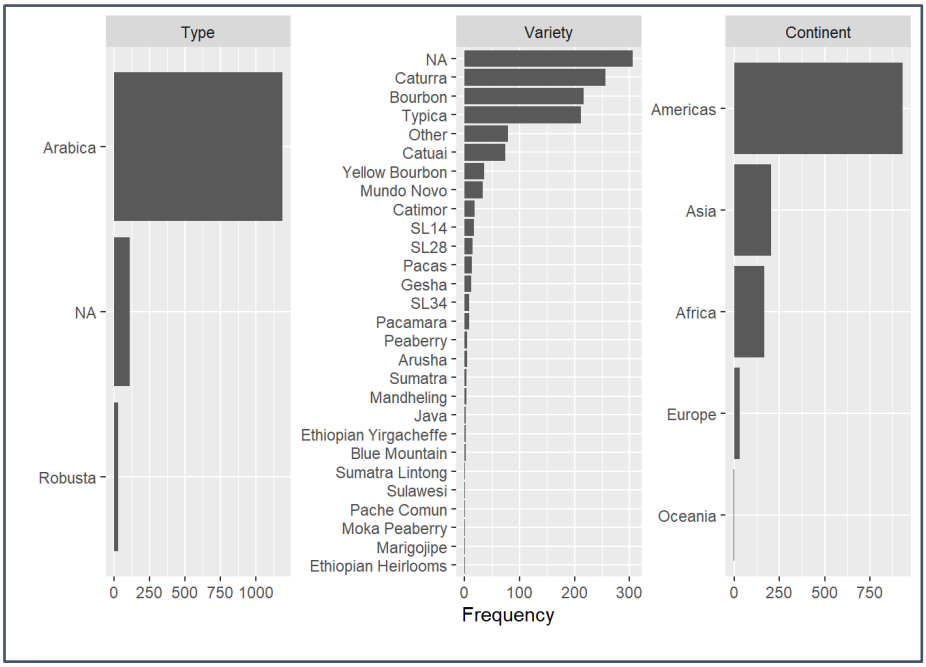


Figura 3. Frecuencia de las variables cualitativas mediante diagramas de barras. Elaboración propia

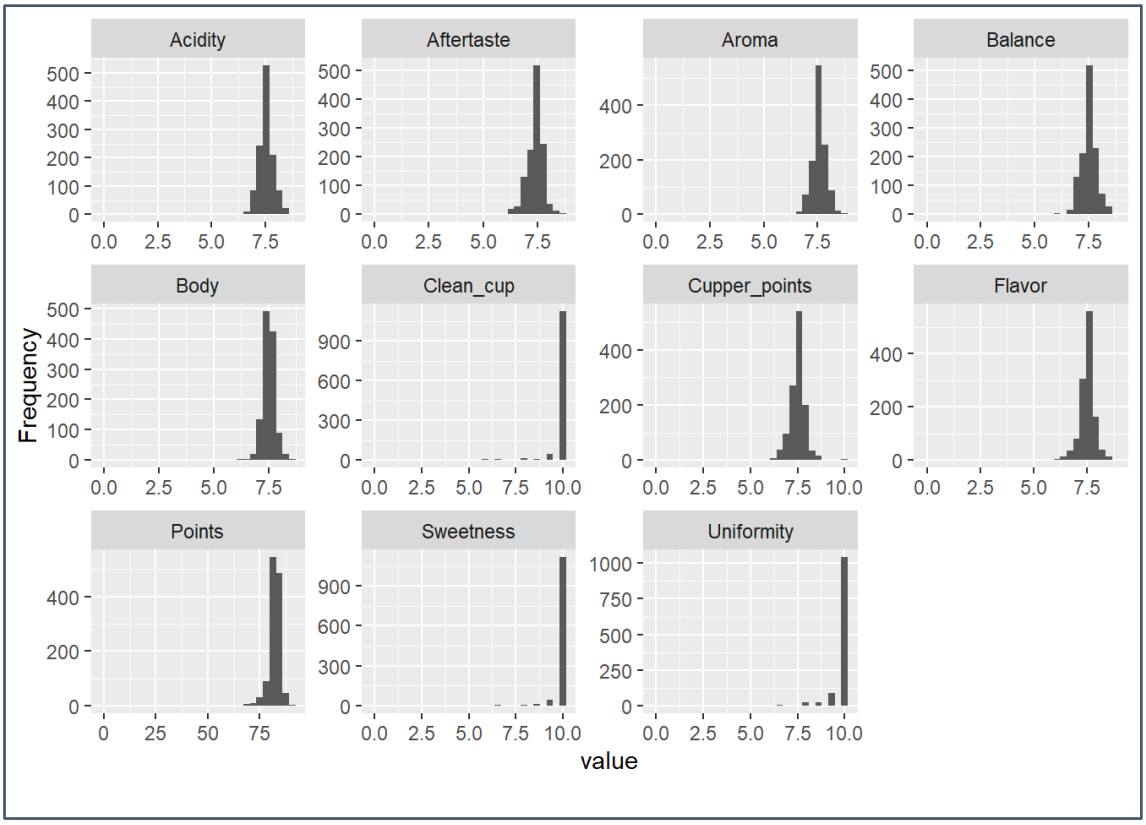


Figura 4. Frecuencia de los datos cuantitativos mediante histogramas. Elaboración propia

En las figuras 3 y 4 evidenciamos la densidad de las variables en el dataset. De la primera se destaca la enorme cantidad de registros de grano tipo arábico con respecto al robusto. En función del origen del grano, la mayor parte es procedente del continente americano, y en menor proporción de Europa y Oceanía. Con respecto a la variedad del grano, se evidencia que Caturra, Bourbon, Typica y "Other" son las más frecuentes.

Por su parte, en la figura 4 no solo observamos la frecuencia continua, sino además la distribución de las variables. A priori, la mayoría de estas presenta un comportamiento normal. Vale destacar que todas estas variables muestran muy altas puntuaciones en la mayoría de los registros analizados (entre 7 y 10). En la sección de anexos se recogen distribuciones en escala logarítmica para que puedan ser estudiadas más de cerca.

Variables cualitativas en profundidad

En el subconjunto estudiado se tienen cuatro variables cualitativas: Country, Continent, Type y Variety.

Por una parte, la variable de Country, se observa que los datos recogen información de granos de cafés que proceden de muchos países del mundo. El país que más se repite en los datos es México, seguido de Colombia y de Guatemala como muestra la tabla 4. Estos son los países que han tenido más muestras calificadas por los catadores.

Tabla 4. País de procedencia de los granos. Elaboración propia

Country	Cantidad
Mexico	230
Guatemala	173
Colombia	141
Brazil	115
Taiwan	69
Honduras	50
Costa Rica	45
Uganda	25
Kenya	23
El Salvador	17
Indonesia	17
China	16
Ethiopia	15
Nicaragua	15
Thailand	14

Sin embargo, la variable de Continent, similar a la información obtenida con la variable Country, la mayoría de las muestras calificadas proceden de América, África y Asia. En Oceanía apenas existe producción de café de acuerdo con este *dataset* (ver tabla 5).

Continent	Cantidad
Americas	928
Asia	203
Africa	165
Europe	30
Oceania	2

La figura 5 nos permite comprobar la cantidad de grano por variedad dentro del tipo arábica. Vale destacar los dos Tipos de café en grano: Arábica y Robusta. El café arábico procede de la especie «*Caffea arabica*» y el robusta del «*Caffea Canephora*». La primera de ellas requiere un clima subtropical fresco y es muy vulnerable al frío y a los insectos lo que hace que su cultivo y cuidado sea mucho más costoso.

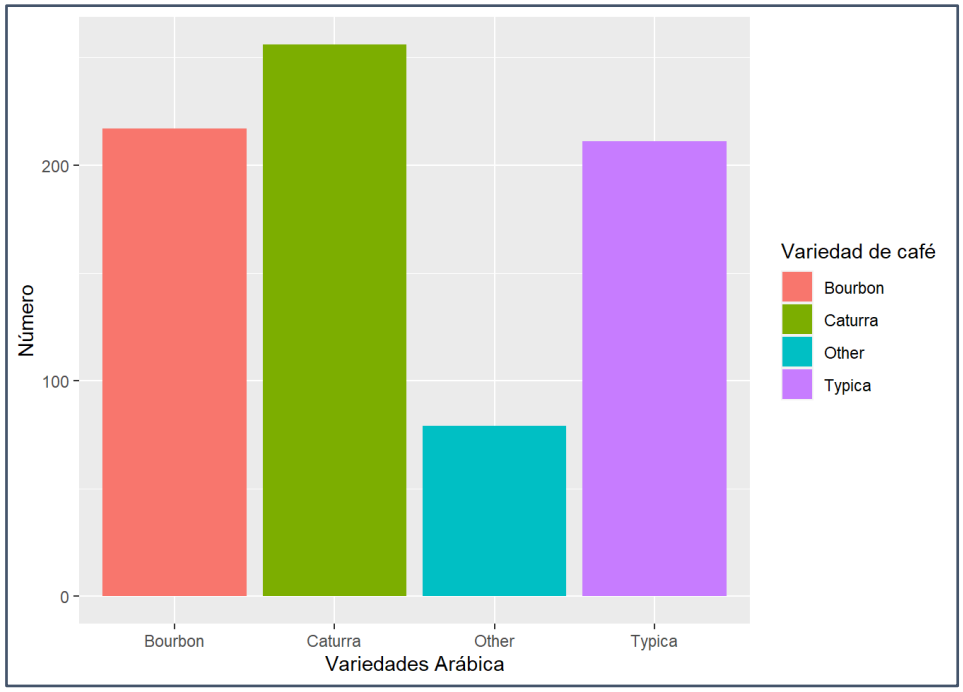


Figura 5. Variedad de grano en función de su cantidad. todas dentro del tipo arábica. Elaboración propia

Se ha decidido trabajar solo con el tipo “arábica” ya que tenemos 1188 registros con este valor y tan solo 28 robusta. De esta forma, solo trabajaremos con las variedades que conforman el tipo de grano arábica, más concretamente aquellas que sean mayoría dentro del dataset: Bourbon, Caturra y Typica. De la figura 5 también se desprende que la variedad “Caturra” es la más frecuente. Con 256 registros, seguida de Bourbon con 217 y Typica con 211.

Covariación entre atributos

Recordando que tenemos 4 variables categóricas: Type, Country, Continent, y Variety. Veamos cómo se relacionan los conjuntos de variables categóricas y continuas en este problema concreto.

Mediante histogramas

De los 4 gráficos siguientes (figuras 6 al 9) se desprende que existe una gran similitud de comportamiento entre ellos. Es importante notar que en la representación de la distribución por Continente (figura 8) la mayoría de los datos se registran en América, África y Asia; esto se debe a que los registros referentes a Oceanía y Europa son muy pequeños, y por tanto poco representativos. Por otro lado, en relación a la distribución de Puntos según el país (figura 9) se destaca que la prevalencia de Etiopía con mayor desplazamiento hacia la derecha, indicando que los granos que proceden de este país son los que presentan la mayor puntuación total obtenida. También podemos destacar

la densidad de Colombia y Brasil son más altas que el resto, esto es debido, a que estos países son los que tienen más registros; y por ende, más valoraciones

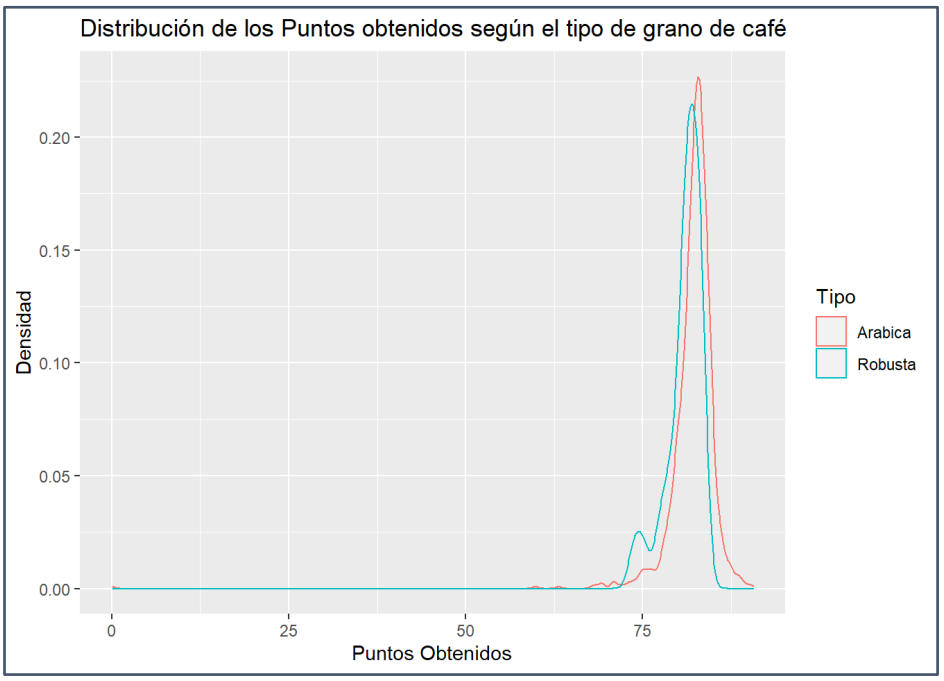


Figura 6. Distribución de puntos de calidad de café según el tipo de grano

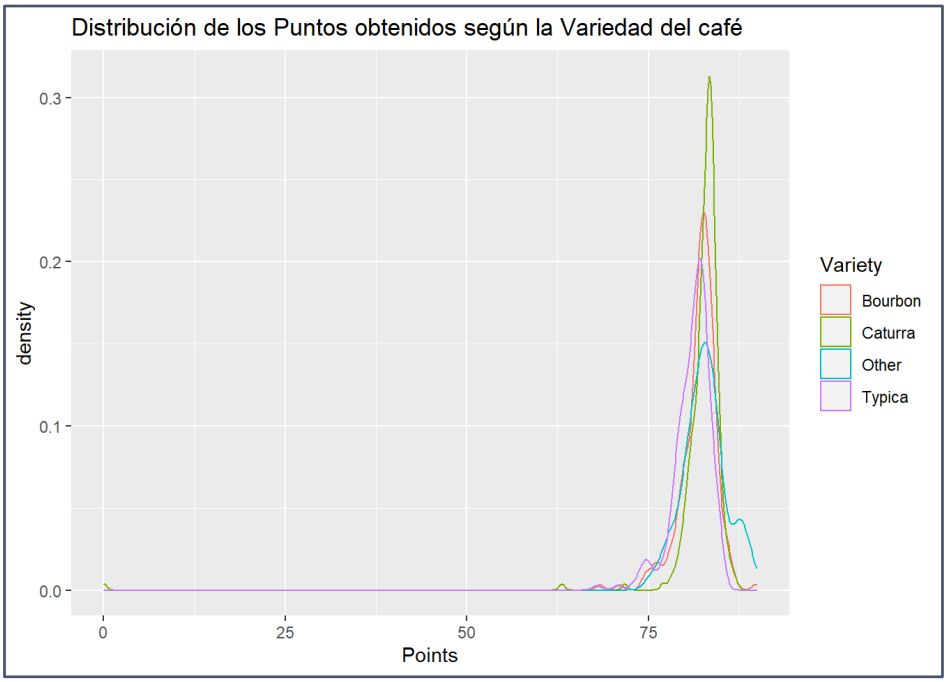


Figura 7. Distribución de puntos según la variedad de café

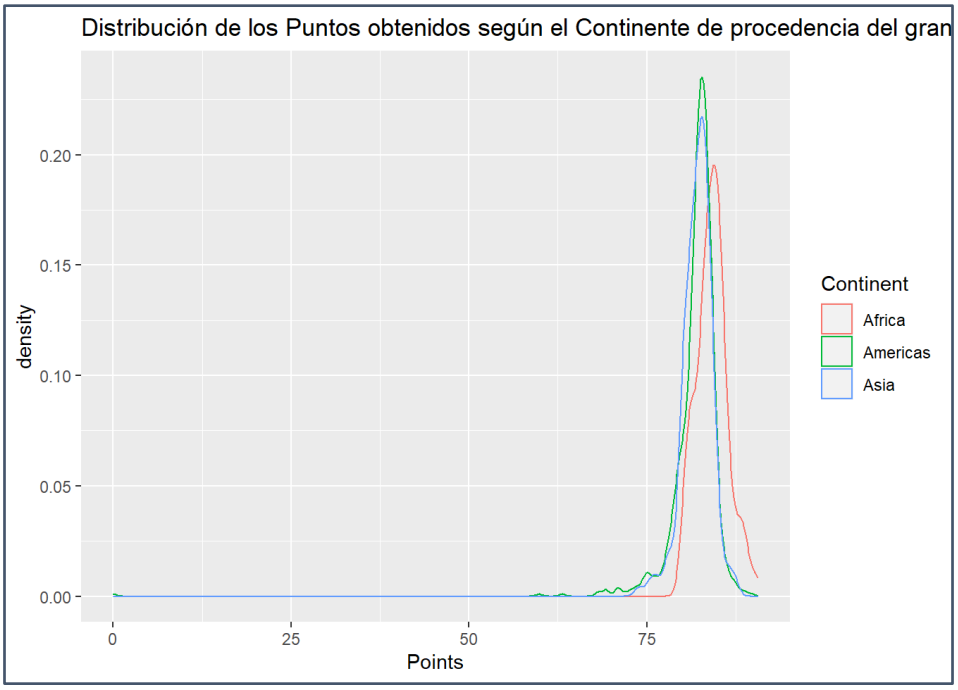


Figura 8. Distribución de los puntos según Continente de procedencia

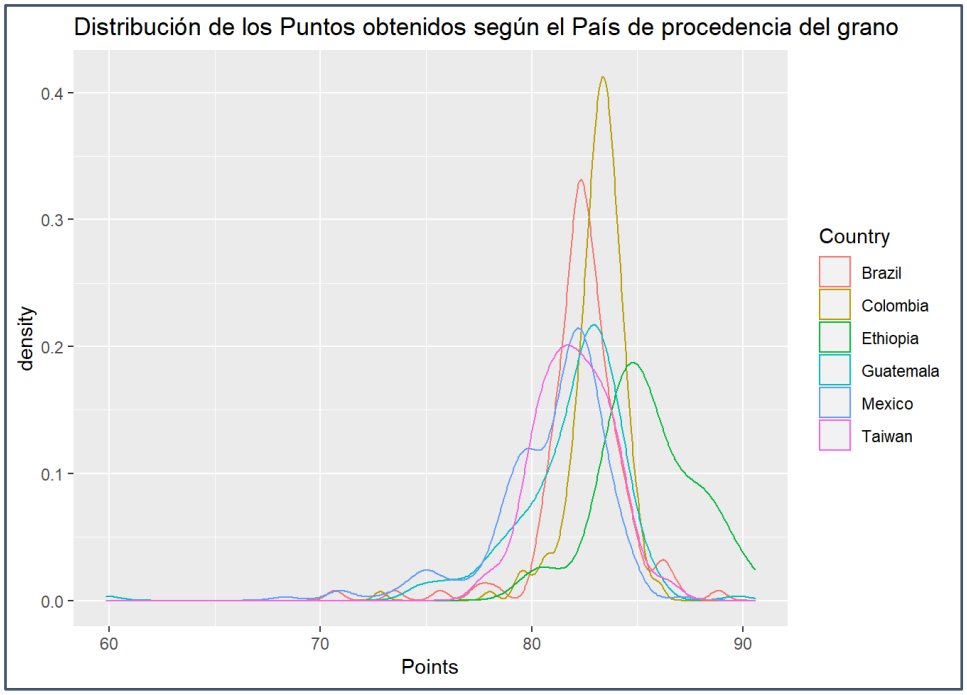


Figura 9. Distribución de puntos según País de origen

Mediante regresión lineal simple

Podemos analizar la covariación entre variables desde un método más numérico y no tan observacional como el caso anterior. Para ello podemos valernos de la regresión simple a fin de encontrar alguna proporción de correspondencia lineal.

Tabla 5. Resultado de regresión lineal simple para continentes

Termino	Estimado	Std Error	Estadístico	p-valor
(Intercepto)	83.992338	0.4176118	201.125372	0.0000000
Américas	-2.172536	0.4371049	-4.970286	0.0000008
Asia	-1.846823	0.5226286	-3.533720	0.0004282

Analizando la correlación entre Continente de procedencia y puntos de catado (tabla 5), se obtuvo un p-valor < 0.05 y un R cuadrado ajustado de 0.03471. Rechazamos la hipótesis nula y vemos que hay una significancia estadística probable entre continente y puntos de catado.

Tabla 6. Resultado de regresión lineal simple para tipos de grano

Termino	Estimado	Std Error	Estadístico	p-valor
(Intercept)	82.129411	0.1039711	789.925582	0.0000000
Robusta	-1.260482	0.6851731	-1.839655	0.0660629

En contraste a lo mostrado en la tabla 5, en el caso del tipo (tabla 6) no parece haber una correlación directa con los puntos. Sin embargo, no podemos rechazar la hipótesis nula ya que del tipo *Robusta* tenemos muy pocos registros en comparación a *Arabica*. Por consiguiente, tendremos que seguir indagando con otras variables categóricas como la variedad y el país.

Tabla 7. Resultados de regresión lineal simple para variedad de grano

Termino	Estimado	Std Error	Estadístico	p-valor
(Intercept)	81.9384793	0.2643402	309.973642	0.0000000
Caturra	0.5005051	0.3593137	1.392948	0.1640435
Other	0.8129131	0.5116765	1.588725	0.1125390
Typica	-0.9177210	0.3764816	-2.437625	0.0150127

La variedad de grano (tabla 7) dentro del tipo Arábica tiene un p-valor <0.05 y un R cuadrado ajustado de 0.02106. Rechazamos la hipótesis nula de significancia estadística ya que la variable Variety influye de manera importante en la variable Points. Al igual que el caso de Continente, encontramos una significancia estadística muy relevante entre Country y Points (ver tabla 8)

Tabla 8. Resultados de regresión lineal simple para origen por país

Termino	Estimado	Std Error	Estadístico	p-valor
(Intercept)	82.5942609	0.2146627	384.762903	0.0000000
Colombia	0.7690015	0.2892458	2.658644	0.0080158
Ethiopia	2.9610725	0.6319500	4.685612	0.0000033
Guatemala	-0.8019487	0.2769682	-2.895454	0.0038981
Mexico	-1.7416522	0.2629071	-6.624592	0.0000000
Taiwan	-0.7015072	0.3505428	-2.001203	0.0457367

De lo anteriormente mostrado se desprende que, si bien los p-values son significativos en continente, país y variedad, no es así para tipo (robusta o arábica). Esta discrepancia se debe a la falta de registros de puntuación para el tipo de grano “robusta”. Tal como se ha mencionado anteriormente, en el caso de continente, país y variedad no se rechaza la hipótesis nula. Sin embargo, los valores de ajuste (R^2 y R^2 ajustada) muestran que la relación entre variables no sigue un comportamiento lineal, por lo que será necesario aplicar otros modelos explicativos y robustos cuyos resultados sean válidos, reproducibles y repetibles.

Del análisis anterior también se tiene evidencia estadística a favor de que los granos procedentes de Brasil, Etiopía, México y Colombia ya que estos obtienen mayores calificaciones por parte de los catadores de café, especialmente si proceden de Etiopía, que tendría en promedio un total de puntos obtenidos aproximadamente de 85.5 puntos (manteniendo el resto de variables constantes).

Por su parte, de la regresión lineal respecto al Tipo de grano obtenemos que la variable que representa al café Robusta es significativa al 10%. Por lo que, en promedio los cafés de Tipo Robusta obtienen un 1.2 puntos menos que las cafés de Tipo Arábica.

Hay que tener en cuenta que el tamaño de la muestra de cafés Robusta es muy pequeño: tan sólo 28 observaciones. Al tener una muestra tan pequeña puede hacer que este resultado no sea del todo fiable

En síntesis, se ha logrado determinar que las variables categóricas que influyen significativamente en los puntos totales de catado son:

- Country
- Continent
- Variety

y en menor medida:

- Type

Por último, Podemos construir un heatmap correlativo que muestre la covariación entre todas las variables cuantitativas (ver figura 10)

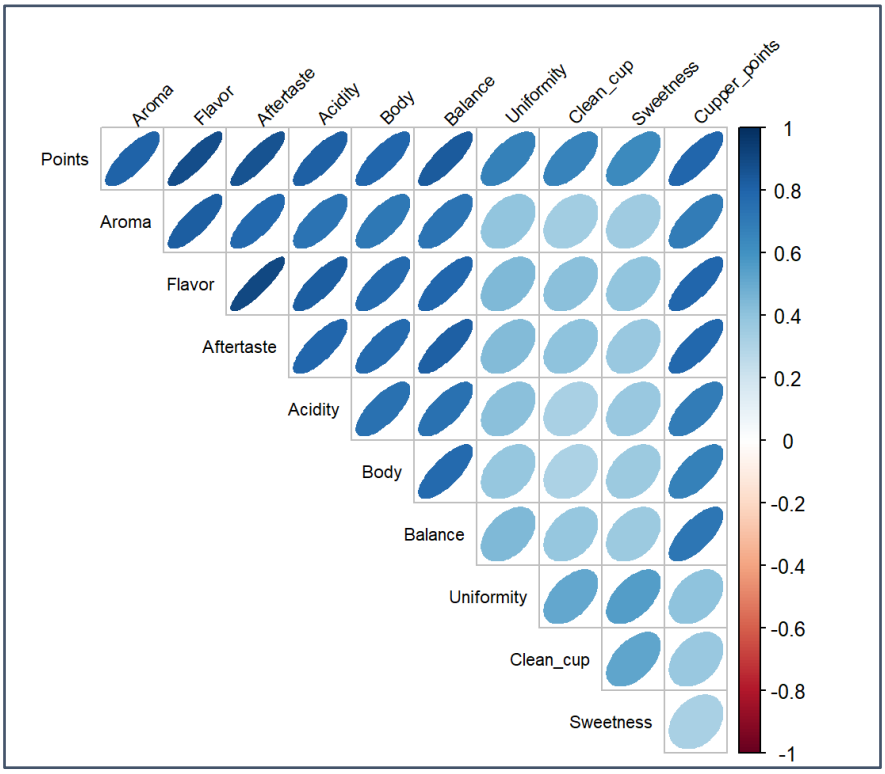


Figura 10. Diagrama de correlación entre las variables cuantitativas

En orden de correlación con la variable Points tenemos: Aroma, Flavor, Aftertaste, Acidity, Body, Balance, Uniformity, Clean_cup, Sweetness.

Modelos Multivariantes

A continuación, presentamos cuatro modelos de análisis multivariantes que explican e los puntajes obtenidos de acuerdo a su variedad de grano de café. Los cuatros modelos desarrollados son:

- Modelo KNN (Vecino más cercano)
- Modelo Random Forest
- Modelo de regresión lineal múltiple
- Modelo Regresión logística

Modelo de Vecino más cercano KNN

Mediante la librería *caret* diseñaremos un modelo de regresión utilizando el algoritmo del vecino más cercano (KNN, siglas en inglés). Para ello utilizaremos la data correspondiente a los 6 países con mayor puntaje de grano a saber: Etiopía, México, Colombia, Guatemala, Brasil y Taiwan.

Huelga decir que en el *dataset* original tenemos muy pocos registros de tipo Robusta por lo que la variable Type no será contemplada como regresor. Para solucionar esto se ha elegido la variable Variety como regresor siendo las variedades Caturra, Bourbon, Typica y Others las que serán analizadas para KNN (figura 11).

```
set.seed(569) ## FIJAMOS SEMILLA PARA OBTENER MISMOS RESULTADOS
features_1 <- features %>%
  filter (Country %in% c ("Ethiopia", "Mexico", "Colombia", "Guatemala", "Brazil", "Taiwan")) %>%
  filter(Variety %in% c("Caturra", "Bourbon", "Typica", "Other"))

#Descartamos la avriable Type puesto que solo tiene un unico nivel
Train <- select(features_1,-Type) %>%
  mutate_if(is.character, factor) #se transforma todas los atributos que sean chr a factor

#cargamos los registros de TEST
Test <- read.xlsx("../src/test.xlsx", sheetName = "Test") %>%
  filter (Country %in% c ("Ethiopia", "Mexico", "Colombia", "Guatemala", "Brazil", "Taiwan")) %>%
  filter(Variety %in% c("Caturra", "Bourbon", "Typica", "Other"))
Test
```

Figura 11. Sección de código con los datos de entrenamiento y prueba para el modelo de KNN

En la figura 12 se muestran los parámetros de control y configuración del KNN, así como la fase de entrenamiento. Vale destacar que se ha utilizado la métrica de evaluación “Cross-validation”

```r
#establecemos los parámetros de control
knn_ctrl <- trainControl(method = "repeatedcv", repeats = 2, classProbs = T)

#Modelo KNN
set.seed(607) #semilla para valores aleatorios
knnFit <- train(Variety ~., data = Train,
                method = "knn",
                trControl=knn_ctrl,
                tuneLength = 15)

knnFit
```

Figura 12. Parámetros de control del modelo KNN y fase de entrenamiento

Encontramos que el mejor clustering es 13 (k=13) con un *Accuracy* del 77%. En la siguiente gráfica se representa este hecho mediante el método del codo (ver figura 13).

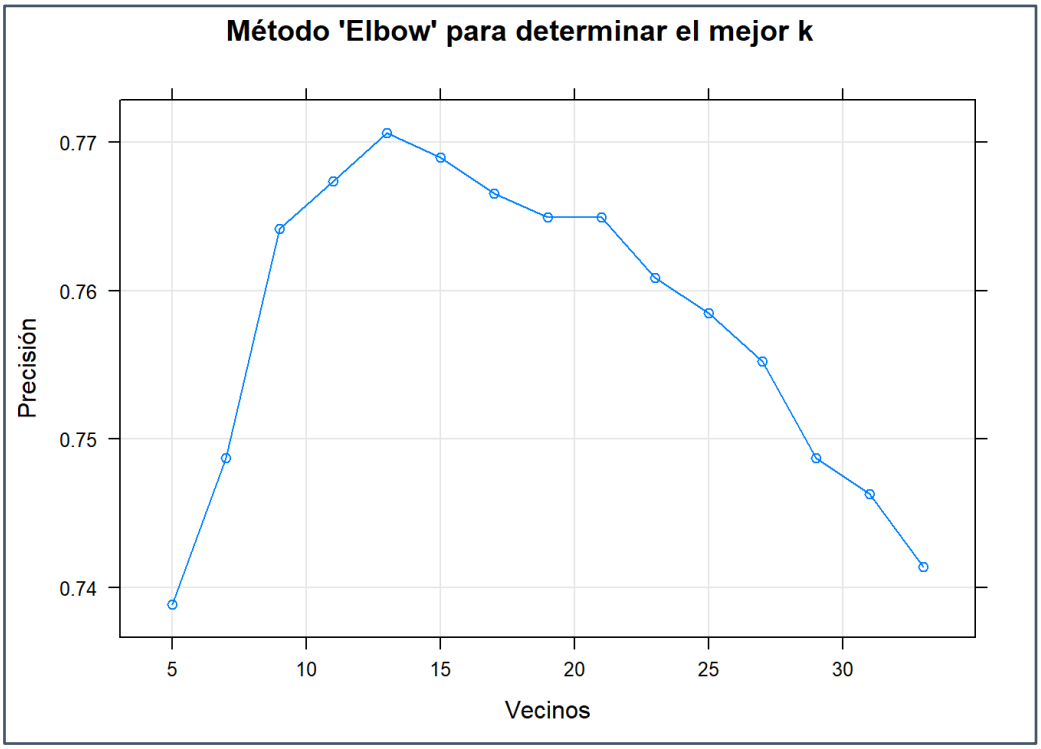


Figura 13. Resultado de la estimación del mejor K siguiendo el método del codo

De manera pormenorizada tenemos el resultado de los cálculos del entrenamiento KNN (ver tabla 9)

Tabla 9. Resultados de la simulación KNN

k	Accuracy	Kappa	AccuracySD	KappaSD
5	0.7388683	0.6220331	0.0516982	0.0741907
7	0.7487044	0.6341250	0.0431110	0.0631225
9	0.7641460	0.6562175	0.0390156	0.0572427
11	0.7673718	0.6602644	0.0431029	0.0632587
13	**0.7706108**	**0.6650476**	**0.0424669**	**0.0623944**
15	0.7689714	0.6626137	0.0417448	0.0612659
17	0.7665521	0.6591091	0.0369072	0.0542044
19	0.7649392	0.6567177	0.0455978	0.0669913
21	0.7649524	0.6568193	0.0416958	0.0612543
23	0.7608673	0.6508746	0.0368205	0.0541155
25	0.7584611	0.6474392	0.0366575	0.0538373
27	0.7552089	0.6426695	0.0425977	0.0625371
29	0.7487176	0.6332207	0.0471308	0.0691897
31	0.7462983	0.6297477	0.0427230	0.0626736
33	0.7414067	0.6225716	0.0486808	0.0714070

Una vez se ha determinado el mejor k, se evalua el poder predictivo del modelo mediante los datos de “test”. El resultado de la predicción del modelo se recoge en la tabla 10.

Tabla 10. Capacidad predictiva del modelo KNN para los datos de testing

Accuracy	0.8805031
Kappa	0.7730449
Accuracy Lower	0.8196937
Accuracy Upper	0.9264996
Accuracy Null	0.6289308
P-Value	0.0000000

De acuerdo a la tabla 10, el modelo tiene un 88% de capacidad de predicción. Una forma de comprobar esto es mediante la matriz de confusión (tabla 11), donde podremos evidencias los verdadero positivos detectados.

Gracias a la matriz de confusión logramos identificar que el modelo de KNN obtiene un *Accuracy* del 88,5%; o lo que es lo mismo, el modelo tiene un nivel de acierto del 88,5% de verdaderos positivos.

Tabla 11. Matriz de confusión del modelo KNN

	Bourbon	Caturra	Other	Typica
Bourbon	97	6	4	0
Caturra	0	21	0	0
Other	0	0	0	0
Typica	3	1	5	22

Para la métrica de evaluación *ROC (ver figura 14)* se requiere el uso de dos categorías, por ende, han utilizado las dos variedades más importantes: Caturra y Bourbon.

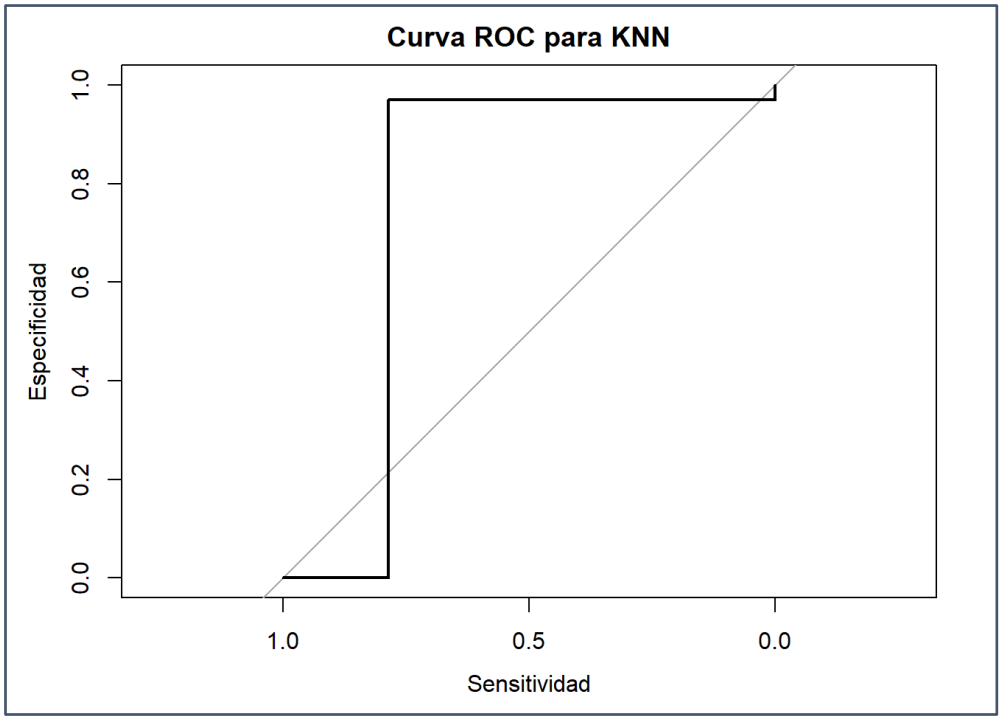


Figura 14. Curva ROC para el modelo KNN variedades Caturra y Bourbon

Modelo de Random Forest

Análogo al caso de KNN, evaluaremos la eficacia del modelo de random forest como método para predecir las variedades de café en función del resto de atributos.

Los parámetros de control y entrenamiento se muestran en la figura 15.

```
rf_control <- trainControl(method = "repeatedcv",
                           number = 10,
                           repeats = 3,
                           search = "random")
set.seed(633)

rf_fit <- train(Variety~.,
                       data=Train,
                       method='rf',
                       metric='Accuracy',
                       tuneLength = 15,
                       trControl=rf_control)
```

Figura 15. parametros de control y entrenamiento del modelo Random Forest. Elaboración propia

El modelo evaluó 13 predictores en 4 clases: Bourbon, Caturra, Other y Typica y como resultado determina el mejor numero de variables muestreadas aleatoriamente como candidatas en cada división (mtry = 3), con un “Accuracy” del 79,67% de efectividad. En la tabla 12 se exponen de manera más detallada los resultados de precisión y Kappa para cada simulación en RF.

Tabla 12. Resultados de la precisión del modelo RF. Elaboración propia

mtry	Accuracy	Kappa	AccuracySD	KappaSD
1	0.7163582	0.5885501	0.0495087	0.0717848
2	0.7945972	0.7014746	0.0558180	0.0819407
3	**0.7967654**	**0.7047314**	**0.0560308**	**0.0822129**
4	0.7913714	0.6970337	0.0574476	0.0841618
7	0.7767936	0.6761182	0.0540480	0.0792527
11	0.7697603	0.6664128	0.0530605	0.0776798
13	0.7687026	0.6651596	0.0549148	0.0803092
14	0.7660321	0.6611956	0.0527030	0.0769941
17	0.7622246	0.6560463	0.0520346	0.0756544

El modelo final contempla una tasa de error del 20,1% y 500 árboles totales. Seguidamente se hizo el testing del modelo comprobando con los valores de prueba para conocer cuantas variedades de grano fue capaz de predecir el modelo RF diseñado. Una forma de averiguarlo es conociendo la matriz de confusión (tabla 13)

Tabla 13. Matriz de confusión del modelo RF

	Bourbon	Catuai	Caturra	Other	Typica
Bourbon	98	0	6	4	0
Caturra	0	0	22	0	0
Other	0	0	0	0	0
Typica	2	0	0	5	22

L a matriz de confusión arroja un 89,31% de precisión predictiva, superior al caso del KNN. Por su parte, la curva ROC arrojó un área bajo la curva AUC de 0,875 y gráficamente puede evidenciarse en la figura 16.

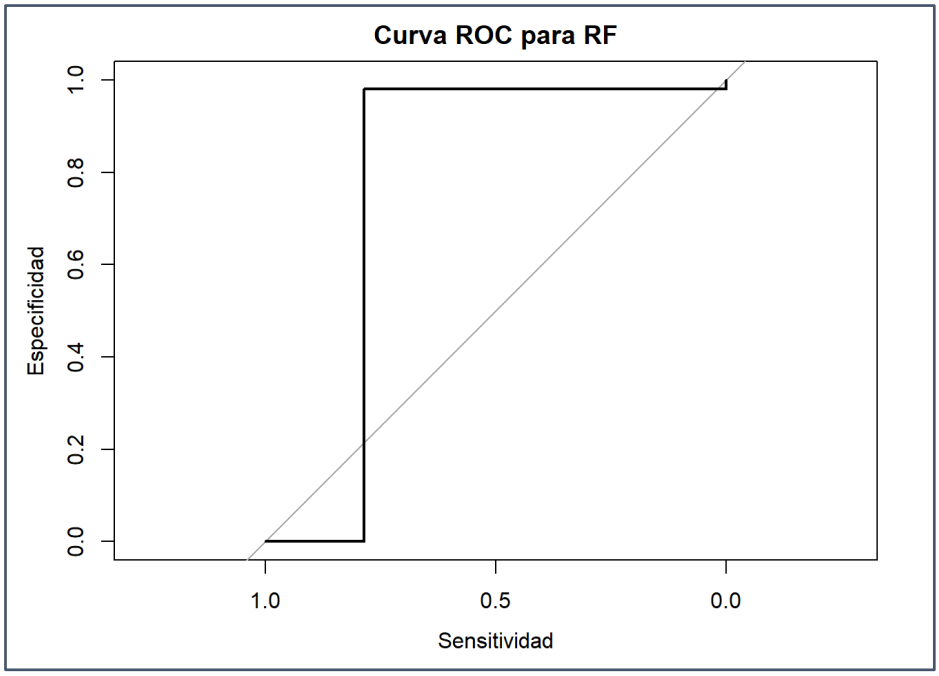


Figura 16. Curva ROC para modelo RF. Elaboración propia

Modelo de regresión logística

En el modelo de regresión logística necesitamos de manera excluyente, una categoría netamente binaria. Es por ello que para la variable Variety solo serán considerados los valores Caturra, Bourbon como categoría *Dummy.* Tanto las funciones de control como de entrenamiento del modelo logístico se muestran en la figura 17.

```
#función de control
logit_control <- trainControl(method = "repeatedcv",
                              number = 10,
                              repeats = 3,
                              search = "random")


set.seed(692)

#modelo de entrenamiento
logit_fit <- train(Variety ~.,
                   data=Train_logit,
                   method='glm',
                   family = "binomial",
                   metric='Accuracy',
                   tuneLength = 15,
                   trControl=logit_control)
```

Figura 17. Parámetros de configuración de control y entrenamiento para el modelo de regresión logística

La precisión calculada para los datos de entrenamiento fue del 83,34% como puede evidenciarse en la tabla 14.

Tabla 14. Resultados de la precisión y rendimiento del modelo de regresión logistica

Accuracy	Kappa	AccuracySD	KappaSD
0.8434092	0.6829208	0.0461952	0.0940685

Con el modelo de RLog también podemos estimar los coeficientes "β" que rigen el modelo completo (ver sección metodología). Estos coeficientes calculados se muestran en la tabla 15.

Tabla 15. Coeficientes de regresión del modelo logístico

	Betas
(Intercept)	-42.67527
Points	31.32242
Colombia	42.60480
Ethiopia	41.06828
Guatemala	19.41584
Mexico	20.16570
Taiwan	20.47911
Aroma	-31.86969
Flavor	-30.86734
Aftertaste	-30.78627
Acidity	-30.90568
Body	-30.94885
Balance	-33.24670
Uniformity	-30.79007
Clean_cup	-29.61023
Sweetness	-31.49246
Cupper_points	-30.45922

De la table 15 se desprende que las variables regresoras que correlacionan positivamente con la variedad del grano son la procedencia por país y los puntos de catado. En otras palabras, de acuerdo al modelo logístico, los puntos de evaluación totales y el país de origen son variables con buen acercamiento para determinar una variedad concreta de grano. Por otro lado, las variables relacionadas a las propiedades organolépticas del grano ofrecen un comportamiento de correlación negativa con la variedad.

Es decir, a mayores valores de puntuación en las propiedades del grano, menor la probabilidad de que el grano sea de la variedad Bourbon, y, por tanto, mayor probabilidad de que sea Caturra. Esta afirmación se ve sostiene cuando se observa el

grafico de la figura 7, en la que se evidencia que las mayores puntuaciones se corresponden con Caturra por encima de Bourbon. Por su parte, la matriz de covariación de RLog se muestra en la tabla 16

Tabla 16. Matriz de covariación del modelo de regresión logística. Elaboración propia

	Points	Aroma	Flavor	After	Acidity	Body	Balance	Unif.	Clean_c.	Sweet	Cupper
Points	-11.52431	-1.3131528	-1.7290565	-1.6027983	-1.3217318	-1.2159320	-1.4726216	-0.5667216	-0.1106365	-0.2221444	-1.9691550
Aroma	0.00000	-0.8992547	-0.0372995	0.1357266	0.0597393	-0.0501416	0.0316402	0.3308865	0.1577176	0.2664337	0.0093938
Flavor	0.00000	0.0000000	0.7928374	0.2185308	-0.0926913	-0.1399022	-0.0520454	-0.3363545	-0.1536866	-0.2319530	-0.0042890
After	0.00000	0.0000000	0.0000000	-0.7869316	0.1342314	0.0730864	-0.0643270	0.2593089	0.1871315	0.2212876	-0.0221752
Acidity	0.00000	0.0000000	0.0000000	0.0000000	-0.9513644	-0.0297028	-0.0031504	0.3732819	0.1560479	0.2433043	0.2161433
Body	0.00000	0.0000000	0.0000000	0.0000000	0.0000000	-1.0272813	-0.2319061	0.6671967	0.1412119	0.4256739	0.0275367
Balance	0.00000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.7036573	-0.3732782	-0.0948660	-0.2841731	0.0426833
Unif.	0.00000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.8996413	-0.1050460	-0.2340453	-0.5551185
Clean_c.	0.00000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	-0.5180491	0.0084578	0.5025644
Sweet	0.00000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	-0.6134845	0.6124777
Cupper_p.	0.00000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	-0.0448137

Podemos ahora observar la matriz de confusión resultante después de evaluar la predicción del modelo con los datos de prueba (tabla 17).

Tabla 17. Matriz de confusión modelo regresión logistica

	Bourbon	Caturra
Bourbon	100	6
Caturra	0	22

La precisión del modelo RLog es de 95,31% al 95% de confianza y un p-value < 0.05. Por su parte, la curva ROC arrojó un área AUC de 0.829 (ver figura 18)

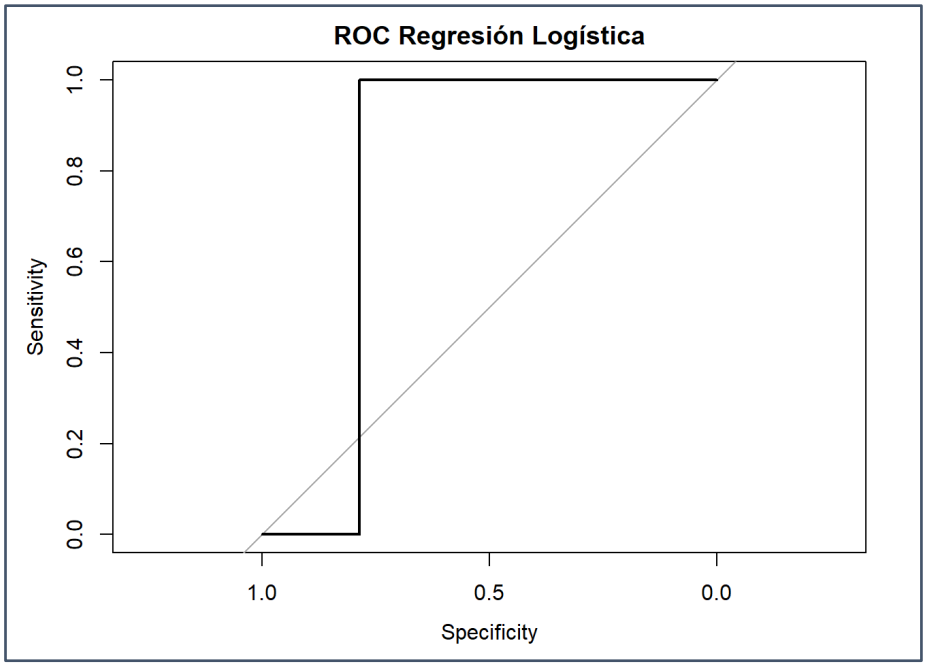


Figura 18. ROC para modelo regresión logística.

Modelo de regresión lineal múltiple

En el modelo de regresión lineal múltiple se toman en cuenta solo variables continas. Por tanto, buscaremos evaluar la significancia regresiva entre los puntos de café y las variables de Aroma, Sabor, Retrosabor, Acidez, Cuerpo, Balance, Uniformidad, Taza limpia, dulzor y puntos de taza. La ecuación a modelar se ha estipulado en el apartado de la metodología.

Tabla 18. Coeficientes beta del modelo de regresión lineal entre puntos de calidad y propiedades del grano

	Betas
(Intercept)	-0.0663120
Aroma	1.0020160
Flavor	1.0030780
Aftertaste	1.0042671
Acidity	0.9972019
Body	1.0041390
Balance	1.0004865
Uniformity	1.0085126
Clean_cup	1.0000970
Sweetness	0.9973011
Cupper_points	0.9899032

La regresión lineal arrojó un R^2 del 0.999; dando a entender que existe una fuerte relación lineal entre los puntos de café y el resto de variables predictoras antes mencionadas. Básicamente podemos interpretar que, a mayor puntaje obtenido en los atributos independientes, mayor el puntaje final de calidad; siendo este un comportamiento que ya se evidenciaba en las técnicas de análisis exploratorio.

Cuando se evalúa el poder predictivo de este modelo múltiple se obtiene un *Accuracy* de alrededor del 100%. Demostrando el carácter fuertemente lineal que tienen los puntos pronosticados y los puntos reales del test. En la figura 19 se presenta la correlación entre los datos reales de puntuación y los calculados

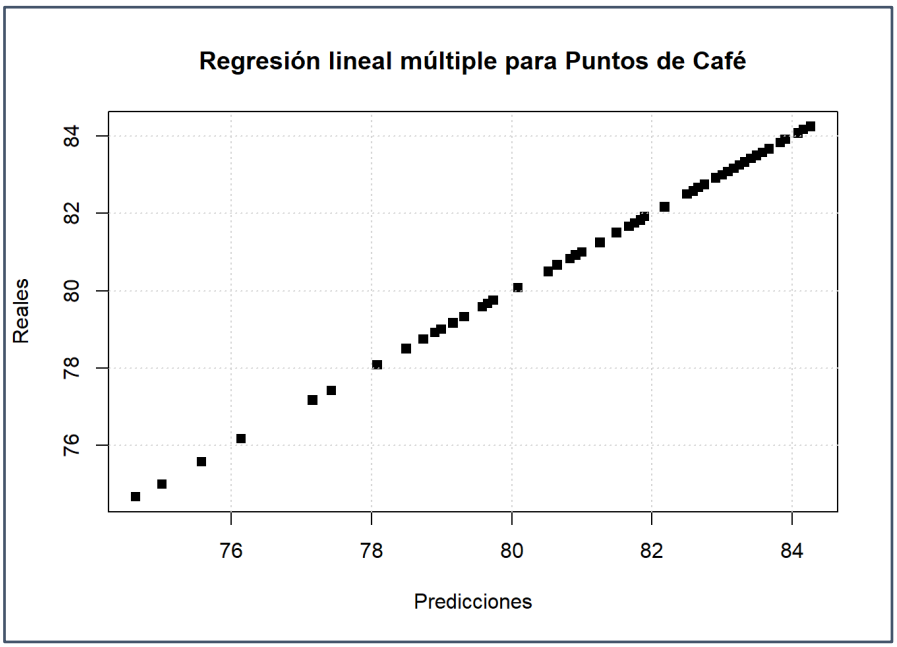


Figura 19. Regresión lineal múltiple para puntos de café

Resumen de Modelos

En este apartado comparamos los parámetros estadísticos obtenidos de cada modelo para determinar la elección de aquel que mejor se ajusta a la naturaleza del problema.

Tabla 19. Resumen de parametros estadisticos por modelo probado

Modelo	Accuracy	AUC	P-Value
KNN	88,50%	0,866	9,76E-13
RF	89,31%	0,877	4,63E-14
RLog	**95,31%**	**0,892**	**5,949E-8**
RLM	99,99%	1,000	0,000

En la tabla 19 se muestran parámetros como precisión, área bajo la curva y valor de p con intervalo de confianza del 95%. De los cuatro modelos, el que mejor se ajusta dentro de los algoritmos de clasificación, es el de regresión logística; ya que cuenta con un nivel de acierto de 95,31% y un área bajo la curva AUC de 0,892. Siendo muy superior que los obtenidos por KNN y RF. No obstante, vale resaltar que este modelo (RLog) solo es aplicable, siempre que los valores a predecir sean en función de dos variedades: Caturra y Bourbon. Esta limitación de valores frontera dentro del modelo RLog puede ser la razón de que su precisión haya sido mayor.

Por consiguiente, si queremos mantener la rigurosidad de la investigación será vital que el algoritmo de clasificación sea múltiple, de tal manera que el *clustering* pueda evaluar más de dos clases. En tal sentido, se ha evidente que el algoritmo que mejor se ajusta a estas exigencias es el de RF. Este modelo obtuvo una mejor precisión y área bajo la curva AUC con respecto a al modelo KNN que también tomaba en consideración la misma cantidad de clases (Variedades de grano)

Ahora bien, si tomamos en cuenta que el costo técnico que implica el uso de algoritmos de clasificación, se podría decir que el modelo KNN es más eficiente que el RF. El tiempo de cálculo de entrenamiento que requiere KNN para hacer estimaciones decentes del 88,50% es de 3,53 segundos, mientras que al modelo RF le toma 109 segundos. Basado en estos puntos resulta claro que el modelo KNN es más eficiente ya que hace predicciones apropiadas y reproducibles en mucho menor tiempo y con buena precisión.

En suma, si el investigador desea únicamente evaluar dos variedades de grano se recomienda ampliamente aplicar el modelo RLog, y en el caso que desee estudiar múltiples opciones de variedad tomando en cuenta la potencia de cálculo y la precisión, el modelo KNN será el más apropiado. De cualquier forma, en ambos casos los modelos están diseñados con base en los puntos de calidad, propiedades organolépticas y procedencia del grano.

Conclusiones

Llegados a este punto del proyecto podemos señalar los resultados más relevantes obtenidos en función con los objetivos del trabajo. En primer lugar, se ha podido analizar de manera completa y rigurosa la base de datos en cuestión, teniendo en consideración la heterogeneidad de los registros, las dimensiones del marco de datos y los diferentes tipos de datos en cada atributo analizado. Con en este análisis riguroso se ha podido responder a uno de los propósitos principales del estudio, a saber, el de averiguar cuáles son las características y factores que hacen que un grano de café se de mayor calidad que los otros.

Gracias a los resultados obtenidos de los 4 modelos multivariantes utilizados, en los 3 de clasificación se evidenció que tanto la procedencia del grano (país y continente de origen) y sus propiedades cata, juegan un papel importante al momento de determinar la variedad del grano. Se encontró que mayores puntuaciones de café correspondían a las variedades Caturra y Bourbon, del tipo arábico, y que en su mayoría provenían de del continente americano. Los países protagonistas fueron México, Guatemala, Colombia y Brasil. En menor medida tendríamos la variedad “Typica” que también mostraba buenas puntuaciones de catado. Otro de los países que mostraron una calidad de café adecuada, pero no tanto como en los anteriores, fueron Taiwán y Etiopía.

De manera pormenorizada, se ha logrado determinar que la variedad predilecta de los catadores de café es Caturra, del tipo arábico, ya que esta ostentaba los mas altos puntajes de cata. Para llegar a esta conclusión, se segmentó apropiadamente el dataset en función de las propiedades organolépticas (aroma, sabor, retrosabor, cuerpo, uniformidad… etc.) y la procedencia del grano.

Finalmente acotar que, la realización de análisis exploratorio de los datos jugó un papel fundamental al ofrecer una información visual y tabulada sobre el comportamiento de los datos y sus posibles desviaciones. Con esta información se pudo construir un juicio de valor robusto para la elección de los modelos multivariantes definitivos: KNN, RF, RLog y RLM. Con la elección de estos modelos se procedió exitosamente a sus implementaciones dentro del entorno RStudio desde donde se pudieron obtener todos los resultados mostrados en este proyecto de investigación.

Limitaciones

Durante la elaboración del estudio se han identificado las siguientes limitaciones:

- El data set solo consta de 1328 registros, lo cual es aún una muestra relativamente pequeña si la comparamos con los esquemas del Big Data. Lo cual hace que los resultados sean únicamente significativos en este caso particular y su reproducibilidad científica sea menos rigurosa.
- De los 1328 registros, solo 28 correspondían al tipo robusta, por lo que fue necesario dirigir el análisis a las variedades de grano en vez de a su tipo. Por consiguiente, los resultados obtenidos solo se limitan al tipo de grano “arabica”. Es muy posible que las puntuaciones de robusta sean muy diferentes a las aquí obtenidas, así esto complementa lo mencionado en el ítem anterior.
- El dataset cuenta con una cantidad importante de valores perdidos, por lo que al depurarlo disminuye la precisión real que tendría cualquiera de los modelos utilizados.
- No se realizaron detecciones de Outliers dentro del dataset.
- No se aplicaron modelos de aprendizaje no supervisado, como K-Means, Clustering jerárquico o gaussiano.

Recomendaciones

- Se recomienda el uso de más modelos de aprendizaje supervisado como el Naive-Bayes Ingenuo, Árbol de Decisión y SVM
- Una manera de comprobar y mejorar los resultados obtenidos sería considerando otros tipos de herramientas de aprendizaje, como pueden ser la no supervisadas: K-Means, Clustering jerárquico y reducción dimensional.
- Debido a que la base de datos es relativamente pequeña, se recomienda repetir los procedimientos aquí descritos con un dataset con cientos de miles de registros, los cuales podrían obtenerse mediante series de tiempo.
- Realizar análisis de detección de outliers en la base de datos, por ejemplo, mediante cálculos de los cuartiles de cada atributo.
- Incluir dentro del dataset una cantidad importante de registros robusta y de esta forma aplicar la clasificación con base al tipo de grano
- Realizar un análisis de serie temporal tomando en cuenta los puntos de cata y variables organolépticas en función de los meses, de esta manera identificar si existe alguna especie de estacionalidad en las puntuaciones de café. Esta idea surge como una forma de evaluar si la época de cosecha del grano de café influye en su puntación final.

Bibliografía

Anthony, F., Bertrand, B., Quiros, O., Wilches, A., Lashermes, P., Berthaud, J., y Charrier, A. (2001). Genetic diversity of wild coffee (Coffea arabica L.) using molecular markers. *Euphytica*, *118*(1), 53-65. https://doi.org/10.1023/A:1004013815166

Barbosa, R. M., Batista, B. L., Varrique, R. M., Coelho, V. A., Campiglia, A. D., y Barbosa, F. (2014). The use of advanced chemometric techniques and trace element levels for controlling the authenticity of organic coffee. *Food Research International*, *61*, 246-251. https://doi.org/10.1016/j.foodres.2013.07.060

Bookman, S. (2014). Brands and urban life: Specialty coffee, consumers, and the co-creation of urban café sociality. En *Space and Culture* (Vol. 17, Número 1, pp. 85-99). https://doi.org/10.1177/1206331213493853

Caporaso, N., Whitworth, M. B., Cui, C., y Fisk, I. D. (2018). Variability of single bean coffee volatile compounds of Arabica and robusta roasted coffees analysed by SPME-GC-MS. *Food Research International*, *108*, 628-640. https://doi.org/10.1016/j.foodres.2018.03.077

Carvalho, N. B., Minim, V. P. R., Nascimento, M., Vidigal, M. C. T. R., Ferreira, M. A. M., Gonçalves, A. C. A., y Minim, L. A. (2015). A discriminant function for validation of the cluster analysis and behavioral prediction of the coffee market. *Food Research International*, *77*, 400-407. https://doi.org/10.1016/j.foodres.2015.10.013

Çelik, E. E., y Gökmen, V. (2018). A study on interactions between the insoluble fractions of different coffee infusions and major cocoa free antioxidants and different coffee infusions and dark chocolate. *Food Chemistry*, *255*, 8-14. https://doi.org/10.1016/j.foodchem.2018.02.048

Condliffe, K., Kebuchi, W., Love, C., y Ruparell, R. (2008). *Kenya coffee: a cluster analysis*. https://www.isc.hbs.edu/Documents/resources/courses/moc-course-at-harvard/pdf/student-projects/Kenya_Coffee_2008.pdf

de Morais, T. C. B., Rodrigues, D. R., de Carvalho Polari Souto, U. T., y Lemos, S. G. (2019). A simple voltammetric electronic tongue for the analysis of coffee adulterations. *Food Chemistry*, *273*, 31-38. https://doi.org/10.1016/j.foodchem.2018.04.136

Deng, Z., Zhu, X., Cheng, D., Zong, M., y Zhang, S. (2016). Efficient kNN classification algorithm for big data. *Neurocomputing*, *195*, 143-148.

https://doi.org/10.1016/j.neucom.2015.08.112

Dranoff, J. A. (2018). Coffee consumption and prevention of cirrhosis: In support of the caffeine hypothesis. En *Gene Expression* (Vol. 18, Número 1, pp. 1-3). https://doi.org/10.5221617/X15046391179559

Dreiseitl, S., y Ohno-Machado, L. (2002). Logistic regression and artificial neural network classification models: A methodology review. *Journal of Biomedical Informatics*, *35*(5-6), 352-359. https://doi.org/10.1016/S1532-0464(03)00034-0

Fernández, P. (2018). *Introducción a Machine Learning con Python (Parte 2)* | *Pybonacci*. https://pybonacci.org/2015/04/06/introduccion-a-machine-learning-con-python-parte-2/

Flament, I. (2002). *Coffee Flavor Chemistry* (p. 424). https://books.google.es/books?hl=es&lr=&id=NQi1LYJxFvUC&oi=fnd&pg=PP11&dq=Flament,+I.+Coffee+Flavor+Chemistry,+1st+ed.%3B+John+Wiley+%26+Sons:+Hoboken,+NJ,+USA,+2001%3B+pp.+1–424.+ISBN+978-0-471-72038-6&ots=dSMe5k4V-t&sig=fprkGK_OE53E1zx7vB2mXBDrNOU

Graus, M. E. G. (2018). Estadística aplicada a la investigación educativa. *Dilemas Contemporáneos: Educación, Política y Valores*.

Gurucharan, M. K. (2020). *Machine Learning Basics: Decision Tree Regression*. toward data science. https://towardsdatascience.com/machine-learning-basics-random-forest-regression-be3e1e3bb91a

Huacasi, H. Y. P. (2020). *Bosques Aleatorios*. Medium. https://medium.com/@hpumah/bosques-aleatorios-482163ace92e

ICO. (2020). *Coffee Market Reports. The Current State of the Global Coffee Trade*. https://www.ico.org/documents/cy2020-21/cmr-0721-e.pdf

Irmeilyana, Ngudiantoro, Samsuri, M. N., y Suprihatin, B. (2021). Logistic regression model on land productivity of Pagar Alam coffee farming. *Journal of Physics: Conference Series*, *1943*(1), 012135. https://doi.org/10.1088/1742-6596/1943/1/012135

Korhoňová, M., Hron, K., Klimčíková, D., y Talanta, L. M. (2019). Coffee aroma—Statistical analysis of compositional data. *Elsevier*. https://www.sciencedirect.com/science/article/pii/S0039914009006249

Maciejewski, G., Mokrysz, S., y Wróblewski, Ł. (2019). Segmentation of coffee

consumers using sustainable values: Cluster analysis on the Polish coffee market. *Sustainability (Switzerland)*, *11*(3). https://doi.org/10.3390/su11030613

Manzo, J. (2014). Machines, people, and social interaction in “third-wave” coffeehouses. *mail.theartsjournal.org*. https://mail.theartsjournal.org/index.php/site/article/view/527

Mitchell, A. C. (2018). *Tracing Economic Sustainability in the Global Coffee Trade: The Rhetoric of the International Coffee Organization*. https://ttu-ir.tdl.org/handle/2346/73818

Nevins, J., Market, N. P.-T. S. A. to, y 2018, U. (2019). Introduction: Commoditization in Southeast Asia. En *Taking Southeast Asia to Market* (pp. 1-24). https://doi.org/10.7591/9781501732270-003

O’Keefe, J. H., DiNicolantonio, J. J., y Lavie, C. J. (2018). Coffee for Cardioprotection and Longevity. En *Progress in Cardiovascular Diseases* (Vol. 61, Número 1, pp. 38-42). https://doi.org/10.1016/j.pcad.2018.02.002

Paliwal, M., y Kumar, U. A. (2009). Neural networks and statistical techniques: A review of applications. En *Expert Systems with Applications* (Vol. 36, Número 1, pp. 2-17). https://doi.org/10.1016/j.eswa.2007.10.005

Palmieri, M. G. S., Cruz, L. T., Bertges, F. S., Húngaro, H. M., Batista, L. R., da Silva, S. S., Fonseca, M. J. V., Rodarte, M. P., Vilela, F. M. P., y Amaral, M. da P. H. do. (2018). Enhancement of antioxidant properties from green coffee as promising ingredient for food and cosmetic industries. *Biocatalysis and Agricultural Biotechnology*, *16*, 43-48. https://doi.org/10.1016/j.bcab.2018.07.011

Park, H., Suh, B. S., y Lee, K. (2019). Relationship between daily coffee intake and suicidal ideation. *Journal of Affective Disorders*, *256*, 468-472. https://doi.org/10.1016/j.jad.2019.06.023

Runtuwene, J. P. A., Tangkawarow, I. R. H. T., Manoppo, C. T. M., y Salaki, R. J. (2018). A Comparative Analysis of Extract, Transformation and Loading (ETL) Process. *IOP Conference Series: Materials Science and Engineering*, *306*(1). https://doi.org/10.1088/1757-899X/306/1/012066

Samoggia, A., y Riedel, B. (2018). Coffee consumption and purchasing behavior review: Insights for further research. En *Appetite* (Vol. 129, pp. 70-81). https://doi.org/10.1016/j.appet.2018.07.002

Sänger, C. (2018). State of the global coffee market. *United Nations Conference on*

Trade and Development 10th Multi-Year expert meeting on commodities and development, 36. https://unctad.org/system/files/non-official-document/MYEM2018_Christoph Saenger_25042018.pdf

Sayad, S. (2021). *KNN Regression*. https://www.saedsayad.com/k_nearest_neighbors_reg.htm

Schonlau, M., y Zou, R. Y. (2020). The random forest algorithm for statistical learning. *Stata Journal*, *20*(1), 3-29. https://doi.org/10.1177/1536867X20909688

Seninde, D. (2020). Coffee flavor: A review. *mdpi.com*. https://www.mdpi.com/763812

Subasi, A., y Erçelebi, E. (2005). Classification of EEG signals using neural network and logistic regression. *Computer Methods and Programs in Biomedicine*, *78*(2), 87-99. https://doi.org/10.1016/j.cmpb.2004.10.009

Sunarharum, W. B., Williams, D. J., y Smyth, H. E. (2014). Complexity of coffee flavor: A compositional and sensory perspective. En *Food Research International* (Vol. 62, pp. 315-325). https://doi.org/10.1016/j.foodres.2014.02.030

Suslick, B. A., Feng, L., y Suslick, K. S. (2010). Discrimination of complex mixtures by a colorimetric sensor array: Coffee aromas. *Analytical Chemistry*, *82*(5), 2067-2073. https://doi.org/10.1021/ac902823w

Toledo, P. R. A. B., Pezza, L., Pezza, H. R., y Toci, A. T. (2016). Relationship Between the Different Aspects Related to Coffee Quality and Their Volatile Compounds. *Comprehensive Reviews in Food Science and Food Safety*, *15*(4), 705-719. https://doi.org/10.1111/1541-4337.12205

Voora, V., Bermúdez, S., y Larrea, C. (2019). *Global market report: Coffee*. https://www.iisd.org/system/files/publications/ssi-global-market-report-coffee.pdf

Wintgens, J. (2004). *Coffee: growing, processing, sustainable production. A guidebook for growers, processors, traders, and researchers.* https://www.cabdirect.org/cabdirect/abstract/20053043070

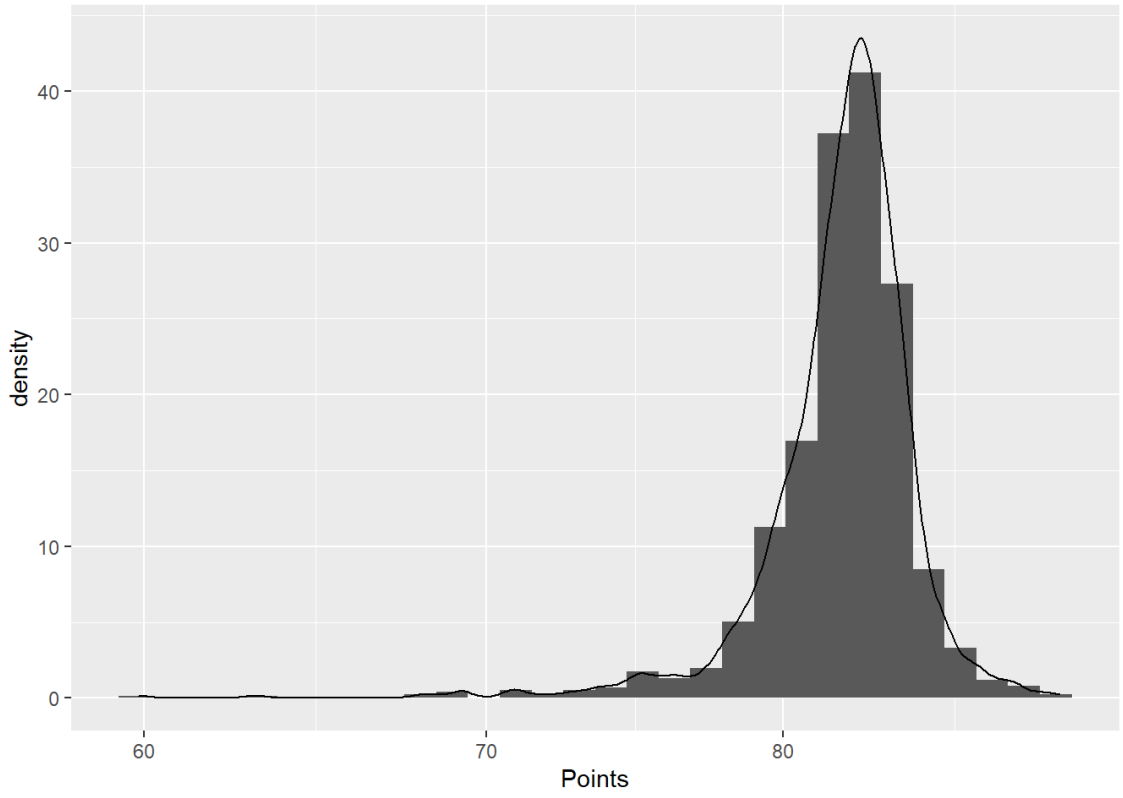
Yannis, P., y Nikolaos, B. (2018). Quantitative and Qualitative Research in Business Technology: Justifying a Suitable Research Methodology. *Review of Integrative Business and Economics Research*, *7*(1), 91-105. https://sibresearch.org/uploads/3/4/0/9/34097180/riber_7-s1_sp_h17-083_91-105.pdf

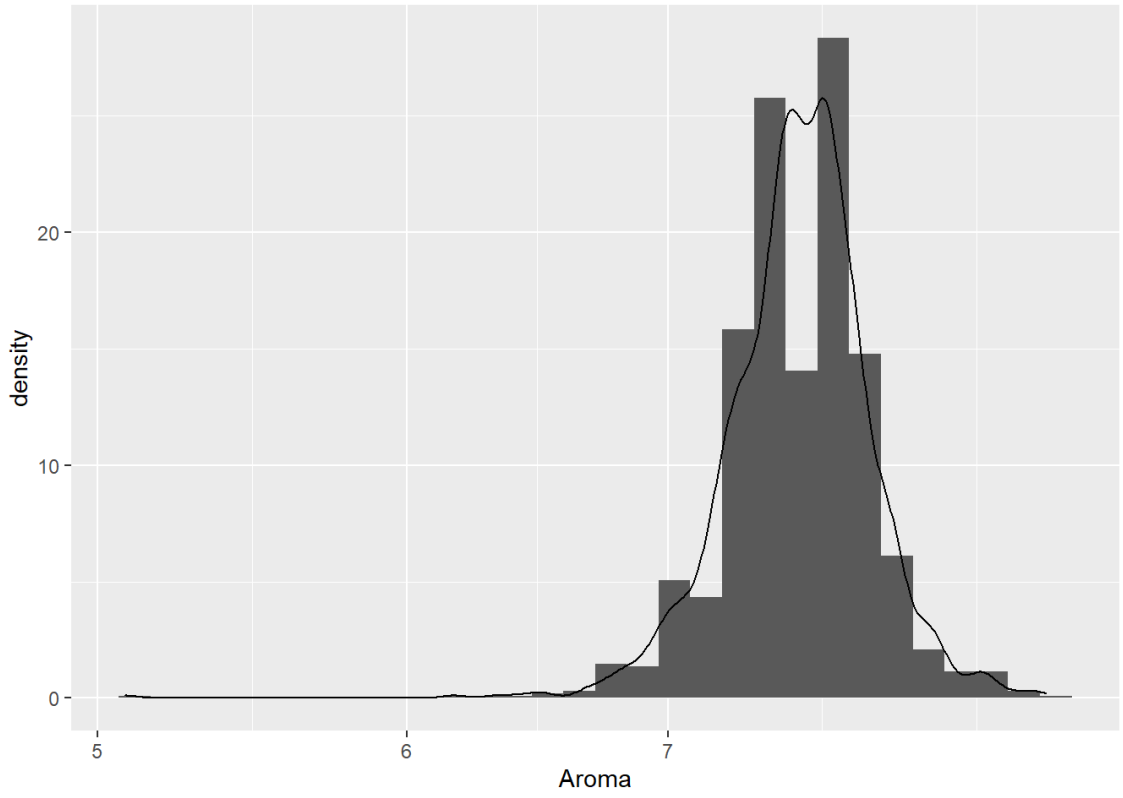
Zhang, S., Li, X., Zong, M., Zhu, X., y Wang, R. (2018). Efficient kNN classification with

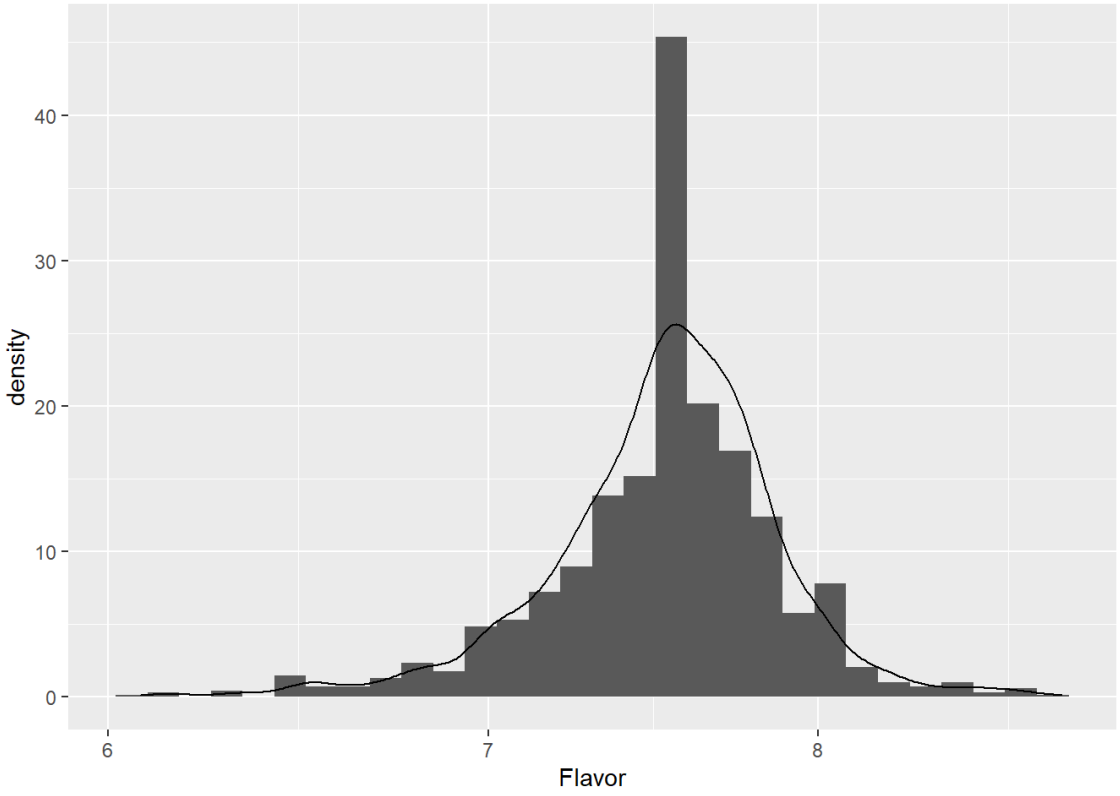
different numbers of nearest neighbors. *IEEE Transactions on Neural Networks and Learning Systems*, *29*(5), 1774-1785. https://doi.org/10.1109/TNNLS.2017.2673241

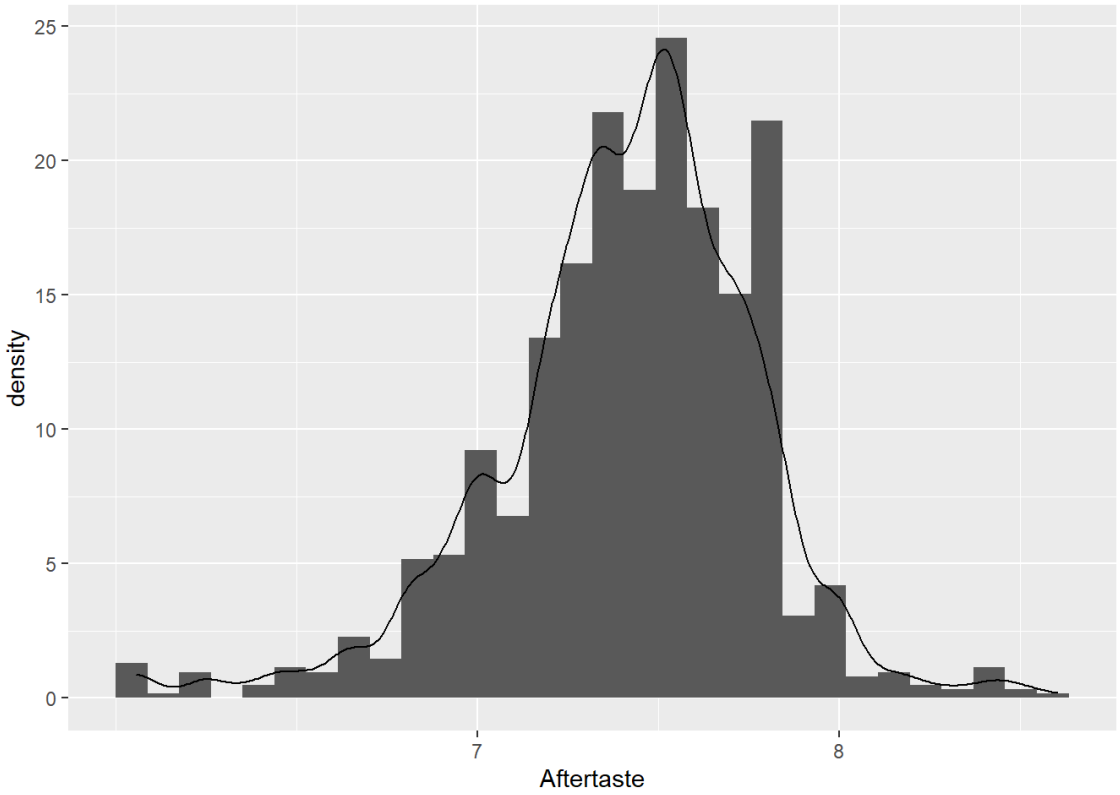
Anexos

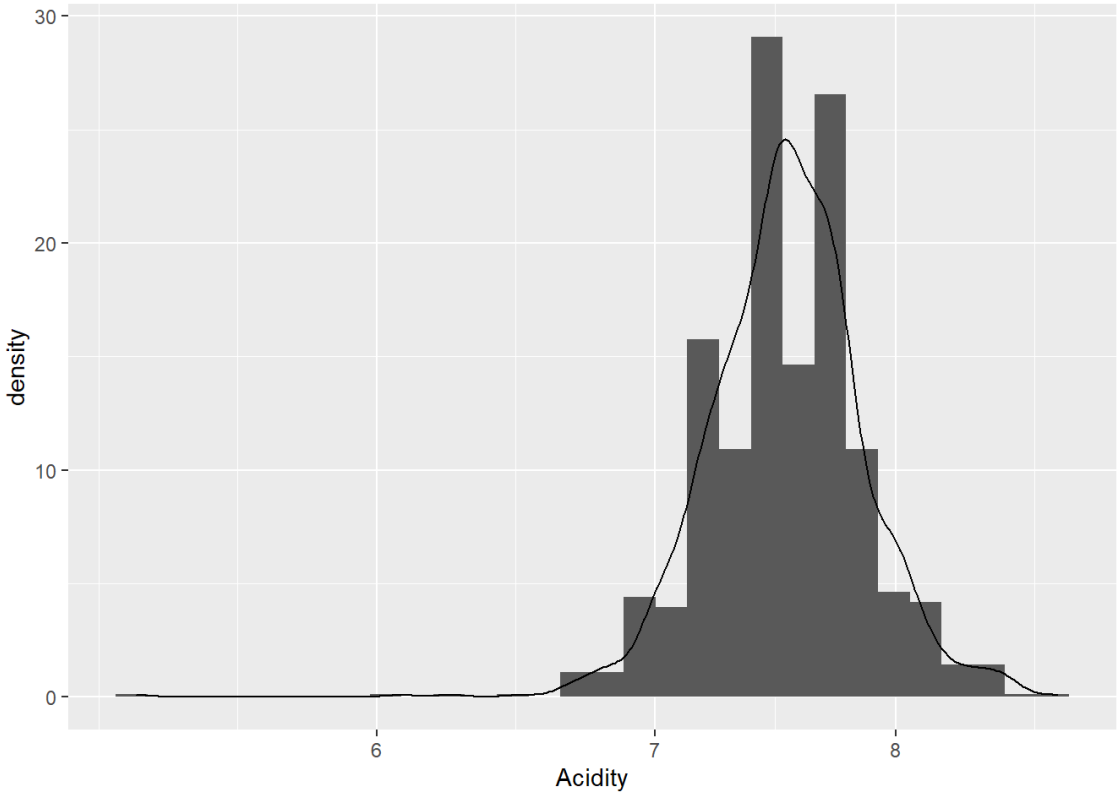
Gráficos adicionales

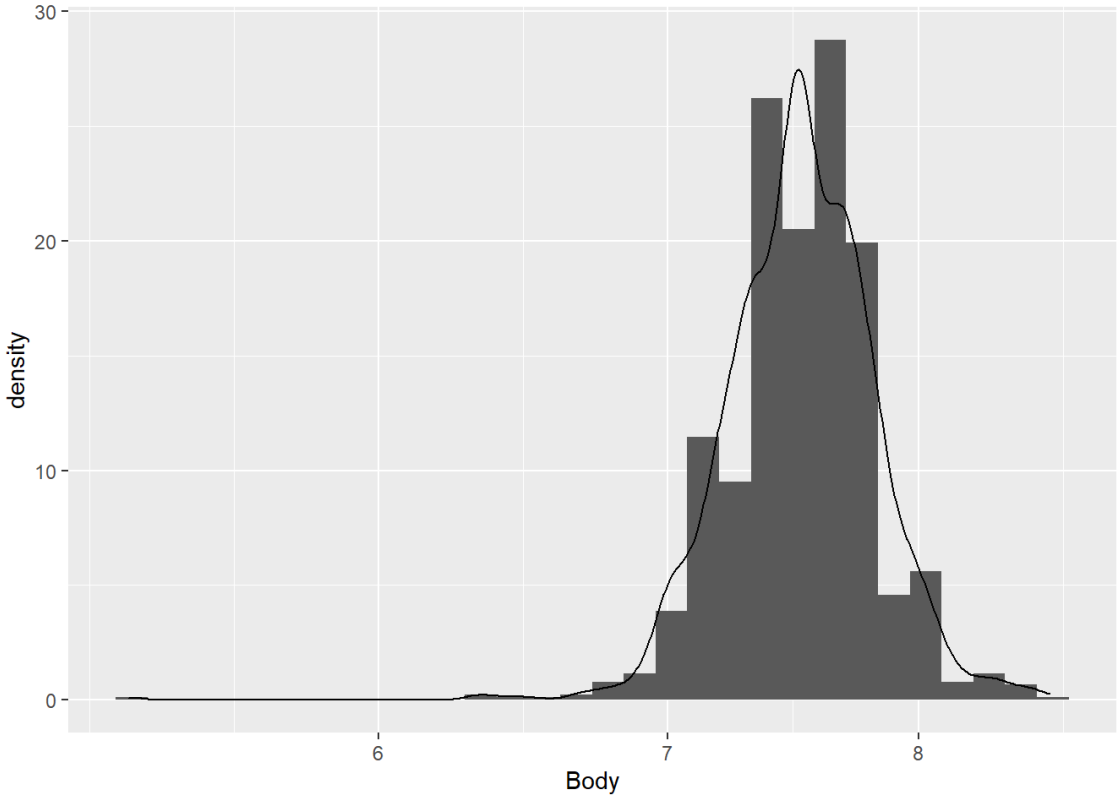


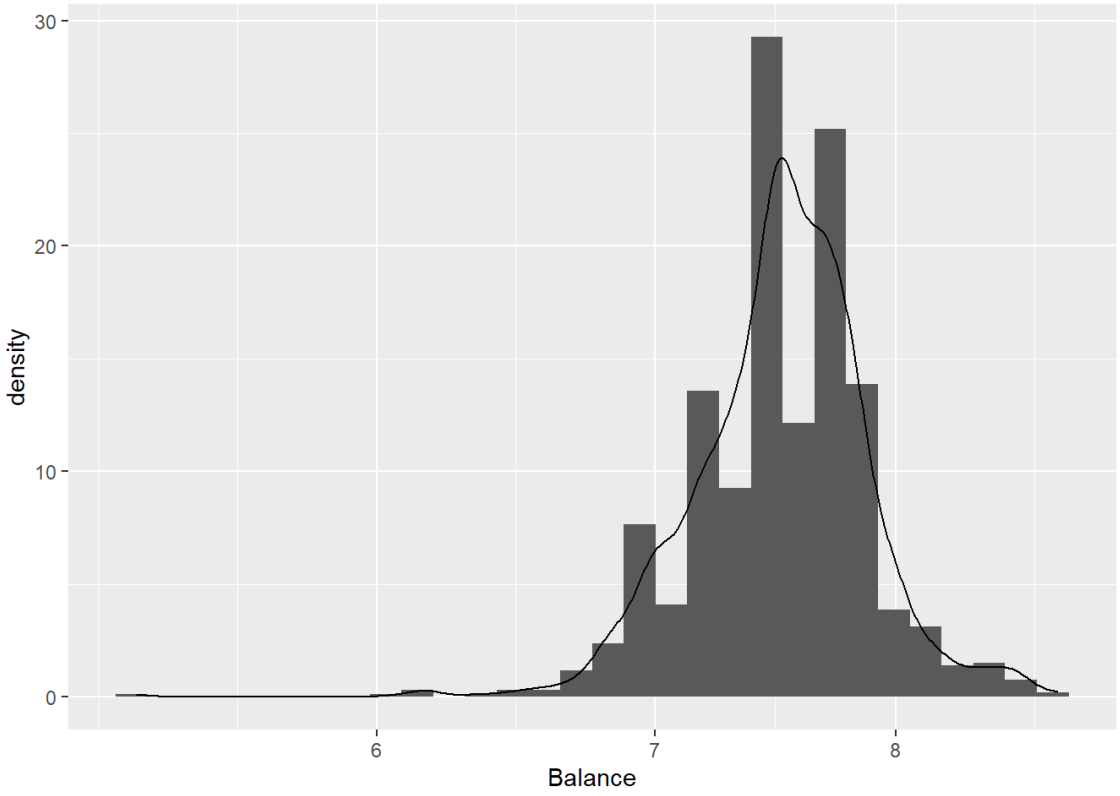


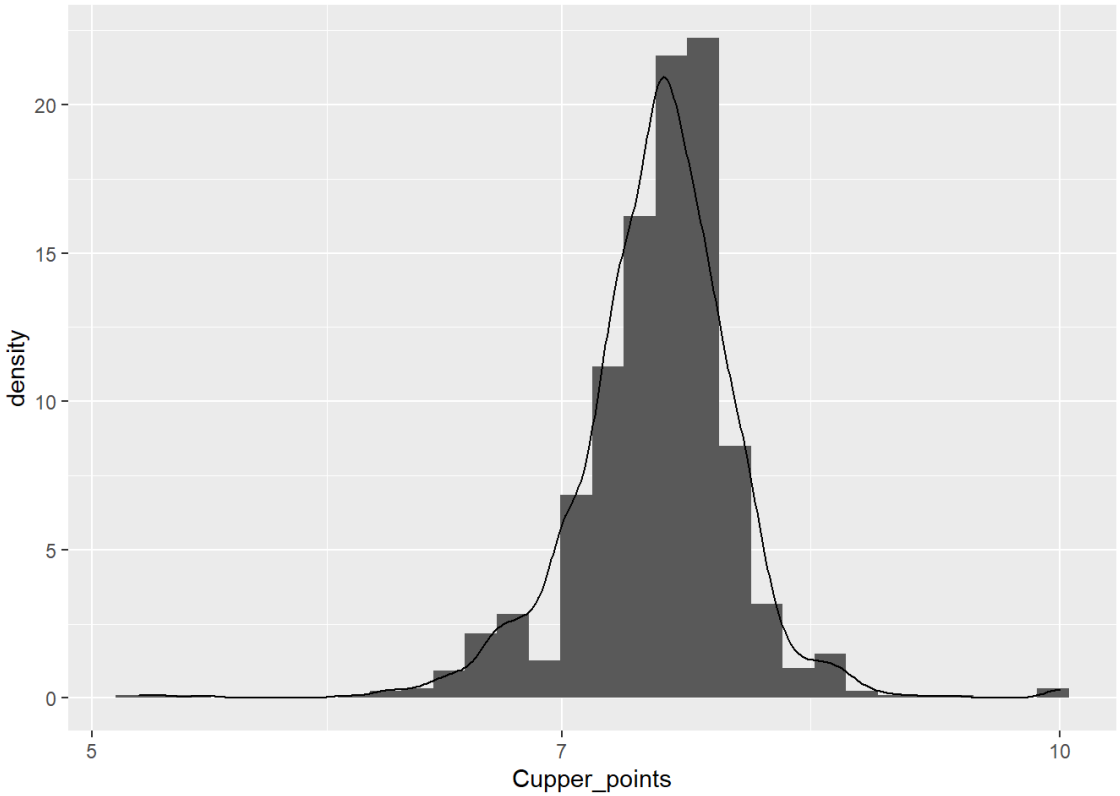

density
40
30
20
10
0
6
7
8
Flavor


density
25
20
15
10
5
0
7
8
Aftertaste


30
20
10
0
density
6
7
8
Acidity


30
20
10
0
density
6
7
8
Body


30
20
10
0
density
6
7
8
Balance


20
15
10
5
0
density
5
7
10
Cupper_points

Script en R

```
# VÍCTOR MANUEL LÓPEZ MENDOZA

## ----setup, include=FALSE----------------------------------------------------

knitr::opts_chunk$set(echo = TRUE)

## fijar directorio

#Establecer directorio de origen


## ----datos, include = FALSE, warning=FALSE, echo=FALSE-------------------------
# llamando las librerÃ-as necesarias para cargar y procesar los archivos

library(tidyverse)
library(tinytex)
library(rvest)
library(janitor)
library(caret) #caret que permite la aplicaciÃƒÂ³n de los mÃƒÂ©todos supervisados
library(dplyr) #funcionalidades adicionales para R
library(xlsx) #lectura y escritura de archivos xls y csv.
library(pROC) #construir grÃƒÂ¡fico ROC
library(MLmetrics) #librerÃƒÂ-a necesaria para randomForest


raw_arabica <- read_csv("../data/arabica_data_cleaned.csv")



raw_robusta <- read_csv("../data/robusta_data_cleaned.csv",
                        col_types = cols(
                          X1 = col_double(),
                          Species = col_character(),
                          Owner = col_character(),
                          Country.of.Origin = col_character(),
                          Farm.Name = col_character(),
                          Lot.Number = col_character(),
                          Mill = col_character(),
                          ICO.Number = col_character(),
                          Company = col_character(),
                          Altitude = col_character(),
                          Region = col_character(),
                          Producer = col_character(),
                          Number.of.Bags = col_double(),
                          Bag.Weight = col_character(),
                          In.Country.Partner = col_character(),
                          Harvest.Year = col_character(),
                          Grading.Date = col_character(),
```

```
                Owner.1 = col_character(),
                Variety = col_character(),
                Processing.Method = col_character(),
                Fragrance...Aroma = col_double(),
                Flavor = col_double(),
                Aftertaste = col_double(),
                Salt...Acid = col_double(),
                Balance = col_double(),
                Uniform.Cup = col_double(),
                Clean.Cup = col_double(),
                Bitter...Sweet = col_double(),
                Cupper.Points = col_double(),
                Total.Cup.Points = col_double(),
                Moisture = col_double(),
                Category.One.Defects = col_double(),
                Quakers = col_double(),
                Color = col_character(),
                Category.Two.Defects = col_double(),
                Expiration = col_character(),
                Certification.Body = col_character(),
                Certification.Address = col_character(),
                Certification.Contact = col_character(),
                unit_of_measurement = col_character(),
                altitude_low_meters = col_double(),
                altitude_high_meters = col_double(),
                altitude_mean_meters = col_double()
              ))


raw_arabica <- raw_arabica %>% clean_names()
raw_robusta <- raw_robusta %>% clean_names()


## ----ratings, include = FALSE, warning=FALSE, echo=FALSE-------------------------
coffee_ratings <- bind_rows(raw_arabica,raw_robusta) %>%
  select(-x1) %>%
  select(total_cup_points, species, everything())



## ----dimension, include = FALSE--------------------------------------------------

dim(coffee_ratings)



## ----continentes, include = FALSE------------------------------------------------
load("../data/continentes.Rdata")
```

```
coffee_ratings <- coffee_ratings %>% rename(country = country_of_origin)


## ---- include = FALSE----------------------------------------------------------
coffee_ratings <- right_join(coffee_ratings, continentes)


## ---- include = FALSE----------------------------------------------------------
library(skimr)

glimpse(coffee_ratings)
skim(coffee_ratings)



## ----caracterÃ-sticas, include=FALSE--------------------------------------------
#seleccionamos el subconjunto
features <- coffee_ratings %>% select(total_cup_points,
                                      species,
                                      variety,
                                      country,
                                      continent,
                                      aroma: cupper_points)


## ---- include=FALSE-----------------------------------------------------------
#renombramos el subconjunto.
features <- features %>% rename(Points = total_cup_points,
                        Type = species,
                        Variety = variety,
                        Flavor = flavor,
                        Country = country,
                        Continent = continent,
                        Aroma = aroma,
                        Aftertaste = aftertaste,
                        Acidity = acidity,
                        Body = body,
                        Balance = balance,
                        Uniformity = uniformity,
                        Clean_cup = clean_cup,
                        Sweetness = sweetness,
                        Cupper_points = cupper_points)




## ----tabla1, fig.show='hold', echo=FALSE, message=FALSE, warning=TRUE----------

library(kableExtra)
```

```
features %>% head(10) %>% kbl(caption = "Tabla 1. Conjunto de Datos") %>% kable_classic_2()


## ----LIBRERIAS, include = FALSE----------------------------------------------------

library(skimr)
library(dlookr)
library(psych)
library(kableExtra)


## ----tabla2, fig.show='hold', echo=FALSE, message=FALSE, warning=TRUE------------

na<- features %>% is.na() %>% summary()
na




## ----prueba, include = TRUE--------------------------------------------------------

library(DataExplorer)

plot_bar(features) #diagrama de barras
plot_histogram(features) #Histograma de frecuencias




## ---- out.width = "50%", echo =FALSE----------------------------------------------
features %>% na.omit() %>% count(Country)%>%
  arrange(desc(n)) %>% head(15) %>%
  kbl(caption = " Tabla 2. PaÃ-s de procedencia de la muestra") %>%
  kable_classic_2()


## ---- out.width = "50%", echo =FALSE----------------------------------------------
features %>% count(Continent) %>%
  arrange(desc(n)) %>%
  head(5) %>%
  kbl(caption = " Tabla 3. Continente de procedencia de la muestra") %>%
  kable_classic_2()


## ---- out.width = "50%", echo =FALSE----------------------------------------------
features %>% ggplot() +
   geom_bar(aes(x=Continent, fill = Continent)) +
   labs(fill="Variedad de cafÃ©") + ylab("NÃºmero") +
```

```
  ggtitle("GrÃ¡fico de barras. Continentes")


## ---- out.width = "50%", echo =FALSE----------------------------------------------
features %>%
  count(Type) %>%
  na.omit() %>%
  head() %>% kbl(caption = "Tabla 4. Total de cafÃ© por tipos") %>%
  kable_classic_2()



## ----variedad, echo =FALSE---------------------------------------------------------
features %>% na.omit() %>%
  filter(Type == "Arabica")  %>%
  count(Variety) %>%
  arrange(desc(n)) %>%
  head(5) %>% kbl(caption = "Tabla 5. Variedad de cafÃ© Arabica") %>%
  kable_classic_2()


## ---- echo =FALSE-----------------------------------------------------------------
features  %>% filter(Type == "Robusta") %>%
  count(Variety) %>%
  arrange(desc(n)) %>%
  na.omit() %>%
  head(5) %>%
  kbl(caption = "Tabla 6. Variedad de cafÃ© Robusta") %>%
  kable_classic_2()


## ---- echo =FALSE-----------------------------------------------------------------
features %>%  filter (Variety %in% c ("Caturra", "Bourbon", "Typica", "Other")) %>%
  ggplot() + geom_bar(aes(x=Variety, fill = Variety)) + labs(fill="Variedad de cafÃ©") +ylab("NÃºmero")
+xlab("Variedades ArÃ¡bica")



## ----HISTOGRAMAS , echo=FALSE, fig.show="hold", out.width= "50%", echo = FALSE, warning= FALSE---
-
#Escala lineal
Points_graph <- features %>% na.omit() %>%  ggplot(aes(x = Points)) +
  geom_histogram(aes (y = ..density..)) +
  geom_density() + coord_cartesian(xlim = c(0,100)) + stat_bin(bins = 30)

Points_graph


## ---- echo=FALSE, fig.show="hold", out.width= "50%", echo = FALSE, warning= FALSE----
#Escala logaritmica
```

```
Points_graph_log<- features %>%  ggplot(aes(x= Points)) +
  geom_histogram(aes(y=..density..))+
  geom_density()  +
  scale_x_log10()
Points_graph_log



## ---- echo=FALSE, fig.show="hold", out.width= "50%", echo = FALSE, warning= FALSE----
#Escala lineal
Aroma_graph <- features %>% na.omit() %>%  ggplot(aes(x = Aroma)) +
  geom_histogram(aes (y = ..density..)) +
  geom_density() + coord_cartesian(xlim = c(0,10)) + stat_bin(bins = 30)
Aroma_graph



#Escala logaritmica
Aroma_graph_log<- features %>%  ggplot(aes(x= Aroma)) +
  geom_histogram(aes(y=..density..))+
  geom_density()  +
  scale_x_log10()
Aroma_graph_log

## ----include = TRUE,  echo=FALSE, fig.show="hold", out.width= "50%", echo = FALSE, warning= FALSE---
-
#Escala lineal
flav_graph <- features %>% na.omit() %>%  ggplot(aes(x = Flavor)) +
  geom_histogram(aes (y = ..density..)) +
  geom_density() + coord_cartesian(xlim = c(0,10)) + stat_bin(bins = 30)
flav_graph



#Escala logaritmica
flav_graph_log<- features %>%  ggplot(aes(x= Flavor)) +
  geom_histogram(aes(y=..density..))+
  geom_density()  +
  scale_x_log10()
flav_graph_log



## ----include = TRUE,  echo=FALSE, fig.show="hold", out.width= "50%", echo = FALSE, warning= FALSE---
-
#Escala lineal
Aftt_graph <- features %>% na.omit() %>%  ggplot(aes(x = Aftertaste)) +
  geom_histogram(aes (y = ..density..)) +
  geom_density() + coord_cartesian(xlim = c(0,10)) + stat_bin(bins = 30)
Aftt_graph



#Escala logaritmica
Aftt_graph_log<- features %>%  ggplot(aes(x= Aftertaste)) +
```

```
  geom_histogram(aes(y=..density..))+
  geom_density()  +
  scale_x_log10()
Aftt_graph_log


## ----include = TRUE,  echo=FALSE, fig.show="hold", out.width= "50%", echo = FALSE, warning= FALSE---
-
#Escala lineal
Ac_graph <- features %>% na.omit() %>%  ggplot(aes(x = Acidity)) +
  geom_histogram(aes (y = ..density..)) +
  geom_density() + coord_cartesian(xlim = c(0,10)) + stat_bin(bins = 30)
Ac_graph


#Escala logaritmica
Ac_graph_log<- features %>%  ggplot(aes(x= Acidity)) +
  geom_histogram(aes(y=..density..))+
  geom_density()  +
  scale_x_log10()
Ac_graph_log


## ----include = TRUE,  echo=FALSE, fig.show="hold", out.width= "50%", echo = FALSE, warning= FALSE---
-
#Escala lineal
body_graph <- features %>% na.omit() %>%  ggplot(aes(x = Body)) +
  geom_histogram(aes (y = ..density..)) +
  geom_density() + coord_cartesian(xlim = c(0,10)) + stat_bin(bins = 30)
body_graph


#Escala logaritmica
body_graph_log<- features %>%  ggplot(aes(x= Body)) +
  geom_histogram(aes(y=..density..))+
  geom_density()  +
  scale_x_log10()
body_graph_log


## ----include = TRUE,  echo=FALSE, fig.show="hold", out.width= "50%", echo = FALSE, warning= FALSE---
-
#Escala lineal
bal_graph <- features %>% na.omit() %>%  ggplot(aes(x = Balance)) +
  geom_histogram(aes (y = ..density..)) +
  geom_density() + coord_cartesian(xlim = c(0,10)) + stat_bin(bins = 30)
bal_graph


#Escala logaritmica
```

```
bal_graph_log<- features %>%  ggplot(aes(x= Balance)) +
  geom_histogram(aes(y=..density..))+
  geom_density()  +
  scale_x_log10()
bal_graph_log


## ----include = TRUE,  echo=FALSE, fig.show="hold", out.width= "50%", echo = FALSE, warning= FALSE---
-
#Escala lineal
CP_graph <- features %>% na.omit() %>%  ggplot(aes(x = Cupper_points)) +
  geom_histogram(aes (y = ..density..)) +
  geom_density() + coord_cartesian(xlim = c(0,10)) + stat_bin(bins = 30)
CP_graph

#Escala logaritmica
CP_graph_log<- features %>%  ggplot(aes(x= Cupper_points)) +
  geom_histogram(aes(y=..density..))+
  geom_density()  +
  scale_x_log10()
CP_graph_log


## ---- echo=FALSE, fig.show="hold", out.width= "40%", echo = FALSE, warning= FALSE----
features %>% ggplot(aes(x = Points)) +
  geom_density(mapping = aes(colour = Type)) + xlab("Puntos Obtenidos") + ylab("Densidad") + labs(
color = "Tipo") + ggtitle("DistribuciÃ³n de los Puntos obtenidos segÃºn el tipo de grano de cafÃ©")



## ---- echo=FALSE, fig.show="hold", out.width= "40%", echo = FALSE, warning= FALSE----
features %>%  filter (Variety %in% c("Caturra", "Bourbon", "Typica", "Other")) %>%  ggplot(aes(x =
Points)) +
  geom_density(mapping = aes(colour = Variety)) +
  ggtitle("DistribuciÃ³n de los Puntos obtenidos segÃºn la Variedad del cafÃ©")


## ---- echo=FALSE, fig.show="hold", out.width= "40%", echo = FALSE, warning= FALSE----
features %>%  filter (Continent %in% c( "Americas", "Africa", "Asia", "Europe", "Oceania")) %>%
ggplot(aes(x = Points)) +
  geom_density(mapping = aes(colour = Continent)) +
  ggtitle("DistribuciÃ³n de los Puntos obtenidos segÃºn el Continente de procedencia del grano")


## ---- echo=FALSE, fig.show="hold", out.width= "40%", echo = FALSE, warning= FALSE----
features %>%  filter (Country %in% c("Ethiopia", "Mexico", "Colombia", "Guatemala", "Brazil",
"Taiwan")) %>%  ggplot(aes(x = Points)) +
  geom_density(mapping = aes(colour = Country))+
  ggtitle("DistribuciÃ³n de los Puntos obtenidos segÃºn el PaÃ-s de procedencia del grano")
```

```
## ------------------------------------------------------------------------
summary(lm(data=features, Points ~ Continent))
summary(lm(data=features, Points ~ Type))



## ------------------------------------------------------------------------
#Para variedad
varieties <- features %>% filter (Variety %in% c("Caturra", "Bourbon", "Typica", "Other"))
summary(lm(data = varieties, Points ~ Variety))



## ------------------------------------------------------------------------
#Para paises
countries <- features %>%  filter (Country %in% c("Ethiopia", "Mexico", "Colombia", "Guatemala",
"Brazil", "Taiwan"))
summary(lm(data = countries, Points ~ Country))


## ----regresiÃ³n simple categoricas, include= TRUE----------------------------------
lm (data = features %>% na.omit(), Points ~ Continent) %>% broom::tidy() %>% kbl(caption =
"RegresiÃ³n Lineal Simple Continentes") %>% kable_classic()

lm(data = countries %>% na.omit(), Points ~ Country) %>% broom::tidy() %>% kbl(caption = "RegresiÃ³n
Lineal Simple Paises") %>% kable_classic()

lm(data = varieties, Points ~ Variety) %>% broom::tidy() %>% kbl(caption = "RegresiÃ³n Lineal Simple
Variedades") %>% kable_classic()


lm(data=features, Points ~ Type) %>% broom::tidy() %>% kbl(caption = "RegresiÃ³n Lineal Simple Tipos")
%>% kable_classic()



## ---- title = "Correlaciones entre variable cuantitativas"----------------------
features %>% plot_correlate()


## ---- warning=FALSE-------------------------------------------------------------
library(tidymodels) ## libreria necesaria para realizar las estimaciones
set.seed(569) ## FIJAMOS SEMILLA PARA OBTENER MISMOS RESULTADOS
features_1 <- features %>%
  filter (Country %in% c ("Ethiopia", "Mexico", "Colombia", "Guatemala", "Brazil", "Taiwan")) %>%
  filter(Variety %in% c("Caturra", "Bourbon", "Typica", "Other"))


## ------------------------------------------------------------------------
```

```
#Descartamos la avriable Type puesto que solo tiene un unico nivel
Train <- select(features_1,-Type) %>%
  mutate_if(is.character, factor) #se transforma todas los atributos que sean chr a factor


## ------------------------------------------------------------------------
#cargamos los registros de TEST
Test <- read.xlsx("../src/test.xlsx", sheetName = "Test") %>%
  filter (Country %in% c ("Ethiopia", "Mexico", "Colombia", "Guatemala", "Brazil", "Taiwan")) %>%
  filter(Variety %in% c("Caturra", "Bourbon", "Typica", "Other"))
Test


## ------------------------------------------------------------------------
#establecemos los parÃ¡metros de control
knn_ctrl <- trainControl(method = "repeatedcv", repeats = 2, classProbs = T)



## ------------------------------------------------------------------------
#Modelo KNN
set.seed(607) #semilla para valores aleatorios
knnFit <- train(Variety ~., data = Train,
                method = "knn",
                trControl=knn_ctrl,
                tuneLength = 15)

knnFit


## ---- include= TRUE----------------------------------------------------------
knnFit$results%>%
  kbl() %>%
  kable_classic_2()


## ------------------------------------------------------------------------
plot(knnFit, main = "MÃ©todo 'Elbow' para determinar el mejor k",
     xlab="Vecinos",
     ylab="PrecisiÃ³n")

## ------------------------------------------------------------------------
#descartamos la variable Type del Test
Test <- select(Test, -Type)


## ------------------------------------------------------------------------

predictions_knn <- predict(knnFit, Test) #calculamos la prediccIÃ³n
predictions_knn
```

```
## ---------------------------------------------------------------------------
CM_KNN <- confusionMatrix(predictions_knn, Test$Variety) #observamos la eficiencia del modelo
mediante la matrÃ-z de confusiÃ³n

## ---- include= TRUE----------------------------------------------------------
CM_KNN$overall %>%
  kbl() %>%
  kable_classic_2()


## ---- include= TRUE----------------------------------------------------------
CM_KNN$table %>%
  kbl() %>%
  kable_classic_2()


## ---------------------------------------------------------------------------
#Seguidamente podemos valernos de las mÃ©tricas de evaluaciÃ³n disponibles en caret, en este caso se
ha usado ROC

knnROC <- roc(Test$Variety, as.numeric(predictions_knn), levels = c("Caturra", "Bourbon"), direction =
">")
knnROC


## ---------------------------------------------------------------------------
#graficamos la curva
plot(knnROC, type="S", main="Curva ROC para KNN", xlab = "Sensitividad", ylab= "Especificidad")


## ---------------------------------------------------------------------------
rf_control <- trainControl(method = "repeatedcv",
                           number = 10,
                           repeats = 3,
                           search = "random")
set.seed(633)

rf_fit <- train(Variety~.,
                data=Train,
                method='rf',
                metric='Accuracy',
                tuneLength = 15,
                trControl=rf_control)
rf_fit


## ---------------------------------------------------------------------------
predictions_rf <- predict(rf_fit, Test)
```

```
## ------------------------------------------------------------------------
confusionMatrix(predictions_rf, Test$Variety)


rf_ROC <- roc(Test$Variety, as.numeric(predictions_rf), levels=c("Caturra", "Bourbon"), direction = ">")
rf_ROC


## ------------------------------------------------------------------------
plot(rf_ROC, type="S", main="Curva ROC para RF", xlab = "Sensitividad", ylab= "Especificidad")


## ------------------------------------------------------------------------
set.seed(659)
linear_fit <- train(Points ~ Aroma + Flavor + Aftertaste + Acidity + Body + Balance +
                      Uniformity + Clean_cup + Sweetness + Cupper_points,
                    data= Train,
                    method = "lm",
                    trControl =rf_control,
                    )

linear_fit


## ------------------------------------------------------------------------
Test_linear <- select(Test, -Variety, -Country, -Continent)


## ------------------------------------------------------------------------
prediction_linear <- predict(linear_fit, Test_linear)


## ------------------------------------------------------------------------
table(prediction_linear, Test_linear$Points)


## ------------------------------------------------------------------------
linear_ROC <- roc(Test_linear$Points, prediction_linear, direction = "<")
linear_ROC


## ------------------------------------------------------------------------
plot(prediction_linear, Test_linear$Points,
    main = "RegresiÃ³n lineal mÃºltiple para Puntos de CafÃ©",
    xlab = "Predicciones",
    ylab= "Reales", pch = 15, col = "black")+grid()
```

```
## ---------------------------------------------------------------------------
features_logit <- features %>%
  filter (Country %in% c ("Ethiopia", "Mexico", "Colombia", "Guatemala",
                    "Brazil", "Taiwan")) %>%
  filter(Variety %in% c("Caturra", "Bourbon"))


## ---------------------------------------------------------------------------
#nueva data de entrenamiento
Train_logit <- select(features_logit,-Type) %>% mutate_if(is.character, factor)


## ---------------------------------------------------------------------------
#Nueva data de test
Test_logit <- Test %>%
  filter (Country %in% c ("Ethiopia", "Mexico", "Colombia", "Guatemala",
                    "Brazil", "Taiwan")) %>%
  filter(Variety %in% c("Caturra", "Bourbon"))
Test_logit


## ---- warning=FALSE-----------------------------------------------------------
#funciÃ³n de control
logit_control <- trainControl(method = "repeatedcv",
                        number = 10,
                        repeats = 3,
                        search = "random")


set.seed(692)

#modelo de entrenamiento
logit_fit <- train(Variety ~.,
                   data=Train_logit,
                   method='glm',
                   family = "binomial",
                   metric='Accuracy',
                   tuneLength = 15,
                   trControl=logit_control)
logit_fit


## ---------------------------------------------------------------------------
predictions_logit <- predict(logit_fit, Test_logit)
predictions_logit

## ---------------------------------------------------------------------------
confusionMatrix(predictions_logit, Test_logit$Variety)
```

```
logit_ROC <- roc(Test_logit$Variety, as.numeric(predictions_logit),
            levels=c("Caturra", "Bourbon"), direction = ">")
logit_ROC


## ------------------------------------------------------------------------
plot(logit_ROC, type= "S", main = "ROC RegresiÃ³n LogÃ-stica")
```