Supervised learning final project

Juan Diego Fonseca
Ricardo Figueroa

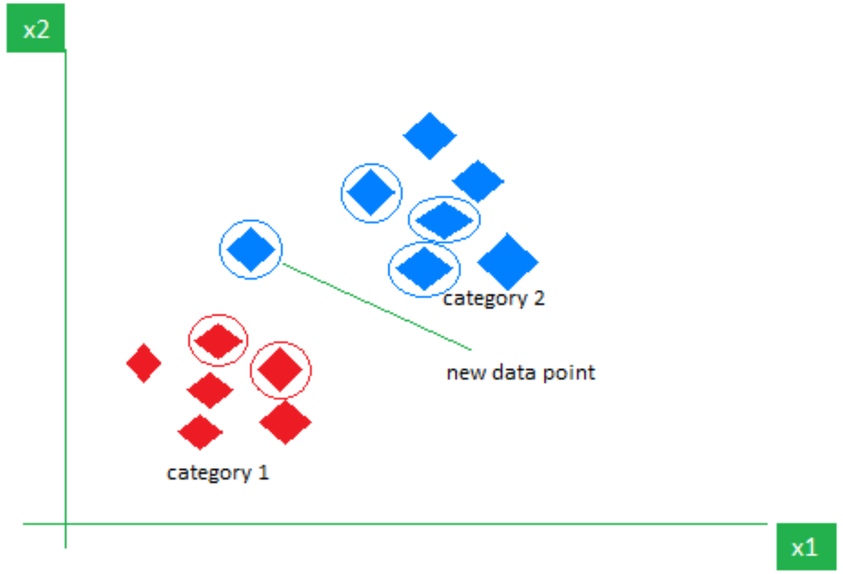
Table of contents

INTRODUCTION

Este proyecto consiste en desarrollar y aplicar modelos de aprendizaje automático supervisado a un conjunto de datos de expresiones de genes del cáncer, con el fin de clasificar con precisión distintos tipos de cáncer y predecir resultados clínicos.

Entrenar modelos sobre un conjunto de datos etiquetados en los que los niveles de expresión de los genes se utilizan como características y el tipo o estadio del cáncer como variable objetivo.

k-Nearest Neighbours Algorithm



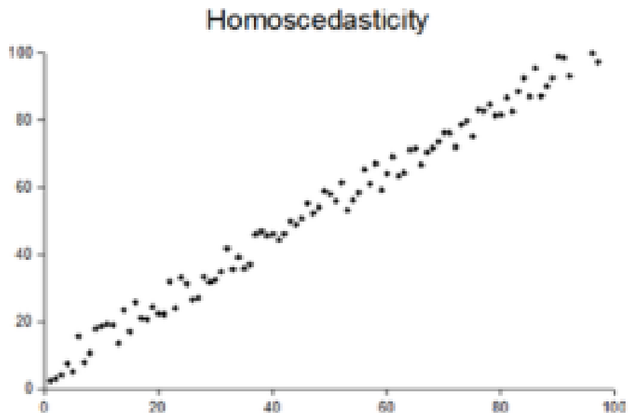
$$d(\mathbf{p}, \mathbf{q}) = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}$$

- El algoritmo K-Nearest Neighbors (KNN) es un método de aprendizaje automático robusto e intuitivo empleado para abordar problemas de clasificación y regresión. Aprovechando el concepto de similitud, KNN predice la etiqueta o el valor de un nuevo punto de datos teniendo en cuenta sus K vecinos más cercanos en el conjunto de datos de entrenamiento

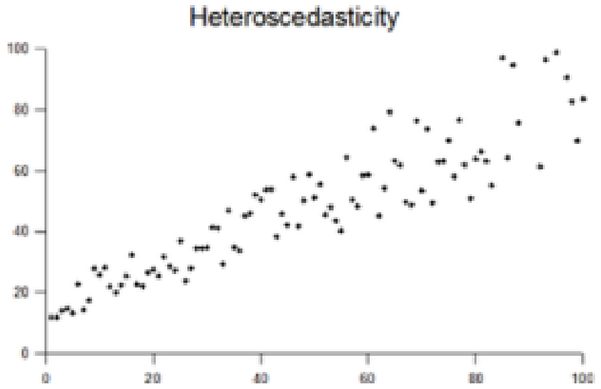
- Puede manejar datos numéricos y categóricos, lo que lo convierte en una opción flexible para varios tipos de conjuntos de datos en tareas de clasificación y regresión

- El algoritmo K-NN funciona encontrando los K vecinos más cercanos a un punto de datos dado basándose en una métrica de distancia, como la distancia euclídea

Linear and multilinear regression

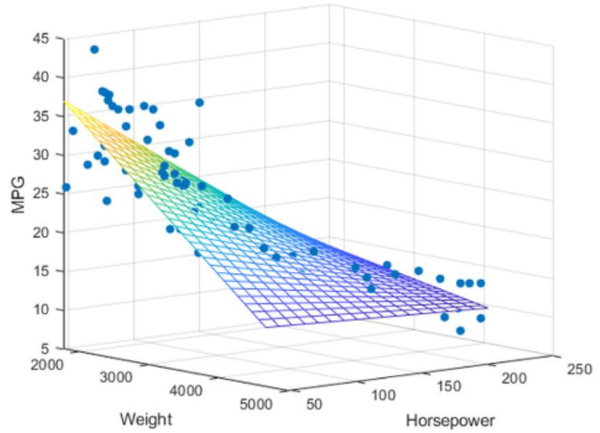


La regresión lineal es una técnica de **modelado estadístico** que se emplea para describir una variable de respuesta continua como **una función de una o varias variables** predictoras. Puede ayudar a comprender y predecir el comportamiento de sistemas complejos o a analizar datos experimentales, financieros y biológicos.



Linealidad. La respuesta puede modelizarse (razonablemente) como una combinación lineal de los predictores.
Homoscedasticity: (varianza constante), la varianza no depende de los predictores.
No hay multicolinealidad perfecta: No existe una relación lineal perfecta (no estocástica) entre los predictores.

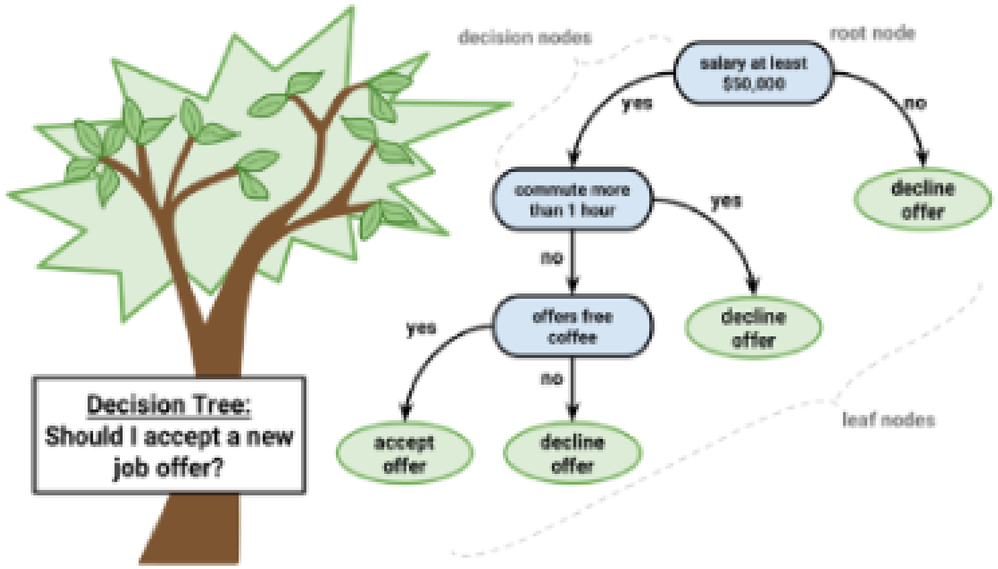
La multicolinealidad no reduce el poder predictivo ni la fiabilidad del modelo en su conjunto



Regresion Multivariable: modelos para varias variables de respuesta. Esta regresión tiene múltiples Yi que derivan de los mismos datos Y. Se expresan con fórmulas diferentes. Este es un ejemplo del sistema con

Decision trees

- Los aprendices de árbol de decisión son potentes clasificadores que utilizan una estructura de árbol para modelar las relaciones entre las características y los resultados potenciales

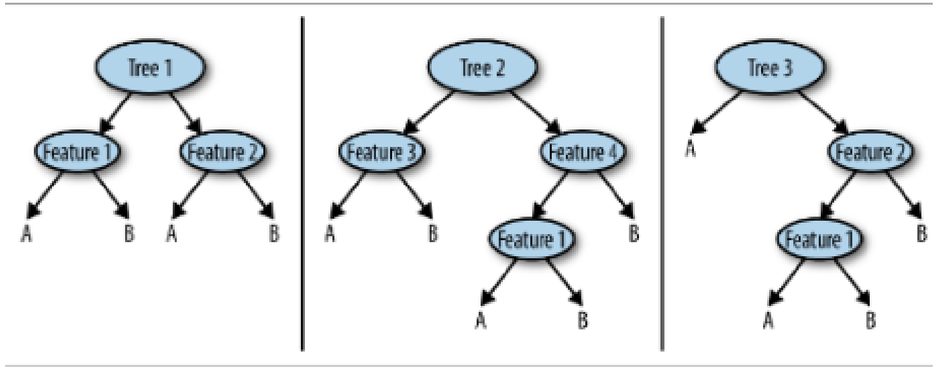


Se construyen:

- Heurística (Partición Recursiva)

Ventajas	Desventajas
• Un clasificador polivalente que funciona bien en muchos tipos de problemas • Proceso de aprendizaje muy automático, que puede manejar características numéricas o nominales, así como datos que faltan • Excluye las características sin importancia • Puede utilizarse tanto en conjuntos de datos pequeños como grandes • El resultado es un modelo que puede interpretarse sin conocimientos matemáticos (para árboles relativamente pequeños) Más eficaz que otros modelos complejos	• Los modelos de árbol de decisión suelen estar sesgados hacia las divisiones de características con un gran número de niveles. • Es fácil sobreajustar o infraajustar el modelo • Puede tener problemas para modelar algunas relaciones debido a su dependencia de divisiones paralelas a los ejes. • Pequeños cambios en los datos de entrenamiento pueden provocar grandes cambios en la lógica de decisión • Los árboles grandes pueden ser difíciles de interpretar y las decisiones que toman pueden parecer contraintuitivas.

Random Forest



- Un método de conjunto, es decir que “pone junto” o combina resultados para obtener un superresultado final.

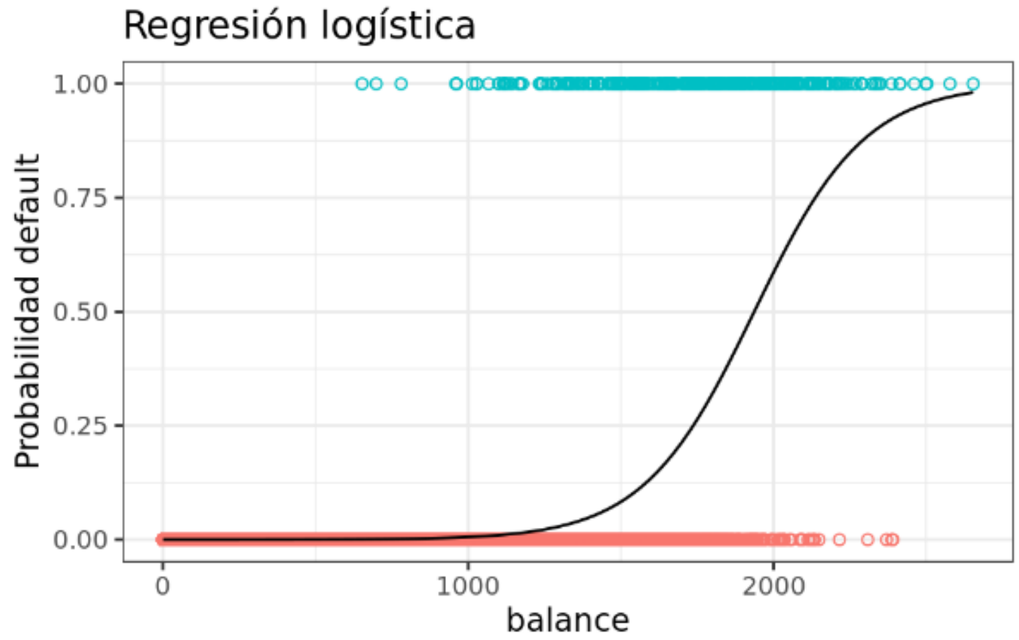
- Cada árbol se entrena en un subconjunto de la serie de datos y da un resultado (sí o no, en el caso de nuestro ejemplo de las setas). Posteriormente, se combinan los resultados de todos los árboles de decisión para dar una respuesta final. Cada árbol “vota” (sí o no) y la respuesta final es la que tenga la mayoría de votos.

Metodologia:

- Dividimos nuestra serie de datos en varios subconjuntos compuestos aleatoriamente de muestras, de ahí el “random” de random forest.
- Se entrena un modelo en cada subconjunto: habrá tantos modelos como subconjuntos.
- Se combinan todos los resultados de los modelos (con un sistema de voto, por ejemplo) lo que nos da un resultado final.

Logistic and multinomial regression

Los métodos de regresión suelen predecir valores numéricos (continuos) numéricos



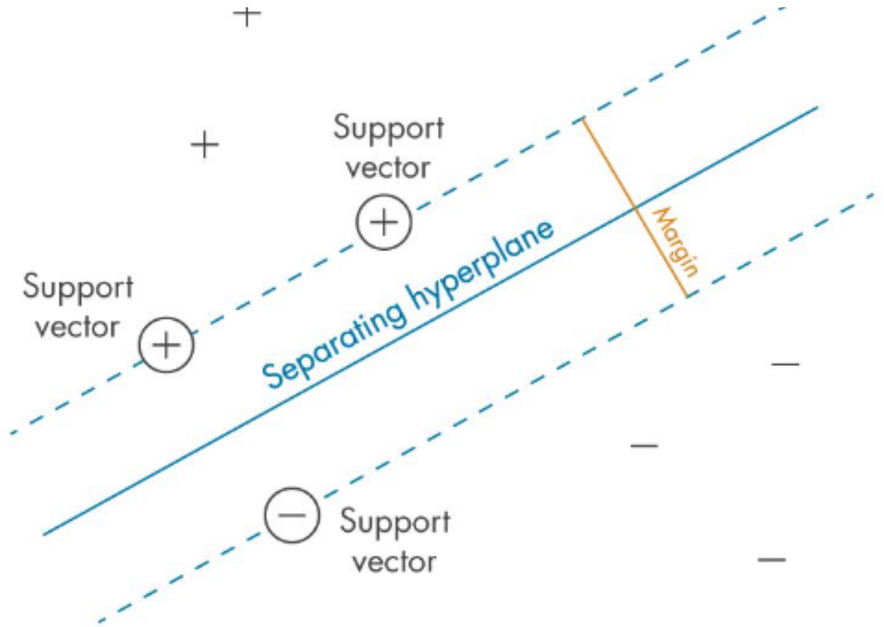
La regresión logística es una técnica de análisis de datos que utiliza las matemáticas para encontrar las relaciones entre **dos factores de datos**. Luego, utiliza esta relación para predecir el valor de uno de esos factores basándose en el otro. Normalmente, la predicción tiene un número finito de resultados, como un sí o un no.

La regresión logística (así como muchos otros tipos de algoritmos de clasificación) funciona con la función sigmoidea (0 y 1).

Wald chi-square: está muy expandido pero pierde precisión con tamaños muestrales pequeños.
Likelihood ratio: usa la diferencia entre la probabilidad de obtener los valores observados con el modelo logístico creado y las probabilidades de hacerlo con un modelo sin relación entre las variables. (estimacion de los parametros)

La regresión **logística múltiple** es una extensión de la regresión logística simple. Se basa en los mismos principios que la regresión logística simple (explicados anteriormente) pero **ampliando** el número de **predictores.** Los predictores pueden ser tanto continuos como categóricos.

Support Vector Classifiers – Naive Bayes

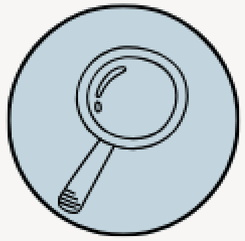


Es un algoritmo de aprendizaje supervisado que se utiliza en muchos problemas de **clasificación y regresión**, incluidas aplicaciones médicas para el procesamiento de señales, el procesamiento del lenguaje natural y el reconocimiento de imágenes y del habla.
El **objetivo** del algoritmo SVM es encontrar un **hiperplano** que separe de la mejor manera posible dos clases diferentes de puntos de datos. "El mejor camino posible" implica el hiperplano con el **margen** más amplio entre las dos clases, representado por los signos más y menos en la siguiente figura.

Naive Bayes es un modelo de predicción basado en la probabilidad bayesiana.

Un clasificador Naive Bayes **asume** que la **presencia o ausencia** de una característica concreta no está relacionada con la presencia o ausencia de cualquier otra característica, dada la clase variable Naive Bayes se utiliza normalmente cuando se dispone de una gran cantidad de datos.

Methodology



- Exploración de dataset
- Identificación de variables

- Preprocesamiento de los Datos
- Selección de los genes que más tenga probabilidad de detección de cancer

- Estratificaión de la data,
- Selección de 123 datos con tumor de cancer y 123 datos sin tumor de cancer

FINALIZACIÓN

- Comparación de los resultados de cada uno de los modelos aplicados y determinar el de mayor eficienca de predicción

- Aplicación de los modelos vistos en clase para determinar cual es mejor modelo de predicción de cancer en este dataset

Results

Para determinar el modelo mas adecuado para la predicción de genes de cancer en el dataset, se tuvo en cuenta variables como de precisión y Kappa, estas dos variables nos indicará cual es el modelo indicado

Models	Accuary	Kappa
K-NN	98.65 %	97.29 %
Random Forest	94 %	-
Soporte Vectorial	95.95 %	91.89 %
Regresión Lineal	95.85 %	-

Conclusion

- La precisión y los resultados detallados proporcionados por los modelos, especialmente en el caso del SVM con un accuracy del 95.95%, la sensibilidad del 92.31%, y la especificidad del 100%, sugieren un rendimiento impresionante en la tarea de clasificación. Sin embargo, la percepción de resultados demasiado exactos puede plantear inquietudes sobre la posibilidad de sobreajuste del modelo o errores en la implementación.
- Es crucial revisar aspectos específicos de la metodología, como la preparación de datos, la selección de variables, y la división entre conjuntos de entrenamiento y prueba, para identificar posibles fuentes de errores o sesgos. La consistencia en la preparación de datos, la estandarización y normalización, y la validación cruzada durante la búsqueda de hiperparámetros son prácticas fundamentales para evitar el sobreajuste y garantizar la generalización del modelo a nuevos datos.
- Además, se debe tener en cuenta la naturaleza del conjunto de datos y la interpretación biológica de los resultados. La elección de variables predictoras y la comprensión de su relevancia biológica son aspectos críticos en la construcción de modelos predictivos en el ámbito biomédico.
- En conclusión, mientras que los altos porcentajes de precisión son alentadores, es imperativo realizar una revisión exhaustiva del código y la metodología de modelado para garantizar la validez de los resultados. La interpretación biológica y la contextualización de los resultados son esenciales para validar la relevancia clínica y biológica de los modelos generados.

Q&A

Para que tipos de datos se usa el metodo de maquina vectorial y Random Forest?

- el metodo de maquina vectorial se utiliza comúnmente en problemas de clasificación, especialmente cuando el espacio de características es de alta dimensión. Es eficaz en conjuntos de datos pequeños a medianos y Random Forest es versátil y puede manejar conjuntos de datos grandes con muchas características. Se adapta bien a datos heterogéneos y es menos propenso a sobreajuste.

¿Cuándo se prefiere utilizar un modelo SVM en comparación con otros modelos de clasificación y qué tipo de variables es compatible?

- El modelo de Máquinas de Soporte Vectorial (SVM) es preferido cuando se busca una separación no lineal entre clases en un espacio multidimensional. Es compatible con variables cuantitativas y cualitativas, y es eficaz en problemas de clasificación binaria y multiclase.

¿En qué escenarios es adecuado emplear Random Forest y qué tipo de variables son apropiadas para este modelo?

- Random Forest es útil en problemas de clasificación y regresión, particularmente cuando hay interacciones no lineales y no se desea sobreajuste. Es compatible con variables cuantitativas y cualitativas, y tiene la capacidad de manejar conjuntos de datos grandes con muchas características.

Gracias

Referencias:

- https://es.mathworks.com/discovery/support-vector-machine.html#:~:text=Support%20vector%20machine%20(SVM)%20es,reconocimiento%20de%20imágenes%20y%20voz.
- https://fervilber.github.io/Aprendizaje-supervisado-en-R/ingenuo.html
- https://rpubs.com/Joaquin_AR/229736
- https://aws.amazon.com/es/what-is/logistic-regression/
- https://books.google.com.co/books?id=iNuSDwAAQBAJ&printsec=frontcover&redir_esc=y#v=onepage&q&f=false
- https://la.mathworks.com/discovery/linear-regression.html
- https://datascientest.com/es/random-forest-bosque-aleatorio-definicion-y-funcionamiento
- https://books.google.com.co/books?id=iNuSDwAAQBAJ&printsec=frontcover&redir_esc=y#v=onepage&q&f=false