The iris data set

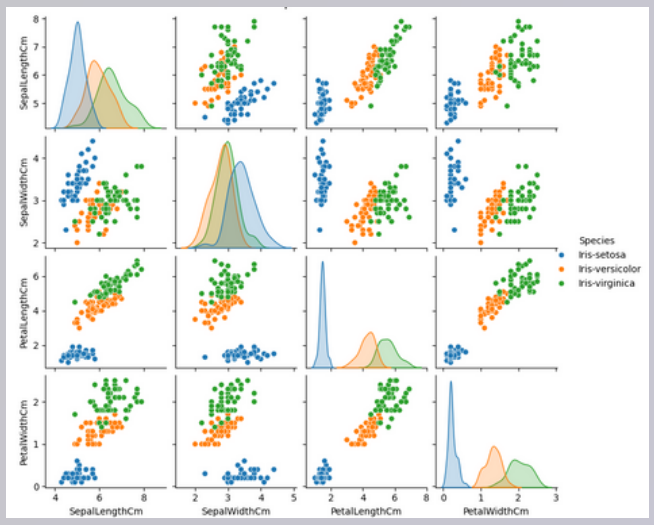
Martín Quijano - Martina Coletto

Vamos a analizar el dataset Iris utilizando técnicas de clasificación. Nuestro objetivo es predecir la especie de flor a partir de cuatro variables: largo y ancho del sépalo, y largo y ancho del pétalo. Para lograrlo, utilizamos dos modelos de aprendizaje supervisado: K-Nearest Neighbors y Regresión Logística.

Entendiendo los datos

Pairplot

Para observar la distribución y posibles relaciones entre las variables.



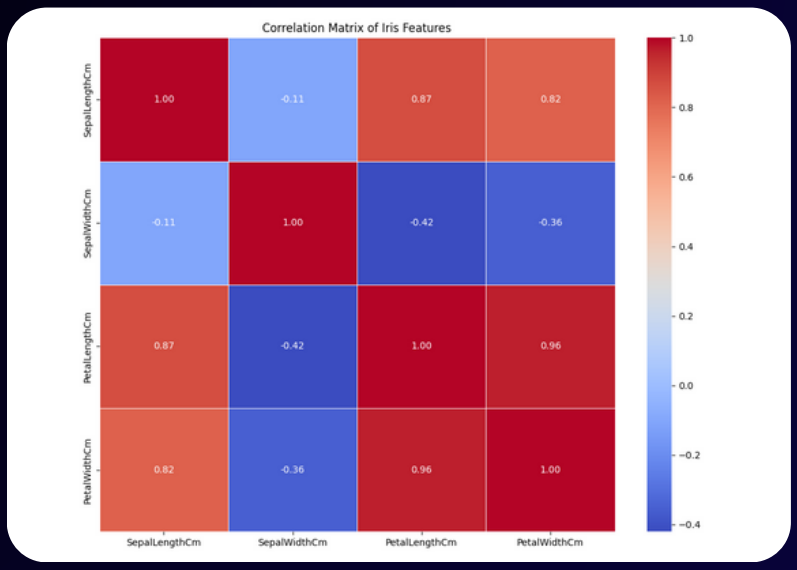
Con el pairplot vemos como se distribuye cada variable y como se relacionan entre si.

Vemos que en sepalLength y sepalWidth las variables tienen valores similares, hay superposicion, mientras que en petalLength y petalWidth hay menos superposicion, y vemos que setosa directamente no tiene superposicion con otra, mientras que versicolor y virginica se superponen en algunos puntos

Entendiendo los datos

Matriz de Correlación

Para identificar dependencias entre las variables.

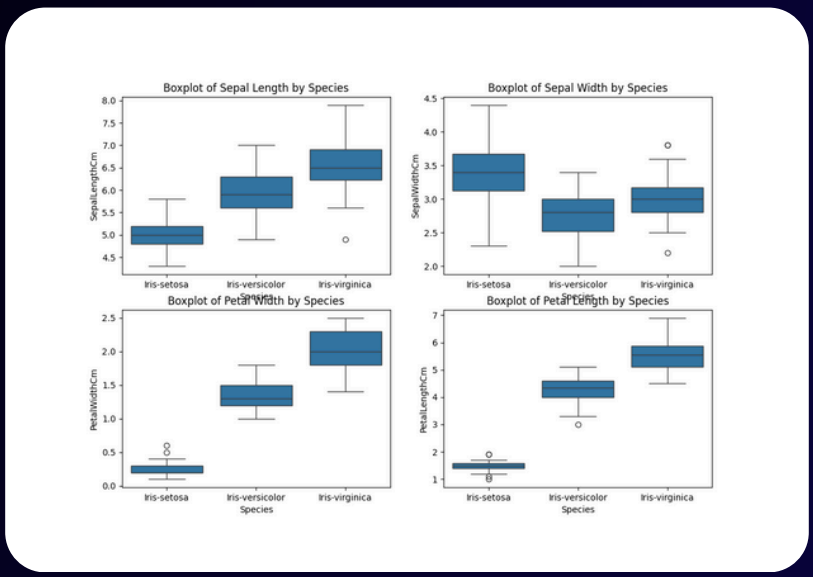


el largo del pétalo y el ancho del pétalo están altamente correlacionados, lo que nos indica que podrían aportar información redundante a los modelos. Sin embargo, esta correlación también puede ayudar a distinguir mejor entre especies

Entendiendo los datos

Boxplot

Para detectar outliers y diferencias en la distribución de las clases.



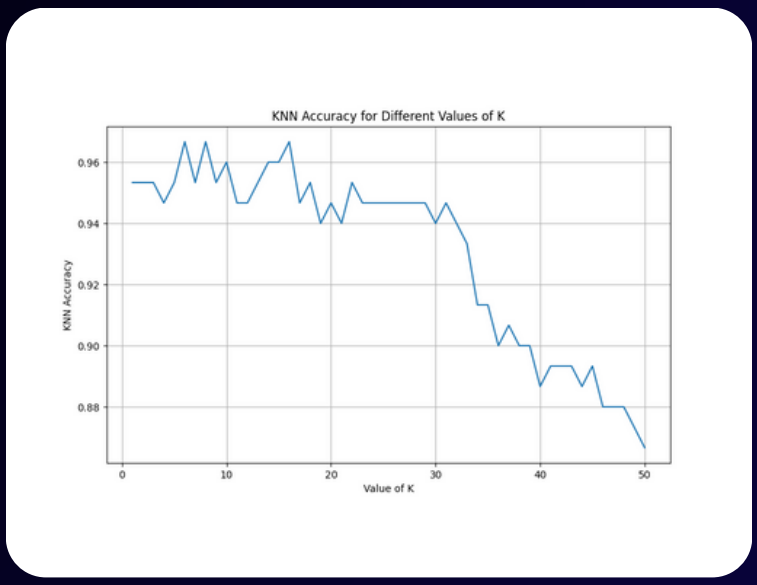
usamos boxplots para analizar la distribución de cada variable por especie. Esto nos permitió identificar posibles valores atípicos y diferencias significativas en la dispersión de las variables entre especies. Como se puede ver, la especie Setosa se diferencia claramente de las otras dos, mientras que Versicolor y Virginica tienen distribuciones más solapadas. Esto nos confirma lo que vimos en el pairplot

Aprendizaje supervisado

Análisis con KNN

Elección del mejor valor de K mediante cross validation.

Accuracy: 0.9667

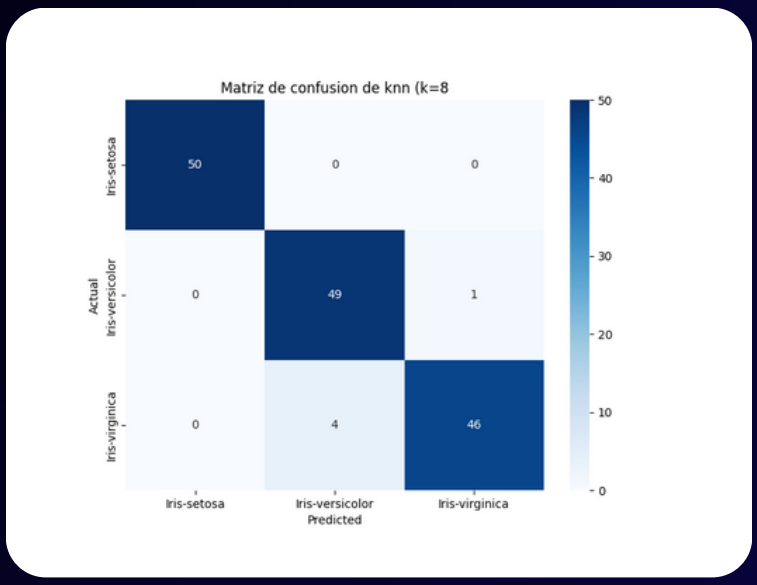


Encontramos que K=5 ofrecía la mayor accuracy. Decidimos usar cross validation y no dividir la muestra entre test y train ya que la muestra no es tan grande, y no queriamos caer en un overfitting o underfitting

Aprendizaje supervisado

Análisis con KNN

Matriz de confusión

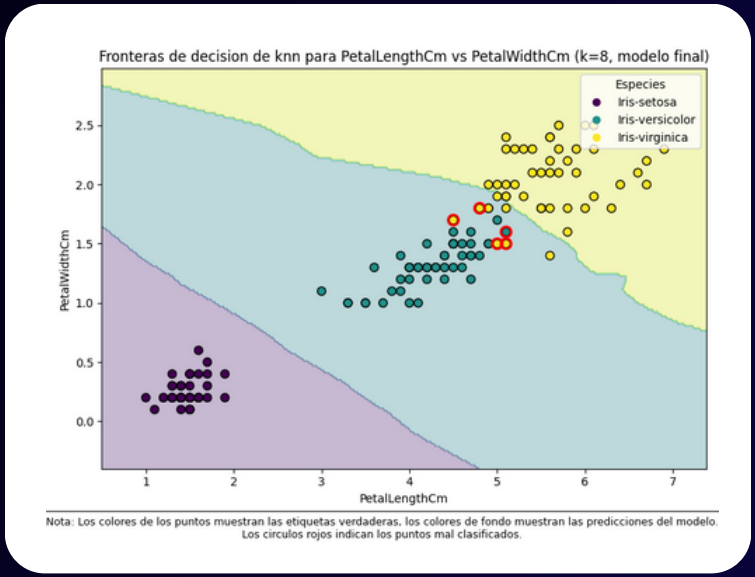


En la matriz de confusión podemos ver que la mayoría de las predicciones fueron correctas, con un accuracy del 96.67%. Esto indica que KNN funciona bien para este dataset. Adicionalmente, vemos que la setosa es clasificada bien siempre, mientras que en versicolor y virginica hay errores. Esto se debe a la superposicion de los valores de las variables de esas especies

Aprendizaje supervisado

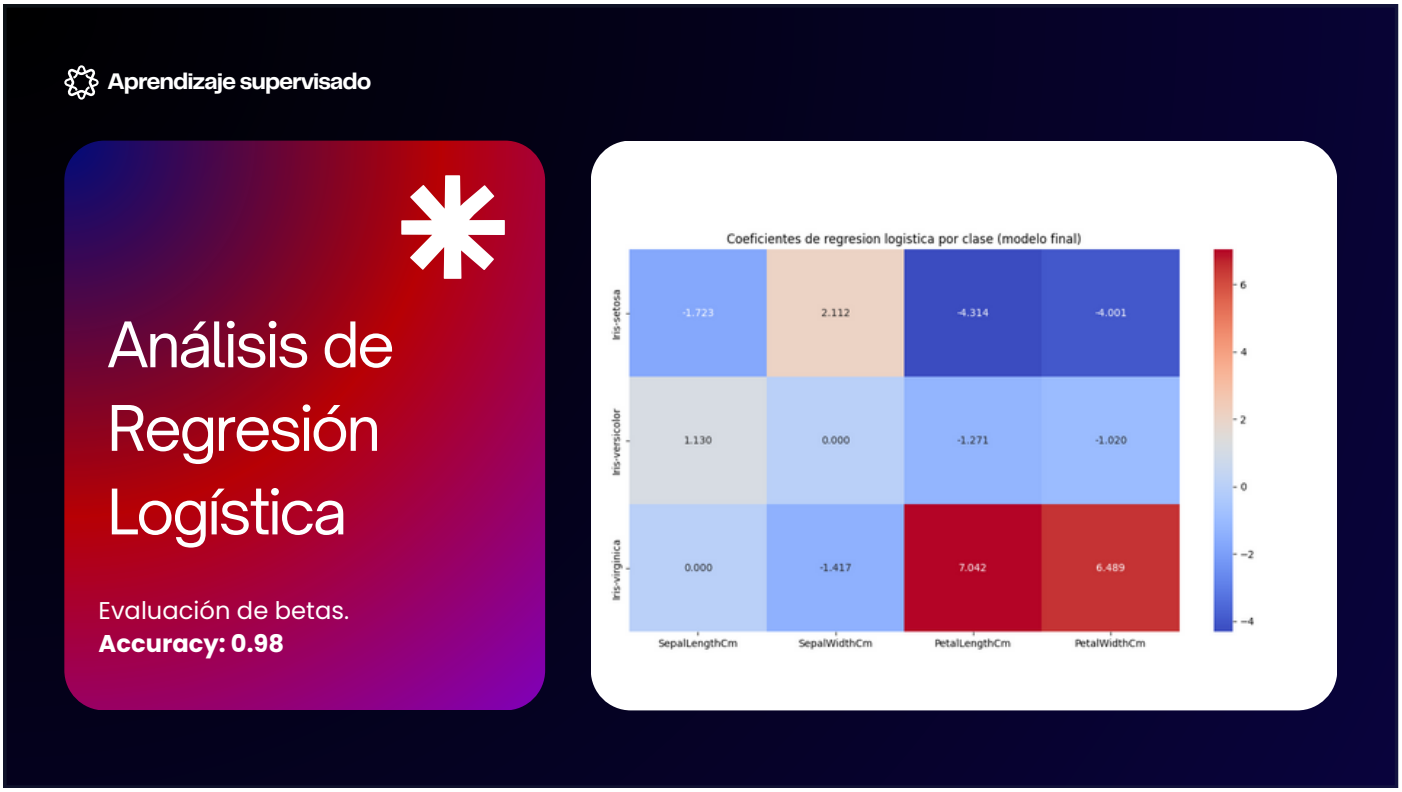
Análisis con KNN

Fronteras de decisión



podemos ver cómo KNN separa las diferentes clases en el espacio de las variables. Observamos que las especies Setosa se separan claramente, pero hay cierto solapamiento entre Versicolor y Virginica, lo que explica los pocos errores de clasificación.

Es importante decir tambien que hay puntos de un color que estan en el area de otro color. Esto se debe a que el grafico solo considera dos de las 4 variables utilizadas



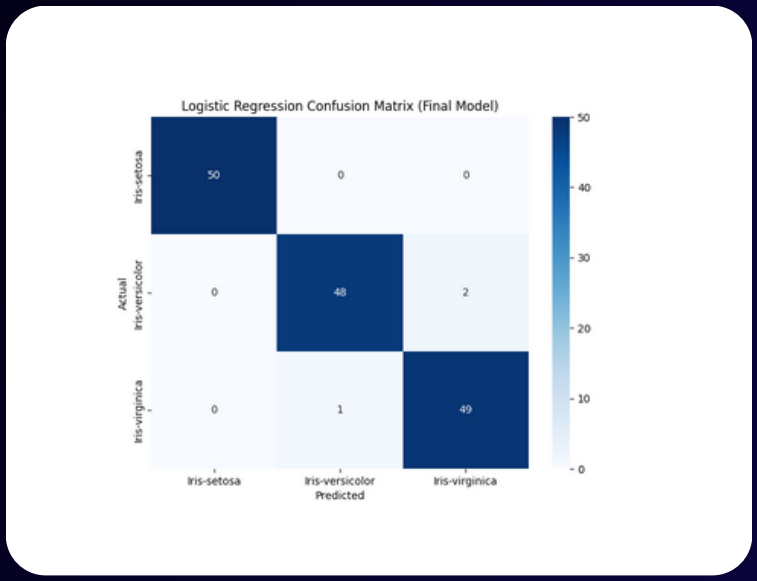
La accuracy obtenida con este modelo fue del 98%, ligeramente superior a la de KNN

Evaluamos los coeficientes beta y observamos dos principales caracteristicas. En primer lugar, las variables de petal tienen mas influencia (beta mas alto) que las variables de sepal. En segundo lugar, vemos que para setosa y para virginica los betas tienen valores bastante altos, mientras que para versicolor los betas no tienen valores muy altos. Esto explica que la mayoria de errores se den versicolor.

Aprendizaje supervisado

Análisis de Regresión Logística

Matriz de confusión

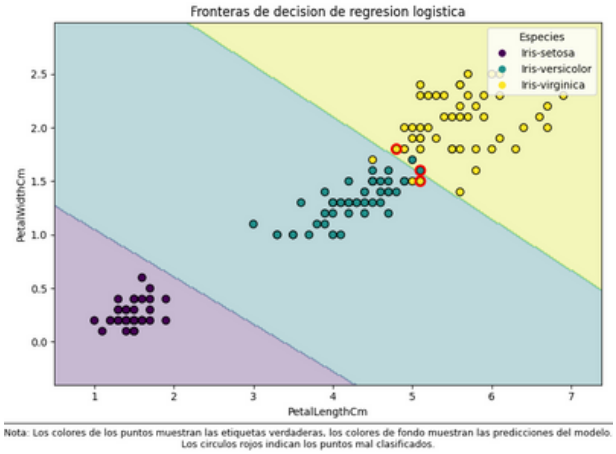


Menos errores que con KNN.

Aprendizaje supervisado

Análisis de Regresión Logística

Fronteras de decision



En comparación con KNN, este modelo tiende a generar límites más lineales, lo que puede ser una ventaja cuando las clases son bien separables. Sin embargo, al igual que en KNN, seguimos viendo cierto solapamiento entre Versicolor y Virginica.

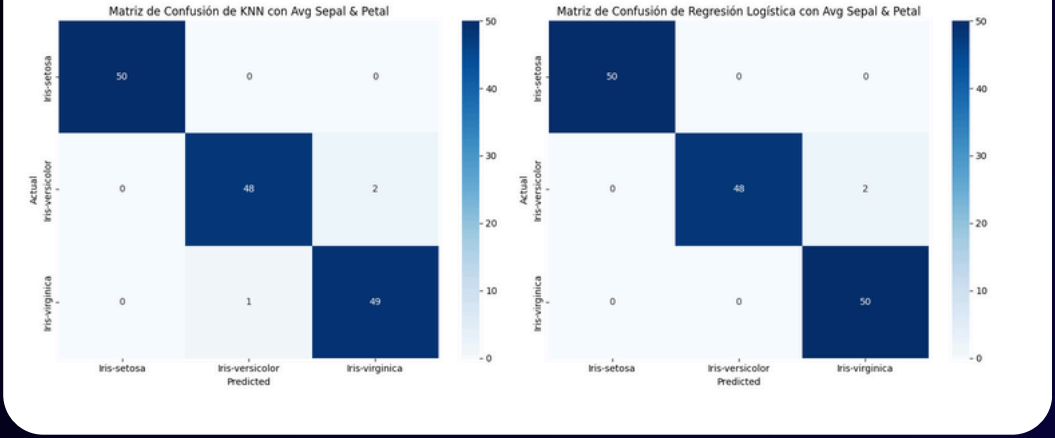
Para mejorar la precisión de nuestros modelos, probamos distintas transformaciones.

Esas transformaciones fueron minMax scaling, promediar sepalLength con sepalWidth y promediar petalLength con petalWidth, calcular ratios de sepal, petal, largo y ancho, poner en escala logaritmica las variables y promediar sepalHeight con petalHeight y sepalWidth con petalHeight.

La mejor transformacion fue MinMaxScaler. Notamos que el accuracy en ambos modelos mejora un 0.0067 al ser evaluados con los datos transformados

Transformaciones y mejora del modelo

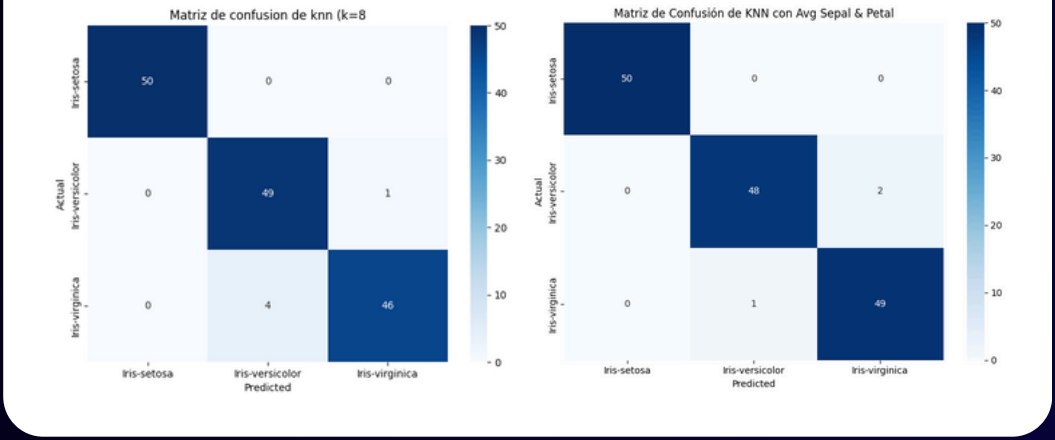
Matriz de confusión de los modelos con mejor transformación



Vemos que con los datos transformados en knn hay 3 errores y en regresion solo 2

Transformaciones y mejora del modelo

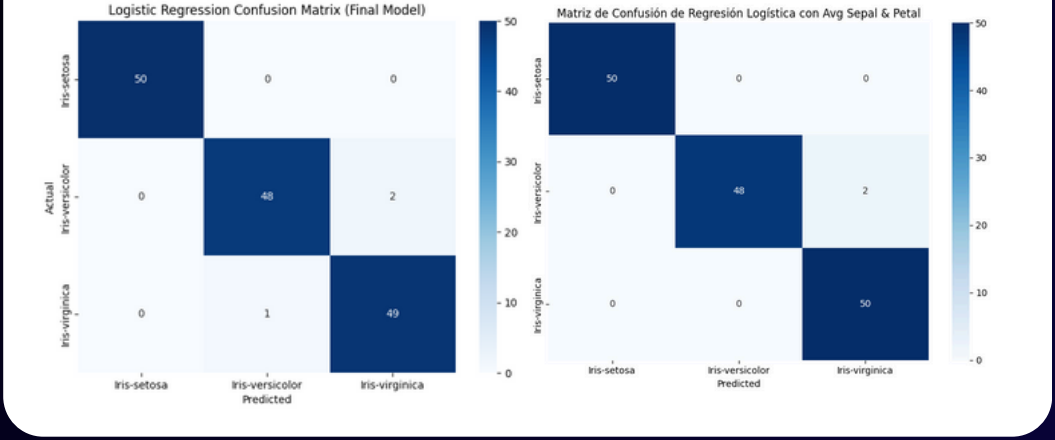
KNN: normal vs. transformación



Vemos que con la transformacion hay menos errores, pasamos de 5 errores a 3 errores en KNN

Transformaciones y mejora del modelo

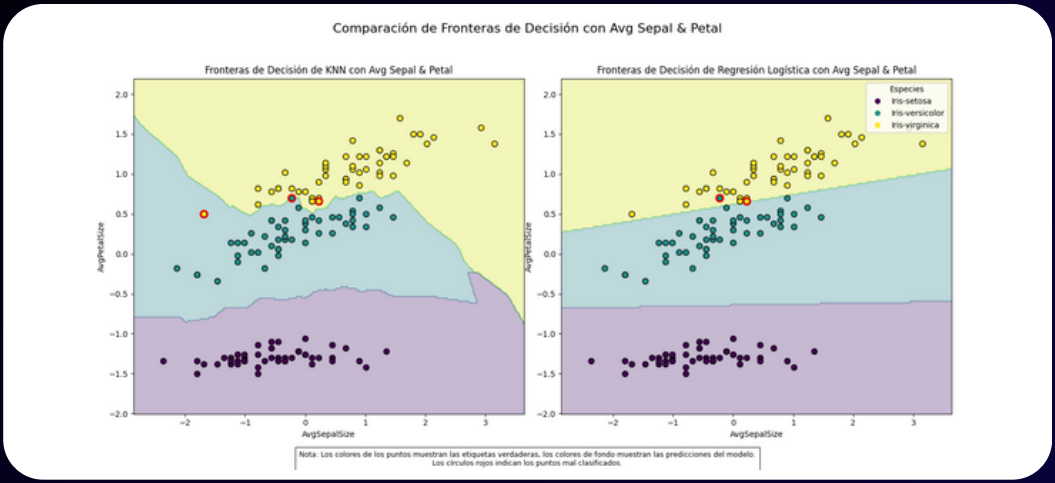
Regresión Logística: normal vs. transformación



Vemos que con la transformacion hay menos errores, pasamos de 3 errores a 2 errores en regresion logistica

Transformaciones y mejora del modelo

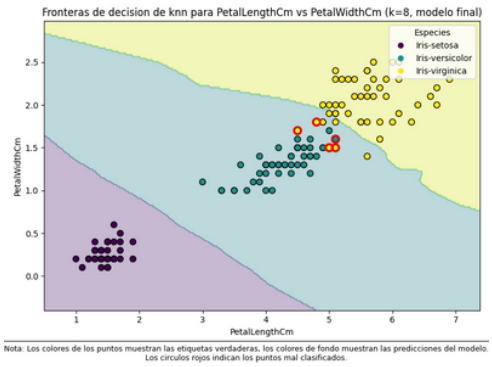
Fronteras de decisión de los modelos con mejor transformación

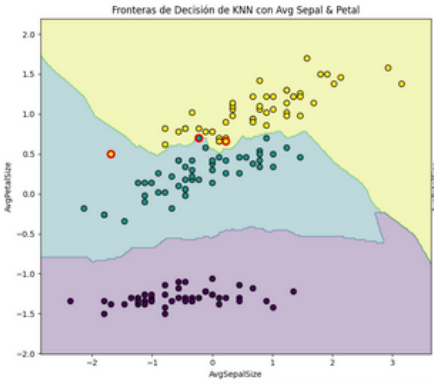


Al ver el grafico de fronteras de decision con datos transformados, vemos que en regresion logistica las fronteras son mas lineales que en knn

Transformaciones y mejora del modelo

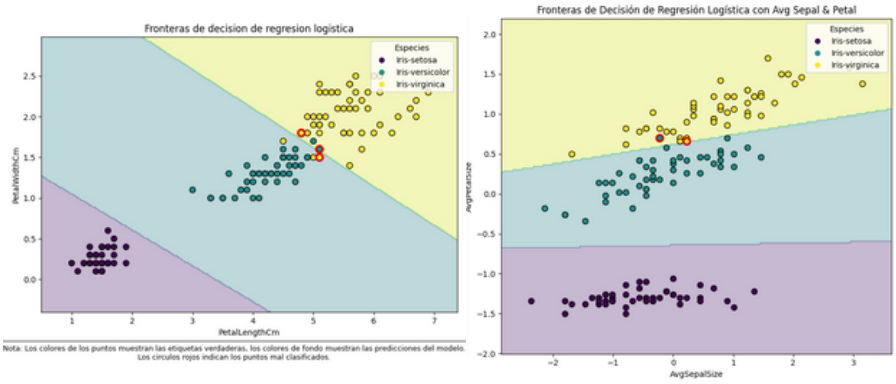
KNN: normal vs. transformación





Transformaciones y mejora del modelo

Regresión Logística: normal vs. transformación



Para concluir, podemos decir que los datos del dataset Iris son altamente clasificables, con un margen de error pequeño. Tanto KNN como Regresión Logística lograron buenos resultados, con una ligera ventaja para la Regresión Logística. Las transformaciones de datos ayudaron a mejorar el rendimiento de ambos modelos, especialmente en la separación de Versicolor y Virginica.
Como posibles mejoras, podríamos probar modelos más avanzados como Random Forest para analizar si obtenemos una clasificación aún más precisa

Gracias

Martín Quijano - Martina Coletto

-