Drug Classification

A Machine Learning Approach

By Praveen Myakala

Mar 03, 2024

Utilizing machine learning for personalized drug classification.

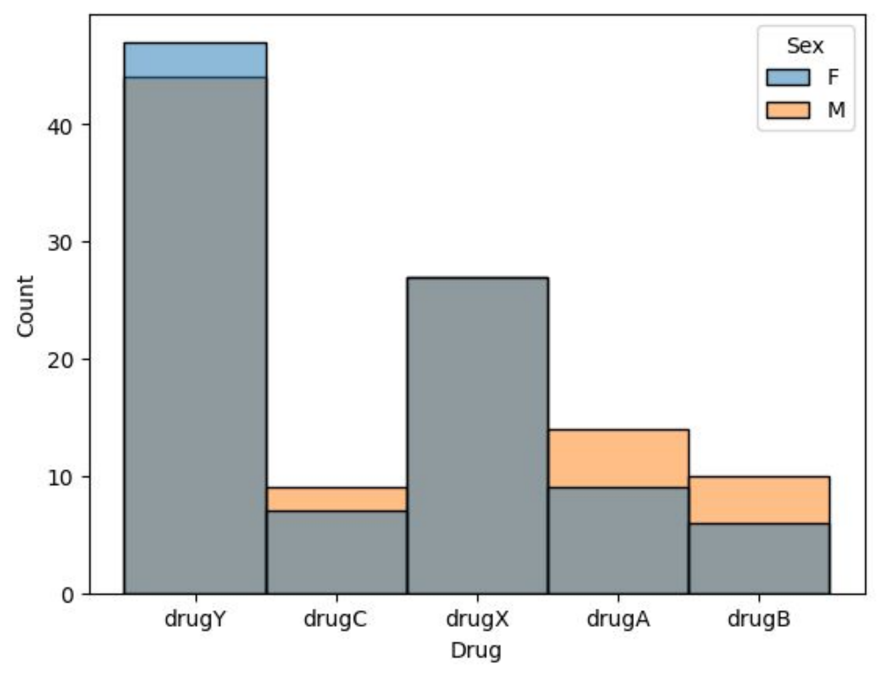
- Analyzed drug dataset

  Kaggle Dataset - https://www.kaggle.com/datasets/prathamtripathi/drug-classification

- Performed EDA and feature selection
- Compared Random forest, Logistic Regression machine learning models.

Sex & Drug

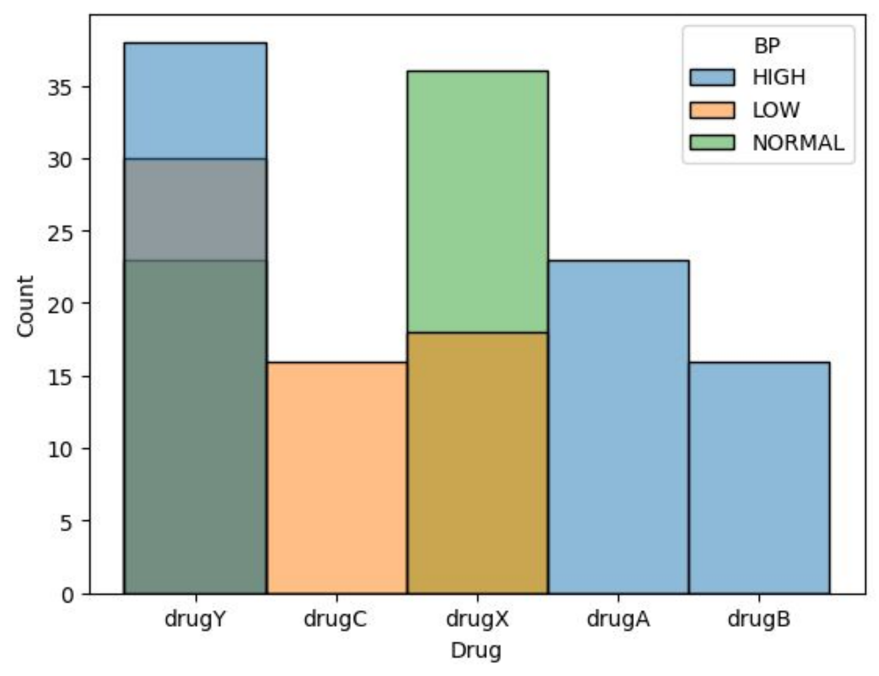
Histogram Plot



We notice that drugY majorly used by Female, and drugC, drugA, drugB is majorly used by Male

BP & Drug

Histogram Plot

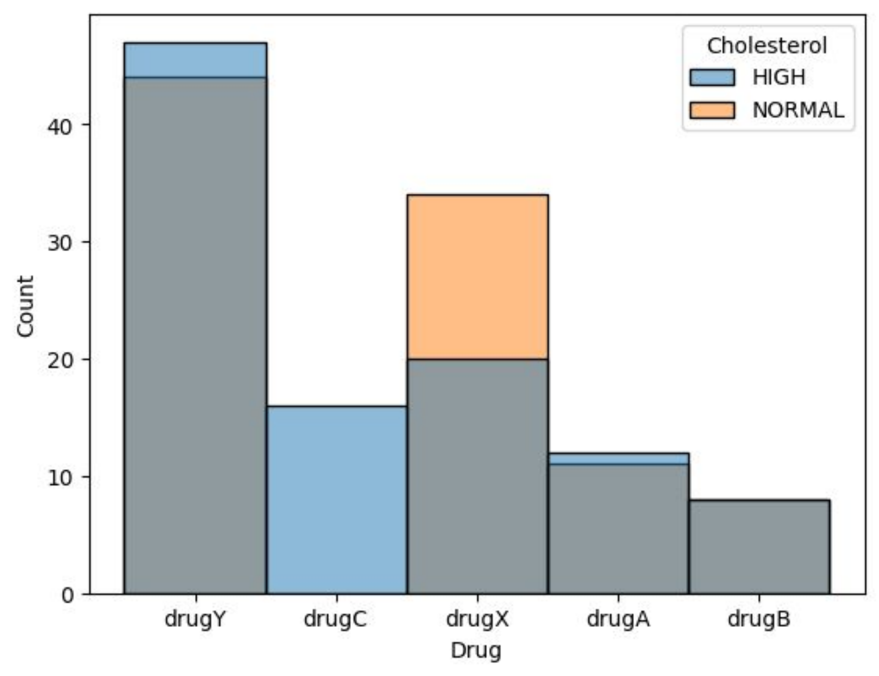


We notice that drugA and drugB are exclusively prescribed for patients with high blood pressure (BP), while drugC is specifically administered to patients with low blood pressure.

On the other hand, drugX appears to be distributed evenly among patients with both low and normal blood pressure.

Cholesterol & Drug

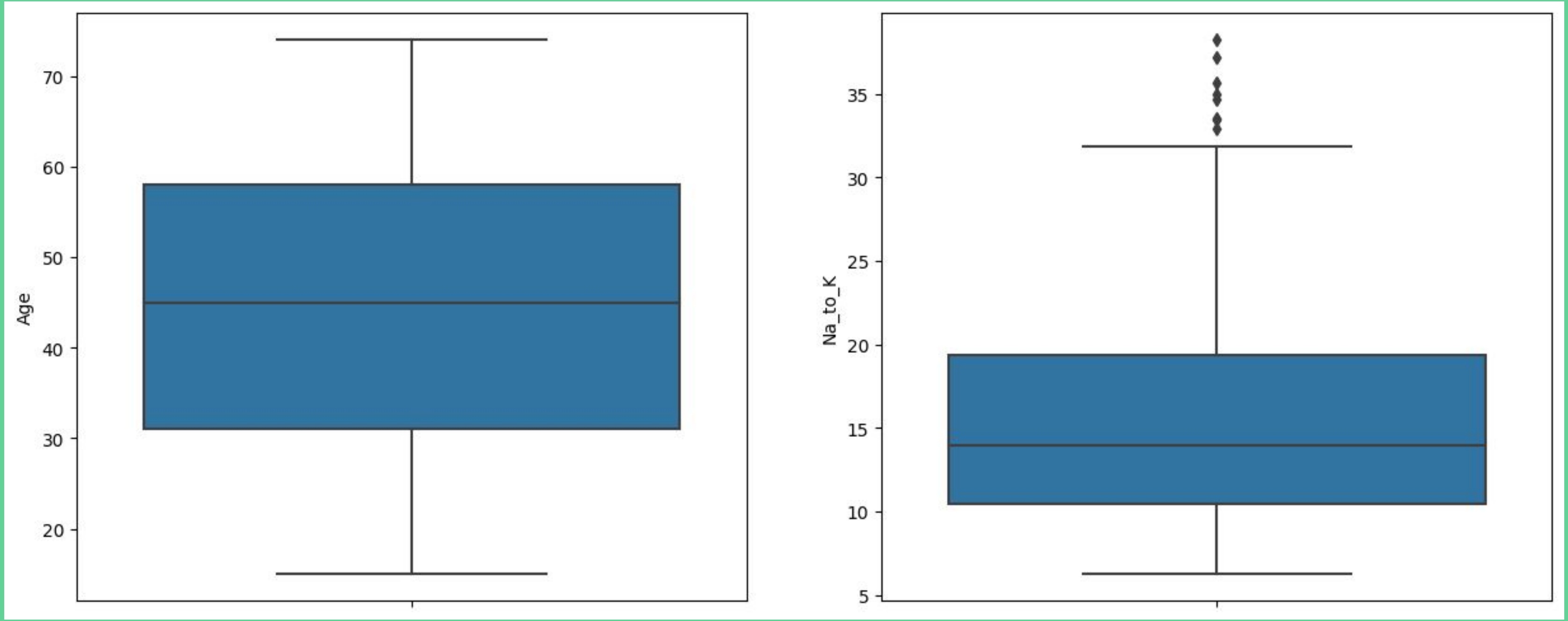
Histogram Plot



drugC was exclusively administered to patients with high cholesterol levels, whereas drugX was more frequently prescribed to patients with normal cholesterol levels.

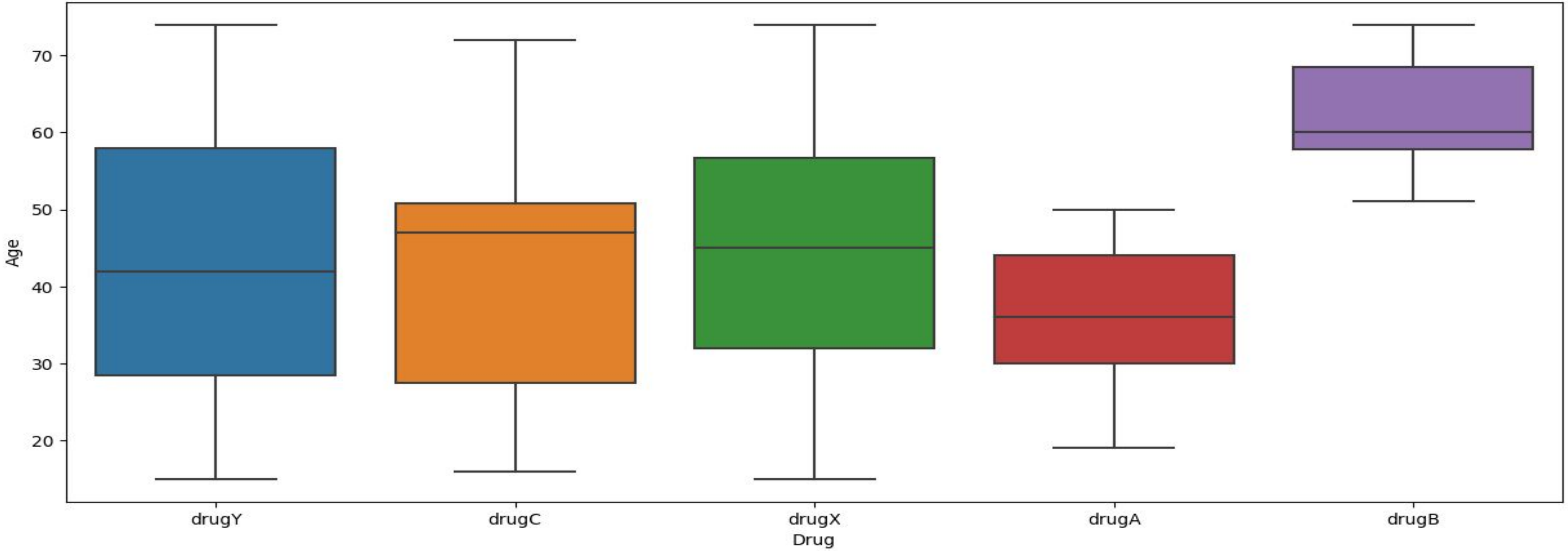
Numerical variables

A side-by-side boxplot using Seaborn boxplot() function compares **Age** and **Na_to_K** variables from the dataset Data.



Boxplot for Age & Drug

It is evident that **drugB** is prescribed to a patient group with a notably higher average age compared to other groups.



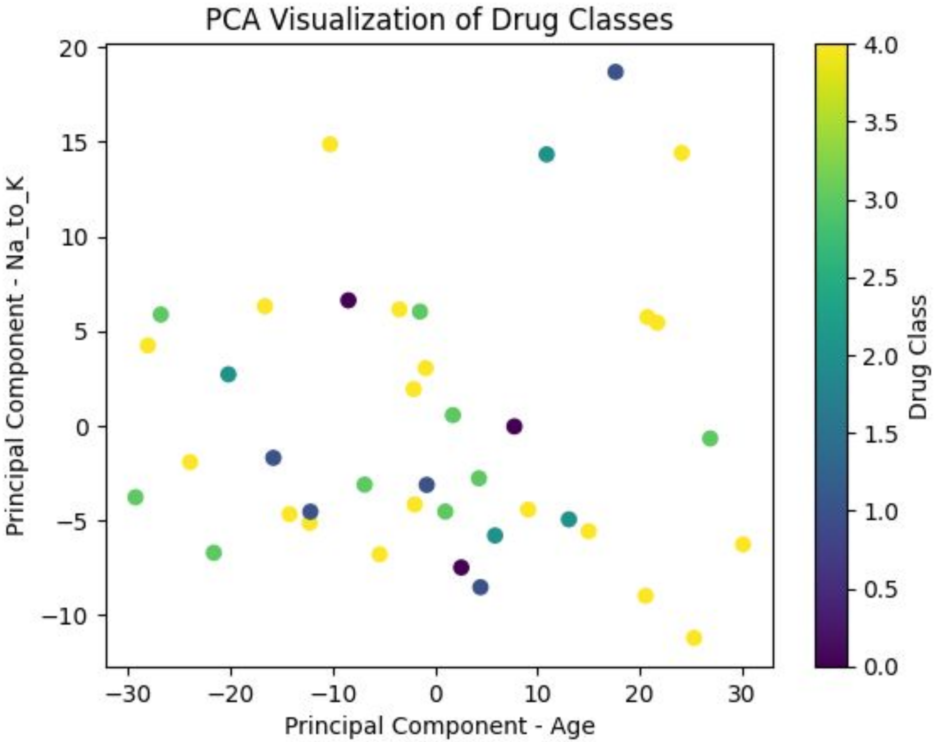
Feature selection

Selecting relevant features from the dataset improves model performance, reduces complexity, and mitigates overfitting by focusing on the most informative variables for predicting the target.

Classification

During classification scoring, I will annotate target labels as categorical, to achieve it I use **LabelEncoder** from scikit-learn to convert them into numerical values

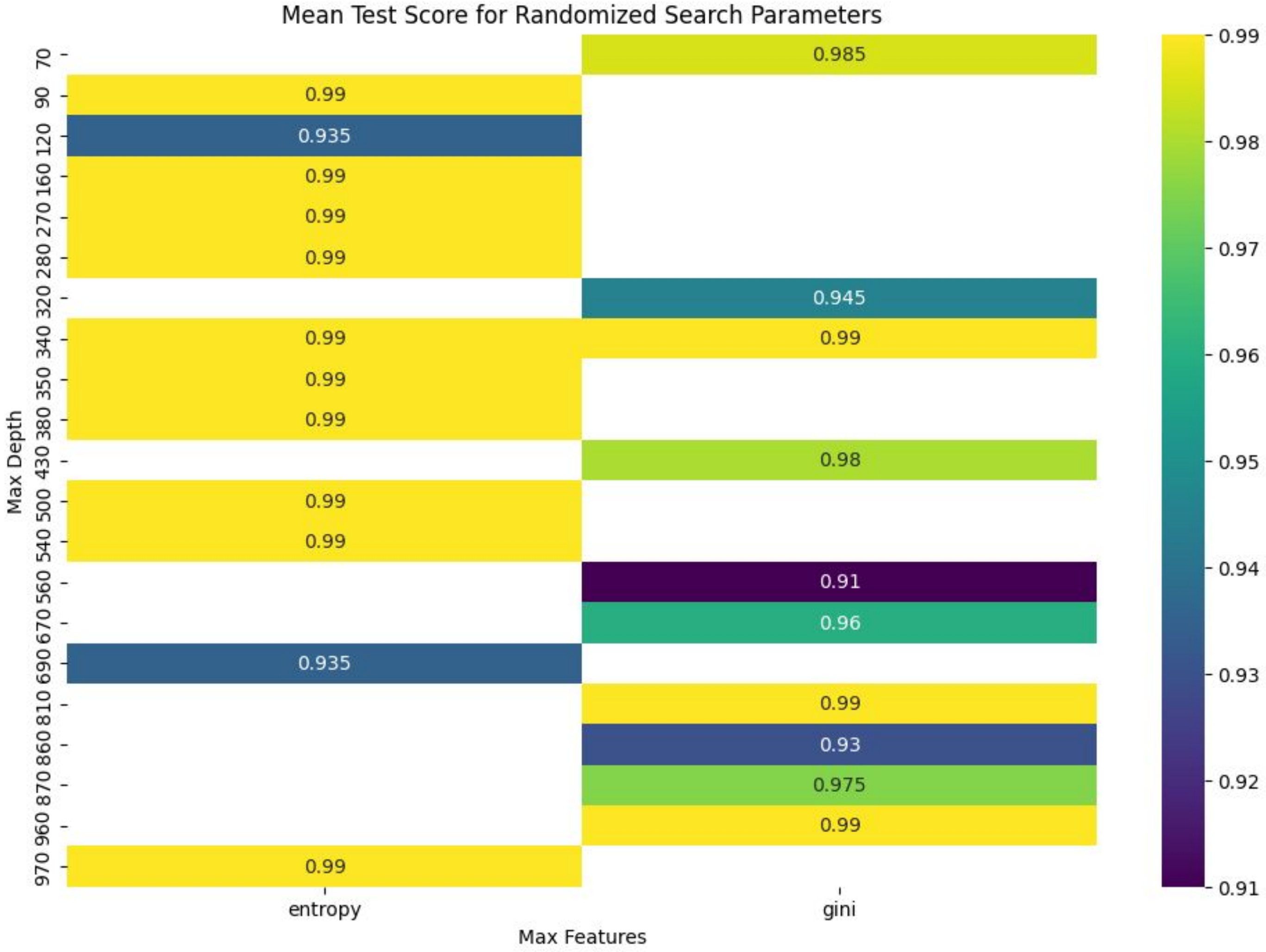
Drug Classes



PCA visualization simplifies the representation of drug classes in classification by reducing dimensionality while retaining essential data points.

KNN - Random forest

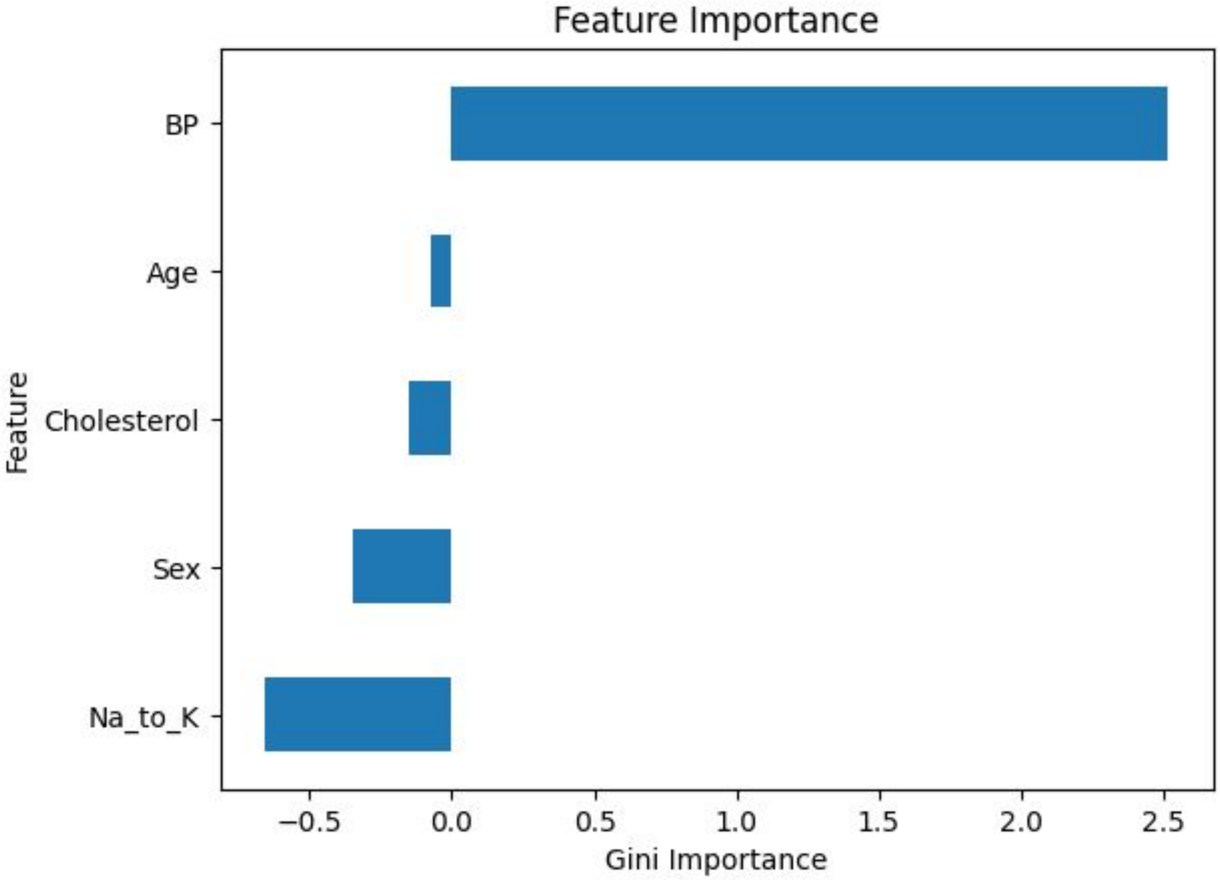
Optimize the hyperparameters of a RandomForestClassifier through RandomizedSearchCV, exploring a grid of parameter values with cross-validation to find the best-performing combination. Limiting parallelism to one worker aims to mitigate potential memory usage or segmentation faults errors during optimization.



Mean Test Score for Randomized Search Parameters

KNN - Logistic Regression

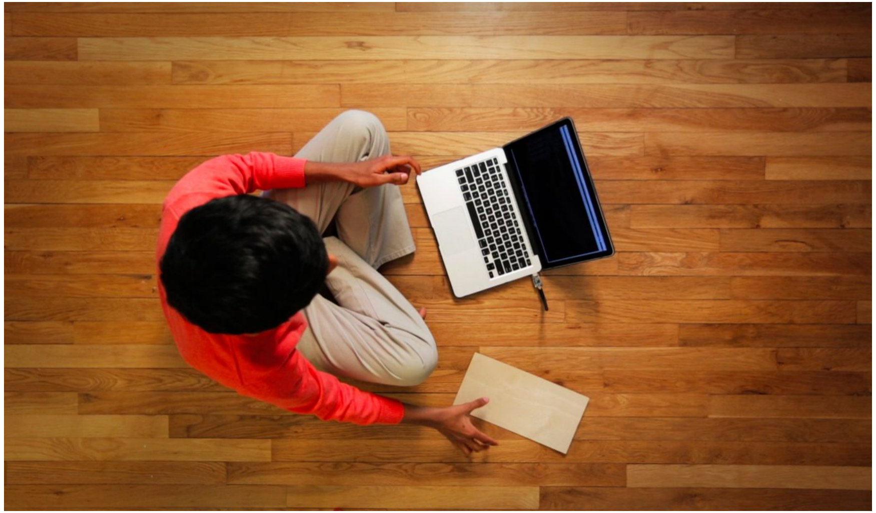
K-Nearest Neighbors (KNN) is a classification algorithm that predicts the class of a data point based on the majority class of its nearest neighbors in the feature space. In contrast, **Logistic Regression** is a linear model used for binary classification, estimating the probability of an outcome by fitting a logistic function to the input features.


Feature Importance
BP
Age
Cholesterol
Sex
Na_to_K
Feature
−0.5
0.0
0.5
1.0
1.5
2.0
2.5
Gini Importance

Aha!

My Learnings

- Exploratory Data Analysis (EDA) provided insights into dataset characteristics and patterns.
- Feature selection aided in identifying the most relevant attributes for drug classification.
- Comparison between Logistic Regression and Random Forest highlighted trade-offs in model complexity and performance.
- Consideration of factors like sex, age, bp and Cholesterol in drug classification for appropriate patients.



Thanks