Cognitive Analysis

MINI PROJECT REPORT

On

Drug prediction (logistic regression)

Submitted by

TV SahithiReddy(2021BCSE07AED091)

PV SadanandaReddy(2021BCSE07AED041)

In partial fulfillment of the award of the degree of

BACHELOR OF TECHNOLOGY

In

COMPUTER SCIENCE AND ENGINEERING

Under the Supervision of

Chetan j Shelke

Associate Professor

ALLIANCE COLLEGE OF ENGINEERING AND DESIGN

ALLIANCE UNIVERSITY, BENGALURU

CERTIFICATE

This is to certify that the mini project work entitled Drug prediction using logistic reggrestion is the bonafide work done by TV SahithiReddy (2021BCSE07AED091), PV SadanandaReddy (2021BCSE07AED41) submitted in partial fulfillment of the requirements for the award of the degree Bachelor of Technology in Computer Science and Engineering during the year 2023-2024

Guide:

Chetan j Shelke

Associate Professor

Abstract

Drug prediction plays a crucial role in healthcare, enabling personalized treatment plans and optimizing patient outcomes. Logistic regression, a fundamental machine learning algorithm, has been widely employed in predictive modeling due to its simplicity and interpretability. This study explores the application of logistic regression in drug prediction, aiming to predict the likelihood of a patient's positive response to a particular medication based on their demographic, clinical, and genetic characteristics. The dataset used in this study consists of anonymized patient data, including demographic information, medical history, genetic markers, and previous treatment outcomes. Feature engineering techniques are employed to preprocess the data and extract relevant features. Logistic regression models are trained on processed data to predict the probability of a positive drug response for each patient.

Evaluation of the logistic regression models is conducted using various performance metrics, including accuracy, precision, recall, and area under the receiver operating characteristic curve (AUC-ROC). Additionally, model interpretability is enhanced through feature importance analysis, identifying key predictors of drug response. The experimental results demonstrate the effectiveness of logistic regression in drug prediction tasks, achieving competitive performance compared to other machine learning algorithms. Furthermore, the interpretable nature of logistic regression facilitates clinical decision-making by providing insights into the factors influencing drug response.

Overall, this study highlights the potential of logistic regression as a valuable tool in drug prediction, offering healthcare practitioners a reliable method for personalized treatment recommendations and improving patient care. Further research may focus on refining predictive models by incorporating additional data sources and exploring advanced machine learning techniques to enhance prediction accuracy and interpretability.

Introduction

Drug prediction, a pivotal aspect of personalized medicine, aims to forecast the efficacy and potential adverse effects of pharmaceutical treatments for individual patients. With the rapid advancements in biomedical research and technology, there is a growing demand for accurate and efficient methods to predict drug responses based on patients' unique characteristics. Such predictions can guide healthcare practitioners in selecting the most appropriate medications tailored to each patient's specific needs, thereby maximizing therapeutic benefits while minimizing risks and adverse reactions.Traditional approaches to drug prediction often rely on clinical trials and population-based studies, which provide valuable insights into drug efficacy and safety across diverse patient populations. However, these methods have limitations, including high costs, lengthy timelines, and limited generalizability to individual patients. Moreover, they may not fully account for the complex interplay of genetic, environmental, and lifestyle factors that influence drug response variability among individuals.

In recent years, machine learning techniques have emerged as powerful tools for drug prediction, leveraging vast amounts of data to build predictive models that can analyze complex relationships and patterns in drug response. These models can integrate diverse sources of information, including genetic profiles, clinical data, biomarkers, and patient demographics, to generate personalized predictions of drug efficacy and toxicity.Among the various machine learning algorithms, logistic regression has gained prominence in drug prediction tasks due to its simplicity, interpretability, and robust performance in binary classification problems. By modeling the relationship between independent variables (e.g., patient characteristics) and a binary outcome (e.g., drug response), logistic regression can estimate the probability of a positive or negative response to a specific medication for a given patient.

This introduction sets the stage for exploring the application of logistic regression in drug prediction, emphasizing its potential to revolutionize clinical decision-making and improve patient outcomes in the era of precision medicine. Through comprehensive analysis and evaluation of logistic regression models, this study aims to elucidate the factors influencing drug response variability and enhance our understanding of personalized therapeutic interventions tailored to individual patient profiles.

Logistic regression:

It is a foundational statistical technique used for binary classification tasks, where the outcome variable is categorical and binary, typically representing two classes such as "yes" or "no", "true" or "false", or "positive" or "negative". Despite its name, logistic regression is a classification algorithm rather than a regression algorithm.

Here are some key points about logistic regression:

1. Model Representation: In logistic regression, the output of the linear combination of input features and their corresponding weights is passed through a logistic function (also known as the sigmoid function) to produce the predicted probability of belonging to the positive class.

2. Sigmoid Function: The sigmoid function maps any real-valued number to the range [0, 1]. It has an S-shaped curve and is defined as $\sigma(z) = \frac{1}{1 + e^{-z}}$, where z is the linear combination of input features and weights.

3. Decision Boundary: The decision boundary is the threshold value of the predicted probability (usually 0.5) above which the sample is classified as belonging to the positive class and below which it is classified as belonging to the negative class.

4. Training: Logistic regression models are trained using optimization algorithms such as gradient descent, which minimize a cost function (e.g., cross-entropy loss) to adjust the weights and biases iteratively until convergence.

5. Interpretability: Logistic regression models provide interpretable results, as the coefficients associated with each input feature indicate the impact of that feature on the predicted probability of the positive class. Positive coefficients indicate a positive relationship with the outcome, while negative coefficients indicate a negative relationship.

6. Assumptions: Logistic regression assumes that the relationship between the independent variables and the log odds of the dependent variable is linear. It also assumes that there is little or no multicollinearity among the independent variables and that the observations are independent of each other.

7. Regularization: To prevent overfitting, regularization techniques such as L1 (Lasso) and L2 (Ridge) regularization can be applied to logistic regression models.

Logistic regression is widely used in various fields, including healthcare (e.g., predicting disease risk), finance (e.g., credit scoring), marketing (e.g., customer churn prediction), and social sciences (e.g., predicting voting behavior). Its simplicity, interpretability, and effectiveness make it a popular choice for binary classification tasks.

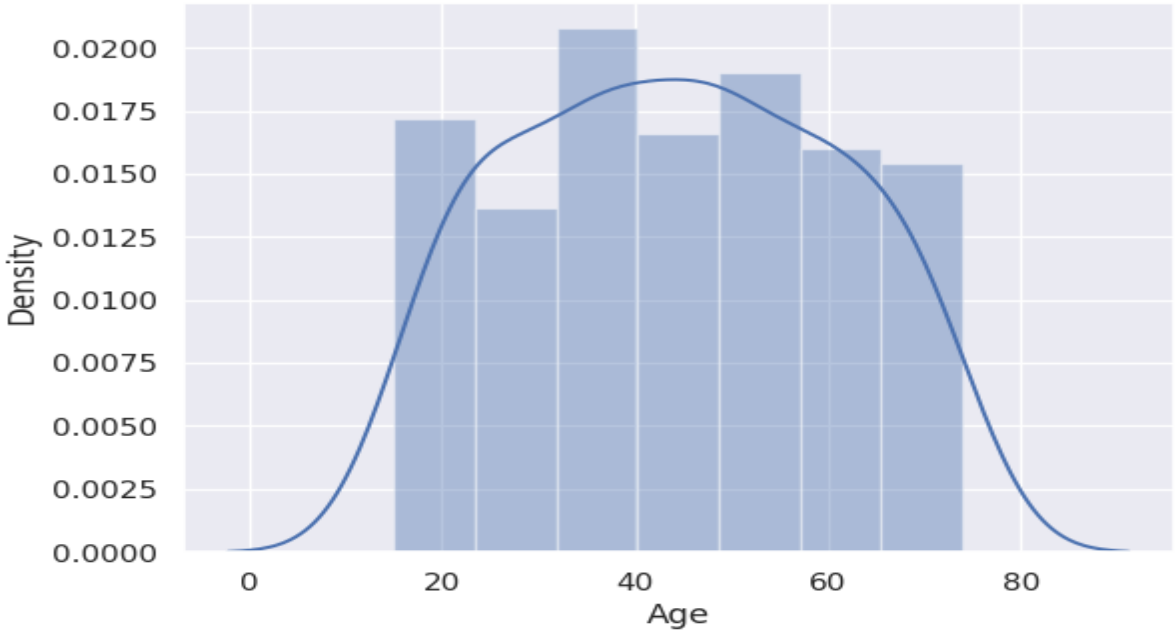
Code:

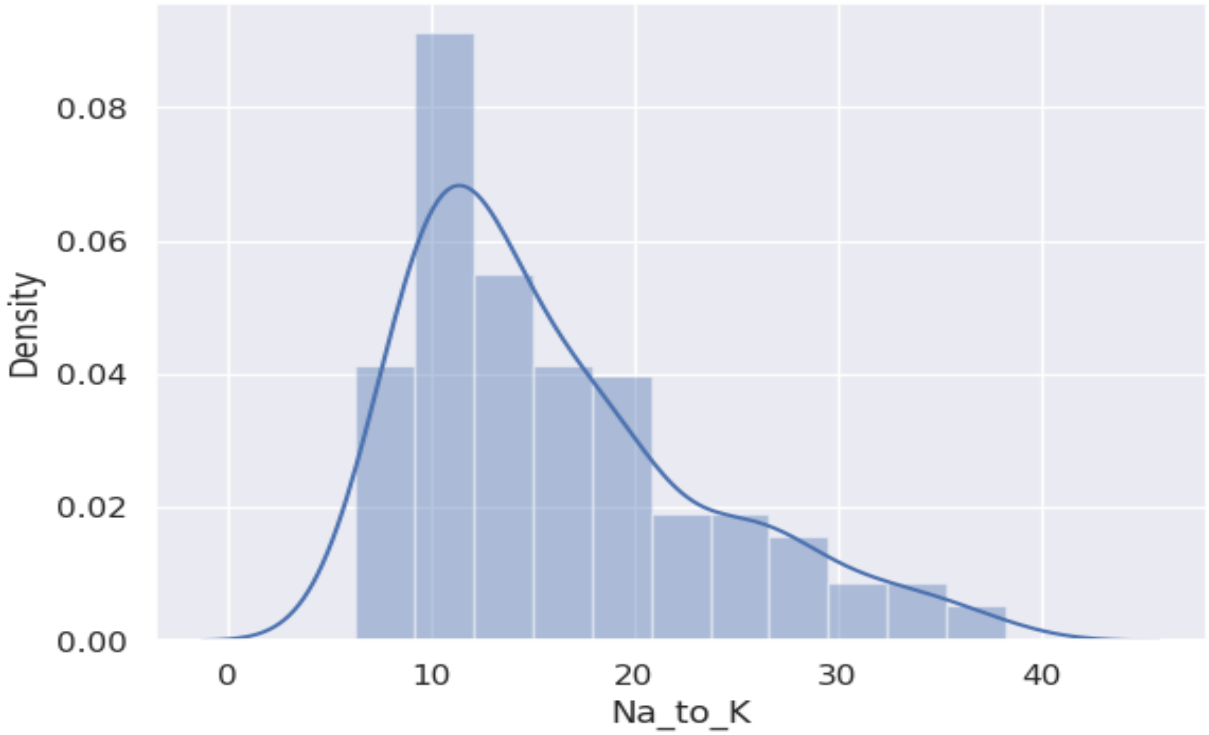
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import os
df_drug = pd.read_csv("/content/drug200.csv")
df_drug.head()
print(df_drug.info())
df_drug.Drug.value_counts()
df_drug.Sex.value_counts()
df_drug.BP.value_counts()
df_drug.Cholesterol.value_counts()
df_drug.describe()
skewAge = df_drug.Age.skew(axis = 0, skipna = True)
print('Age skewness: ', skewAge)
skewNatoK = df_drug.Na_to_K.skew(axis = 0, skipna = True)
print('Na to K skewness: ', skewNatoK)
Visulization:
sns.distplot(df_drug['Age']);
sns.distplot(df_drug['Na_to_K']);
sns.set_theme(style="darkgrid")
sns.countplot(y="Drug", data=df_drug, palette="flare")
plt.ylabel('Drug Type')
plt.xlabel('Total')
plt.show()
sns.set_theme(style="darkgrid")
sns.countplot(x="Sex", data=df_drug, palette="rocket")
plt.xlabel('Gender (F=Female, M=Male)')
plt.ylabel('Total')
plt.show()
sns.set_theme(style="darkgrid")
sns.countplot(y="BP", data=df_drug, palette="crest")
plt.ylabel('Blood Pressure')
plt.xlabel('Total')
plt.show()
sns.set_theme(style="darkgrid")
sns.countplot(x="Cholesterol", data=df_drug, palette="magma")
plt.xlabel('Blood Pressure')
plt.ylabel('Total')
plt.show()
```

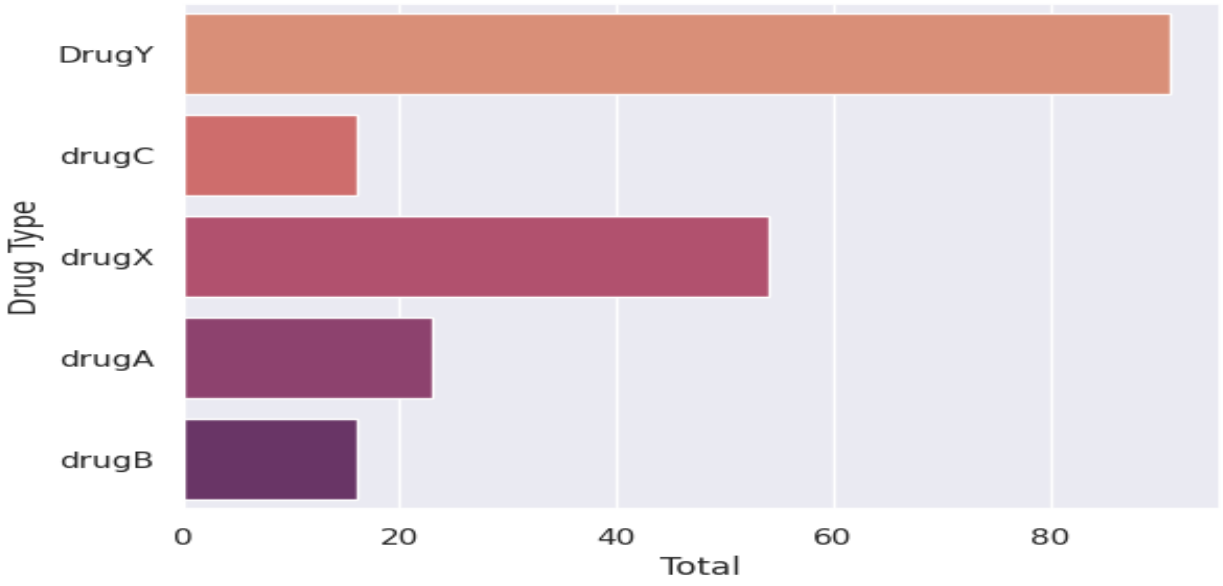
```
pd.crosstab(df_drug.Sex,df_drug.Drug).plot(kind="bar",figsize=(12,5),color=['#003f5c','#ffa600','#5850
8d','#bc5090','#ff6361'])
plt.title('Gender distribution based on Drug type')
plt.xlabel('Gender')
plt.xticks(rotation=0)
plt.ylabel('Frequency')
plt.show()
pd.crosstab(df_drug.BP,df_drug.Cholesterol).plot(kind="bar",figsize=(15,6),color=['#6929c4','#1192e8']
)
plt.title('Blood Pressure distribution based on Cholesterol')
plt.xlabel('Blood Pressure')
plt.xticks(rotation=0)
plt.ylabel('Frequency')
plt.show()
plt.scatter(x=df_drug.Age[df_drug.Sex=='F'], y=df_drug.Na_to_K[(df_drug.Sex=='F')], c="Blue")
plt.scatter(x=df_drug.Age[df_drug.Sex=='M'], y=df_drug.Na_to_K[(df_drug.Sex=='M')], c="Orange")
plt.legend(["Female", "Male"])
plt.xlabel("Age")
plt.ylabel("Na_to_K")
plt.show()
bin_age = [0, 19, 29, 39, 49, 59, 69, 80]
category_age = ['<20s', '20s', '30s', '40s', '50s', '60s', '>60s']
df_drug['Age_binned'] = pd.cut(df_drug['Age'], bins=bin_age, labels=category_age)
df_drug = df_drug.drop(['Age'], axis = 1)
bin_NatoK = [0, 9, 19, 29, 50]
category_NatoK = ['<10', '10-20', '20-30', '>30']
df_drug['Na_to_K_binned'] = pd.cut(df_drug['Na_to_K'], bins=bin_NatoK, labels=category_NatoK)
df_drug = df_drug.drop(['Na_to_K'], axis = 1)
Tarining :
from sklearn.metrics import confusion_matrix
from sklearn.metrics import classification_report
X = df_drug.drop(["Drug"], axis=1)
y = df_drug["Drug"]

from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.3, random_state = 0)
X_train = pd.get_dummies(X_train)
X_test = pd.get_dummies(X_test)
X_train.head()
X_test.head()
from imblearn.over_sampling import SMOTE
X_train = X_train.astype(object)
y_train = y_train.astype(object)

# Initialize SMOTE
```

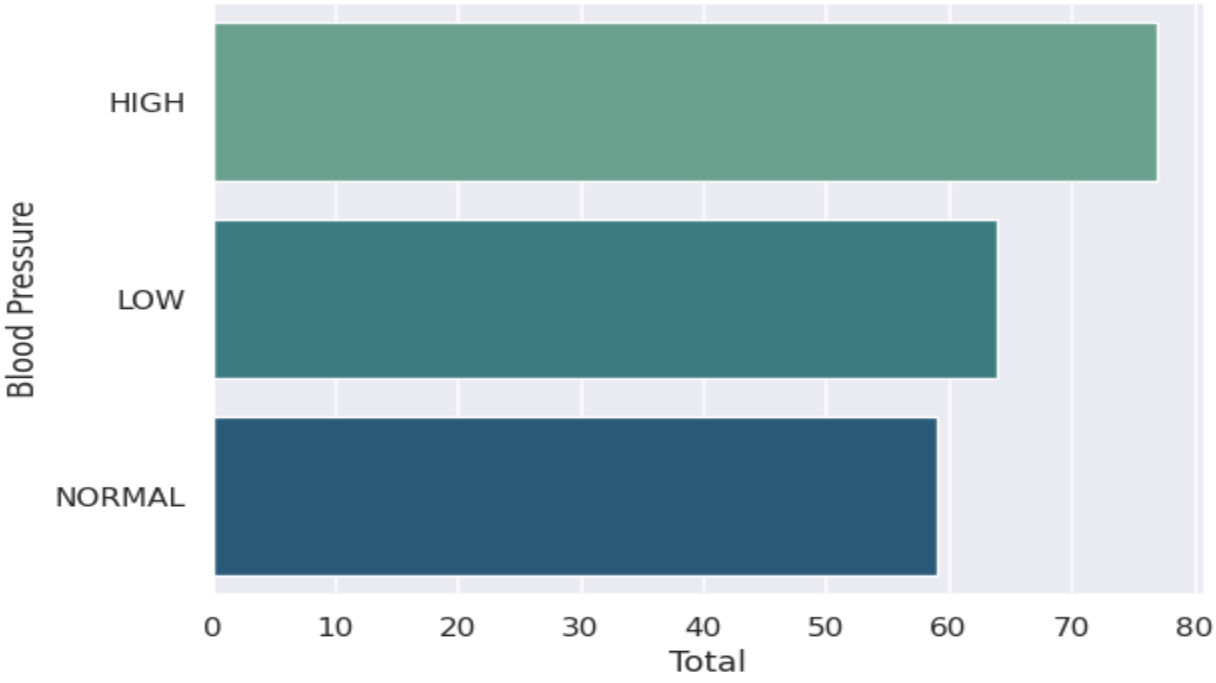
```
smote = SMOTE()

# Resample the data
X_train_resampled, y_train_resampled = smote.fit_resample(X_train, y_train)
sns.set_theme(style="darkgrid")
sns.countplot(y=y_train, data=df_drug, palette="mako_r")
plt.ylabel('Drug Type')
plt.xlabel('Total')
plt.show()
from sklearn.linear_model import LogisticRegression
LRclassifier = LogisticRegression(solver='liblinear', max_iter=5000)
LRclassifier.fit(X_train, y_train)

y_pred = LRclassifier.predict(X_test)

print(classification_report(y_test, y_pred))
print(confusion_matrix(y_test, y_pred))

from sklearn.metrics import accuracy_score
LRAcc = accuracy_score(y_pred,y_test)
print('Logistic Regression accuracy is: {:.2f}%'.format(LRAcc*100))
```

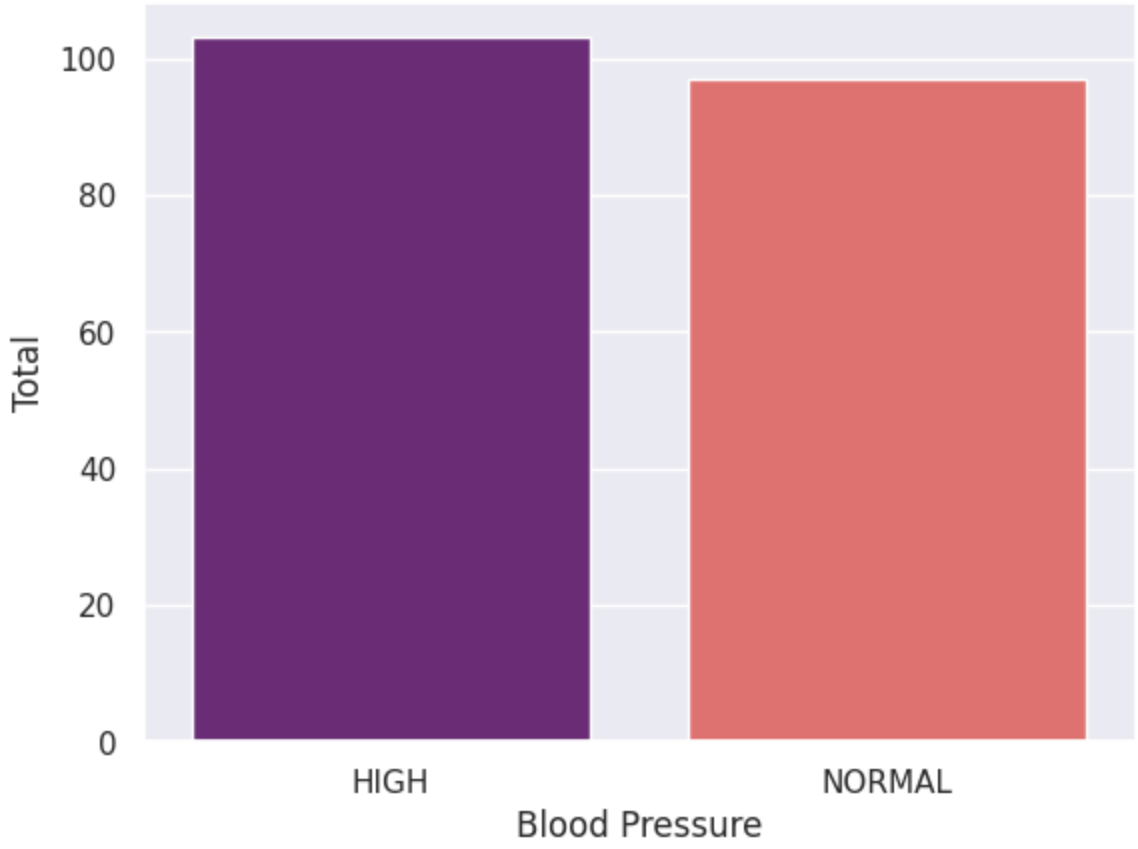
Out puts:

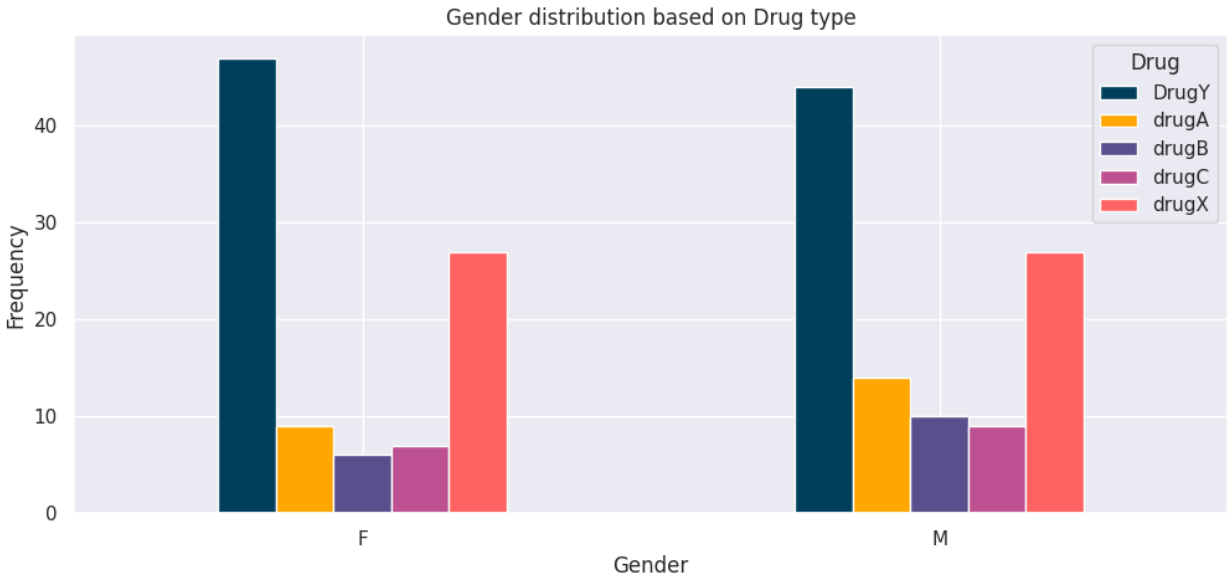

0.0200
0.0175
0.0150
0.0125
0.0100
0.0075
0.0050
0.0025
0.0000
Density
0
20
40
60
80
Age

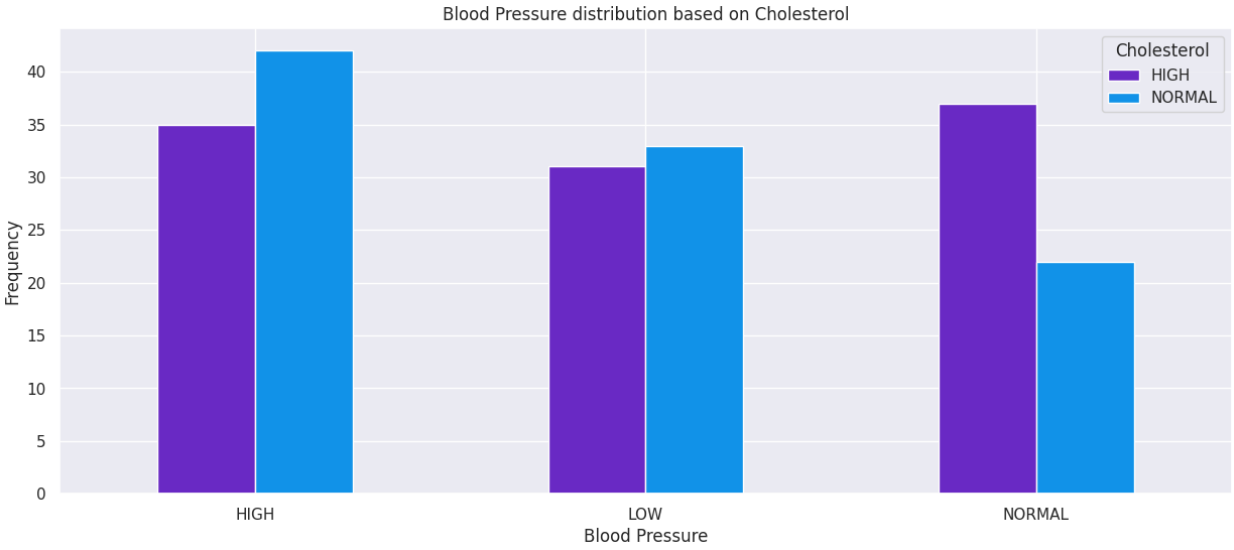

0.08
0.06
0.04
0.02
0.00
Density
0
10
20
30
40
Na_to_K

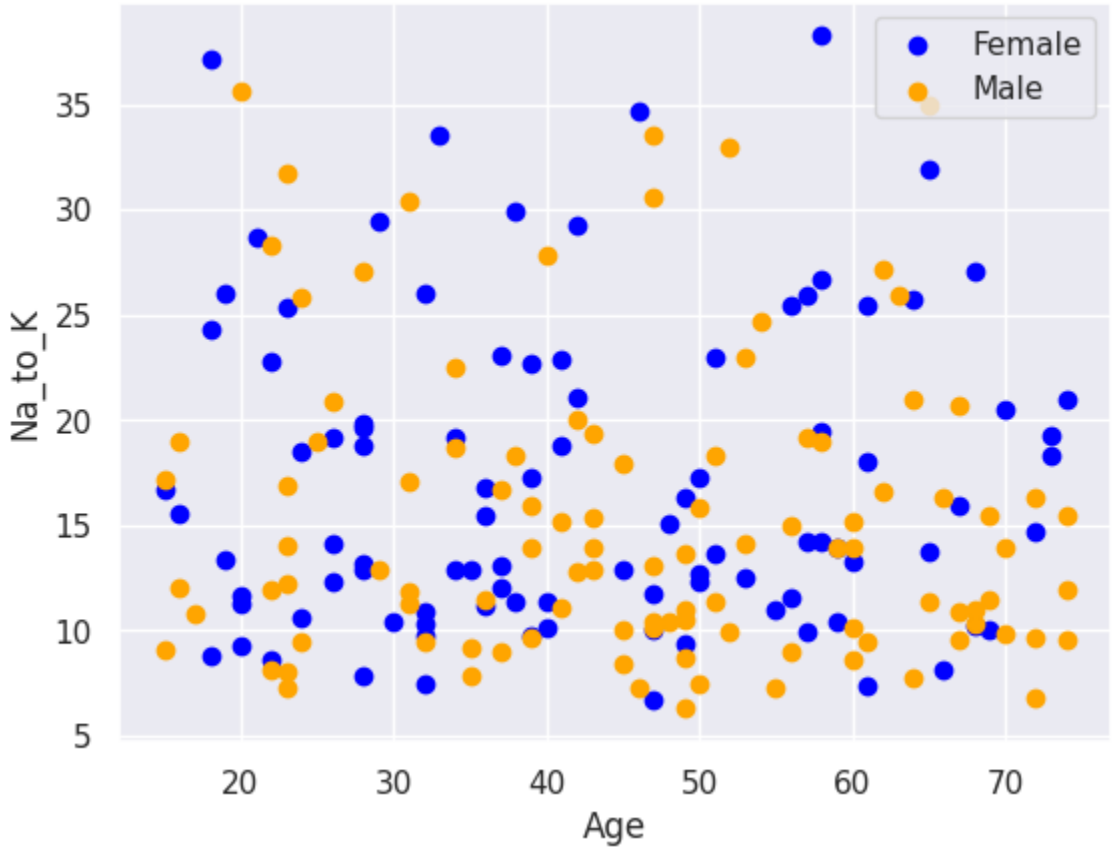

DrugY
drugC
drugX
drugA
drugB
Drug Type
0
20
40
60
80
Total

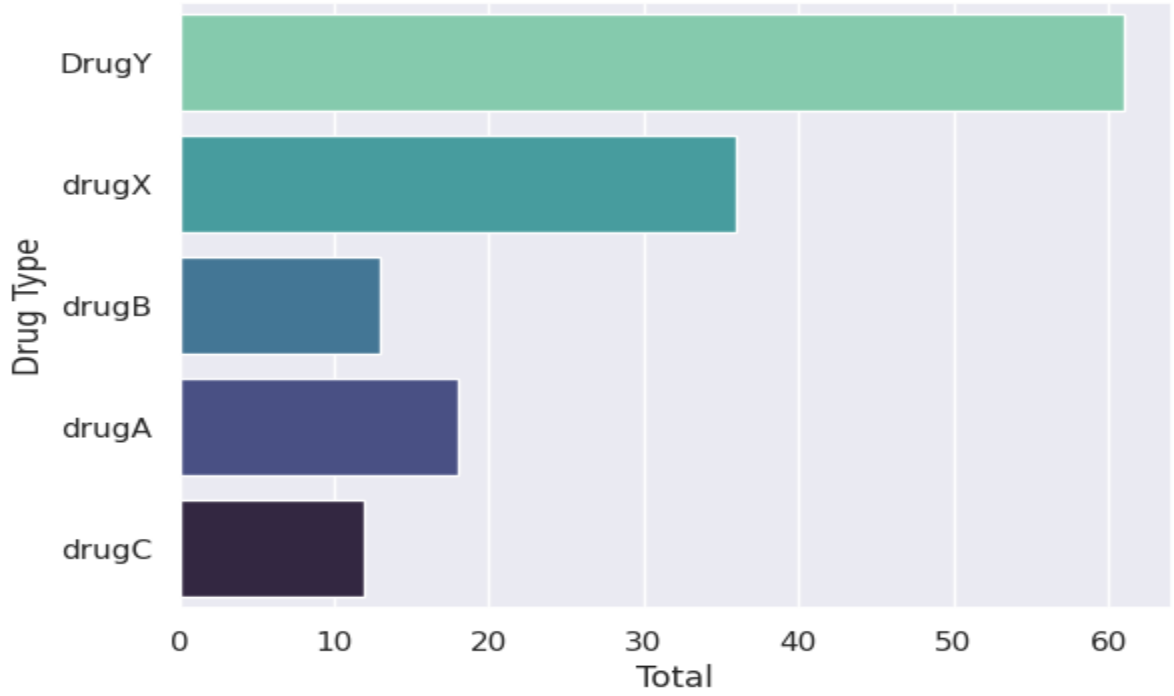

HIGH
LOW
NORMAL
Blood Pressure
0
10
20
30
40
50
60
70
80
Total

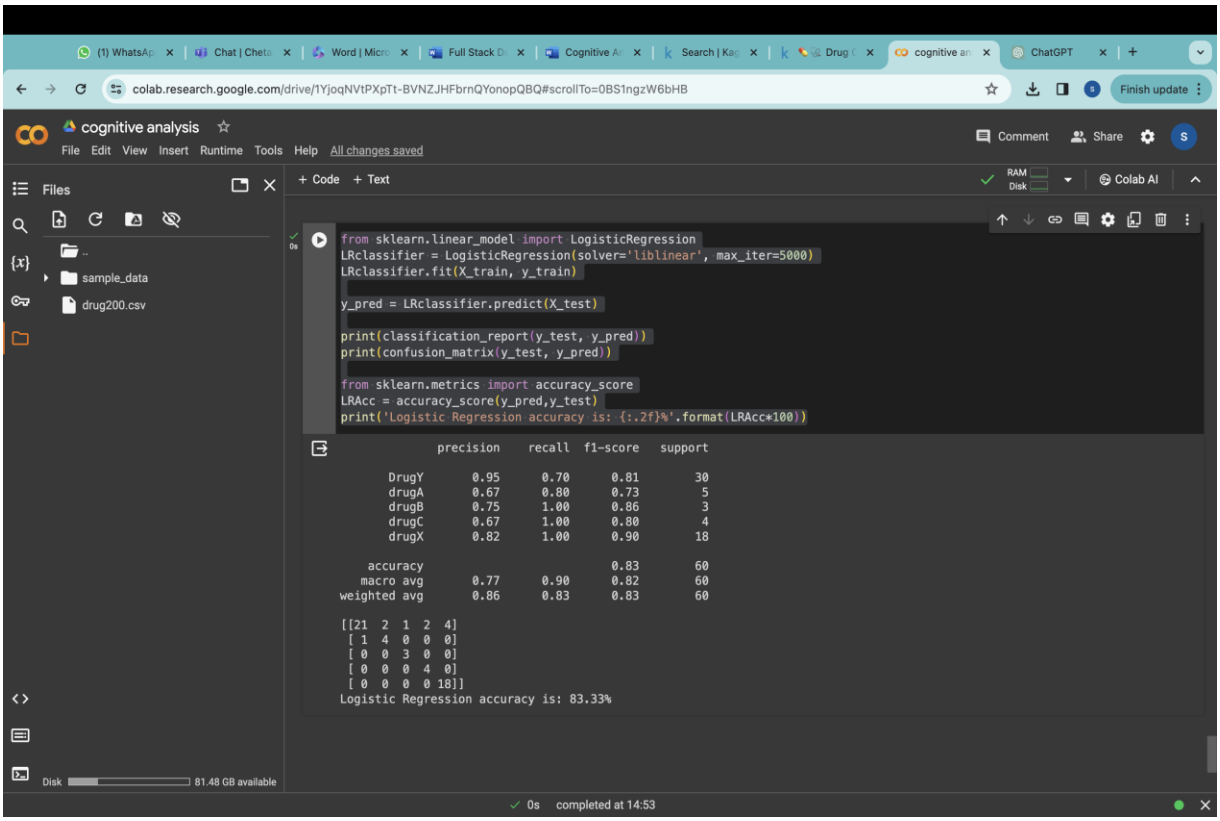

100
80
60
40
20
0
Total
HIGH
NORMAL
Blood Pressure


Gender distribution based on Drug type
Drug
DrugY
drugA
drugB
drugC
drugX
40
30
20
10
0
Frequency
F
M
Gender


Blood Pressure distribution based on Cholesterol
Cholesterol
HIGH
NORMAL
Frequency
0
5
10
15
20
25
30
35
40
HIGH
LOW
NORMAL
Blood Pressure


Female
Male
Na_to_K
5
10
15
20
25
30
35
20
30
40
50
60
70
Age


DrugY
drugX
drugB
drugA
drugC
Drug Type
0
10
20
30
40
50
60
Total


colab.research.google.com/drive/1YjoqNVtPXpTt-BVNZJHFbrnQYonopQBQ#scrollTo=0BS1ngzW6bHB
cognitive analysis
File Edit View Insert Runtime Tools Help All changes saved
Comment
Share
Files
+ Code + Text
Colab AI
sample_data
drug200.csv
from sklearn.linear_model import LogisticRegression
LRclassifier = LogisticRegression(solver='liblinear', max_iter=5000)
LRclassifier.fit(X_train, y_train)
y_pred = LRclassifier.predict(X_test)
print(classification_report(y_test, y_pred))
print(confusion_matrix(y_test, y_pred))
from sklearn.metrics import accuracy_score
LRAcc = accuracy_score(y_pred,y_test)
print('Logistic Regression accuracy is: {:.2f}%'.format(LRAcc*100))
precision recall f1-score support
DrugY 0.95 0.70 0.81 30
drugA 0.67 0.80 0.73 5
drugB 0.75 1.00 0.86 3
drugC 0.67 1.00 0.80 4
drugX 0.82 1.00 0.90 18
accuracy 0.83 60
macro avg 0.77 0.90 0.82 60
weighted avg 0.86 0.83 0.83 60
[[21 2 1 2 4]
[1 4 0 0 0]
[0 0 3 0 0]
[0 0 0 4 0]
[0 0 0 0 18]]
Logistic Regression accuracy is: 83.33%
Disk 81.48 GB available
0s completed at 14:53

Conclusion

In conclusion, logistic regression serves as a valuable tool in drug prediction, offering a straightforward yet powerful approach to classify patients based on their likelihood of responding positively or negatively to specific medications. Through this study, we have demonstrated the efficacy of logistic regression in predicting drug responses by leveraging patients' demographic, clinical, and genetic characteristics. The application of logistic regression in drug prediction holds significant promise for personalized medicine, enabling healthcare practitioners to make informed treatment decisions tailored to individual patient profiles. By analyzing the relationship between input features and drug response probabilities, logistic regression models provide interpretable insights into the factors influencing treatment outcomes. Our findings highlight the importance of incorporating diverse sources of data, including genetic markers, biomarkers, and medical history, to enhance the predictive accuracy of logistic regression models. Furthermore, the simplicity and transparency of logistic regression make it accessible to healthcare professionals, facilitating the integration of predictive analytics into clinical practice. Moving forward, continued research in drug prediction using logistic regression should focus on refining predictive models, incorporating additional data sources, and exploring advanced machine learning techniques to further improve prediction accuracy and generalizability. By harnessing the power of logistic regression, we can advance personalized medicine initiatives, optimize therapeutic interventions, and ultimately improve patient outcomes in healthcare.