Determining the most effective machine learning model on segmented vs. unsegmented patient data to assign Drug Classification

Nino Saeki

Abstract

Throughout history and in the modern world, accurate drug prescription has been one of the most important tasks that a medical professional does in their everyday tasks. Through recent innovations in the machine learning environment, algorithms and models can more accurately predict the physiological activity of drugs and further classify drugs based on their physiological properties. This project focuses on the latter, and leverages a sample patient dataset and runs logistic regression, k-neighbors, support vector machine, naïve Bayes, decision trees, and the random forest models to determine accuracy between models. Afterwards, the data is segmented by sex, and the models are implemented on each dataset, and the accuracies are compared. The accuracy for the models that were applied to the entire dataset are the following (greatest to least accurate): decision tree (100%), random forest (100%), SVM (98%), naive Bayes (83%), logistic regression (83%), and k-Neighbors (66.67%). Overall, segmentation had the smallest effect on the Random Forest and Decision Tree models as both produced a 0% difference in accuracy between male and female datasets, and had the biggest effect on the k-neighbors model with a 38.03% between male and female datasets.

Key Words

Machine Learning, k-Neighbors, Random Forest, Decision Tree, Support Vector Machine, Drug Classification, Female vs. Male, Patient Data

Introduction

Drug Implementation and classification techniques have long been one of the more difficult elements of the process which begins at the inception of the drug, and ends with the user taking the drug to cure their ailment. Traditionally, medical practitioners have manually given prescriptions to patients, but with recent developments in artificial intelligence (AI), the use of a machine learning (ML) model may remove human error from the equation (Medlinskiene et al. 2021). There has also been plenty of argument against the use of ML models, as models such as decision trees can easily fall into the trap of overfitting data or also arriving at incorrect conclusions due to biases that remained undetected to the creators of the data (Sae-Ang et al. 2022). This push for more traditional techniques, however, is also argued by scientists that believe these models can be used in tandem with medical practitioners to reach accurate conclusions(Wang et al., 2023). Model application has also led to decreased healthcare professionals workload through more effective data analysis and decision-making models. (Kumar et al. 2022). There are also arguments that “AI’s capacity to process vast amounts of clinical data, which can aid in ‘making more accurate diagnosis and developing personalized treatment plans’”(Karalis, 2024). By testing and implementing logistic regression, k-neighbors, support vector machine, naïve bayes, decision tree, and the random forest algorithms on both male and female datasets, a theoretical determination can be made about which model is the best applicable for which sex.

Methods

The dataset that was used for training and deploying the several different machine learning models is a UCI ML Dummy Training Set, https://www.kaggle.com/code/caesarmario/drug-classification-w-various-ml-models/input?select=drug200.csv. The dataset contains four categorical variables and two continuous variables: the age of the patient, the sex of the patient, the blood pressure of the patient, the cholesterol levels of the patient and the Sodium to Potassium Ratio of the patient. The age variable is an integer value (Figure 5) and the sex is a binary variable, however it is technically a string value of either “M” or “F” (Figure 2) . Furthermore, the blood pressure (BP) is a string value of either “HIGH”, ”MEDIUM”, or ”LOW” (Figure 3). The cholesterol is also binned similarly to the blood pressure, however, there are no “LOW” cholesterol values (Figure 1). The Sodium to Potassium values are all float values (Figure 6). The final variable is the drug variable which has five different values which are all strings. The drugs that are used are “drugA”, “drugB”, “drugC”, “drugX” and “drugY” (Figure 4). There were two-hundred total patients.

Feature Engineering was an integral of this project. The primary motivation for this “data cleaning” step was to allow operations to be done on the data in future steps. Some of these machine learning models require all passed values to be integer values. For this reason, the dataset had to be modified to be effective when being used for all models. A function called “strToIntConverter” was used to convert all “HIGH”, ”MEDIUM”, and ”LOW” to integer values. This function can still work for variables other than just blood pressure. Since variables such as cholesterol also have both “HIGH” and “NORMAL”, this function was also used for cholesterol as well. Sex was unchanged as it was going to be used to separate the data set.

Multiple feature engineering approaches were explored, and finally, the function described below was selected for final implementation. Instead of converting all variables into integer values, a previous iteration method expands the single column into 3 columns and uses each column to keep track of “HIGH”, ”MEDIUM”, and ”LOW” (Figure 25). While this may have its own advantages, a different, and perhaps, more simple approach was used for this project (Figure 27).

An integral part of this research project was to run different machine learning models on the entire data set, then run the ML models on segmented female and male data sets individually. In order to do that, the data had to be separated. First a comparison variable was initialized and set. Afterwards, a new list is created and every entry that passes the condition we set makes it into the new list. This is then done with both genders, so now we have 2 data frames of different sexes.

For the research project, six different ML models were used to compare to each other. Logistic Regression, K. Neighbors, Support Vector Machine, Naïve Bayes, Decision Tree, and the Random Forest. The base library that was used was called “sklearn.linear_model”. From this library, each of the different models were imported and used. Since there is no “Industry” standard ratio for separating data evaluation vs training, a 30:70 ratio was used for training to use. For each of the six different models, each was run on the three respective lists (Male, Female, Both). The library that was imported has a feature called the “classification_report”. In this report, information such as precision, accuracy, recall, f1-score, and support are shown. These are all valuable metrics that will be discussed further in a later section. A confusion matrix is also created by the library. This was an additional metric explored in the analysis of the machine-learning approaches (Figure 26).

For the graphing of the different components of the dataset, Matplotlib was used. It is a library for python that has many visualization methods. For this specific project, bar graphs were the most used. There were several problems when graphing specific variables such as the Cholesterol level. After implementing the strToIntConverter, it is important to overwrite the titles to the appropriate names. When implementing the function, the title of the column also becomes a value, and MatPlotLib does not take an integer as a column name when graphing.

Link to full repository can be found at the following address:
NSaeki1103/Drug-Classification-Project (github.com)

Results

The first model that was implemented was the random forest ML model. The number of leaf nodes was forced to stay under 30 total leaf nodes. Random forest ML Models are an industry standard initial testing model. Random forest ML models create multiple decision trees where each tree is built from a random subset of the data/features. These trees are all random forest models that are less effective at predicting traits of new data, however they are still a good way of benchmarking whether the data can be tested with other ML models. For those reasons, the random forest ML model was used initially. When the random forest was used on all the data, the accuracy was 100% and the recall value and f1-score were both 1.0 (Figure 7). This remained the same for both male and female data sets. All three datasets were at 100% effectivity (Figure 8) (Figure 9).

$$\boldsymbol{\Delta}_{(accuracy\ difference\ \ male\ and\ female)} = \boldsymbol{a}_{male\ accuracy} - \boldsymbol{a}_{female\ accuracy}$$

By applying the equation above, the difference in percent accuracy between the model implemented on male and female patients can be determined. The difference was 0%.

The second implemented AI Model was the K-Neighbors ML model. This model classifies data points based on the majority of their k closest neighbors. While being a simple algorithm, K-Neighbors is very computationally expensive and is not effective for large data sets. However, this UCI Data set is relatively small, so the model was effective. For the dataset with both male and female, the accuracy was 66.67%, the recall value fluctuated between 0.00 to 0.9 for each drug, and the f1 score also fluctuated between 0.0 and 0.95 for each drug (Figure 10). The accuracy for the only male dataset was 34.38% (Figure 11). The f1-score and recall values were both fluctuating between 0.00 and 0.62 with an outlier of a recall value of 1.00 for drug B. For the female dataset, the accuracy was 72.41% (Figure 12). The f1-scores and the recall values were higher for this dataset.

By applying the equation above, the difference in percent accuracy between the model implemented on male and female patients can be determined. The difference was -38.03%

The third model that was implemented was logistic regression. While this model assumes that there is a correlation between input and output variables, it was still effective to test on the data. The model estimates the probability of a data point belonging to a specific class by attempting to fit the data point on a logarithmic regression. This is a relatively popular choice in the Industry because of its simplistic nature and its relative accuracy. When the model was implemented on the data set with both male and female, the accuracy was 83% (Figure 13). The recall values were 1.00 for three out of five drugs. The f1-scores were more variable, but fluctuated between a smaller range of 0.67 to 0.89. When the model was run on the dataset with just males, the accuracy was 78% (Figure 14). The recall values remained similar to the previous model, but the recall values for the other two drugs were way lower. The f1-scores were slightly lower overall for this dataset (when compared to the original dataset testing). When the model was tested on the female data set, the accuracy was 89% (Figure 15). The recall values/f1-score were similar to original dataset testing.

By applying the equation above, the difference in percent accuracy between the model implemented on male and female patients can be determined. The difference was -11%

The fourth model that was used was the Support Vector Machine. The SVM Model separates data points into very different categories. When the dataset becomes more complicated, the SVM uses a kernel to separate the values. While the SVM is effective at separating data and has high accuracy, the use of a kernel can lead to very slow processing which is very computationally taxing. When implementing the SVM on the dataset with both male and female, the accuracy was 98% (Figure 16). The recall values and f1-scores are almost all 1.00. Implementation of the SVM on the male only dataset resulted in an accuracy of 100% (Figure 17). After implementing the SVM Model on the Female data set, the resulting implementation led to an accuracy of 89.66% (Figure 18). Drug B had a very low precision and f1 score of 0.33 and 0.5 compared to the others.

By applying the equation above, the difference in percent accuracy between the model implemented on male and female patients can be determined. The difference was 10.34%

The next Model that was implemented was the Naïve Bayes model. Naïve Bayes Models are probabilistic machine learning algorithms that assume that all features are independent of each other, which allows the algorithm to simplify computational tasks. The program is Naïve as it assumes that there are no correlations between data, even when there most likely is. As a result of this algorithm, the model can be implemented on larger datasets while not adding high amounts of stress to the computational ability of the computer. For the overall dataset with both male and female, the accuracy was 83% (Figure 19). The f1-scores and the recall values mostly ranged between 0.30 to 0.85. The accuracy for the dataset with only males was 75% (Figure

20). The recall values were mostly all 1.00 while the f1-scores were mostly ranging between 0.65 to 0.76. When the model is implemented on the female only dataset, the accuracy was 89.66% (Figure 21).

By applying the equation above, the difference in percent accuracy between the model implemented on male and female patients can be determined. The difference was -14.66%

The final model that was used was the Decision tree ML model. Although a random forest model was already done, the decision tree model was also used to set a baseline for a model with reduced variance when compared to the random forest model. This model separates the data into binary categories and categorizes the data using a flowchart. While it is not computationally draining, there are several drawbacks to this model of data analysis. One of the big issues with a decision tree is that the model is prone to overfitting. Overfitting is when the ML Model finds the trends and incorporates them into predictions, but also accidentally incorporates background noise/random details into predictions. While it may work for this specific dataset, there are issues that can arise when using the model on new data. For the dataset with all male and female data, accuracy was 100% (Figure 22). When using the model on only male patients, the accuracy was also 100% (Figure 23). Finally, when implemented on the female dataset, the accuracy was 89% (Figure 24).

By applying the equation above, the difference in percent accuracy between the model implemented on male and female patients can be determined. The difference was 11%.

Discussion

Separating the data into male and female patients was very important for this project. There were 2 reasons for this segmentation in the data: The main objective of this project was to determine how each selected Machine learning approach would absorb and deploy its own analysis of the data onto “new” data. By separating the data into male and female, insights on ML models’ accountability on the sex differences in humans can be had. Furthermore, accuracies of respective genders in relation to the whole data set can also amplify previously discovered trends. Science in the past has already shown that there are vast biological differences between males and females (Arnold et al., 2024). As a result of these biological/fundamental differences between males and females, the data was separated to better represent each group.

It is important to note that the negative percent differences in some of the accuracy differences between male and female is intentional. The negative difference implies that the model accuracy on the female data was higher than the model accuracy on the male dataset. The accuracy for the models that were applied on the entire dataset are the following (greatest to least accurate): Decision Tree (100%), Random Forest (100%), SVM (98%), Naive Bayes (83%), Logistic Regression (83%), and K-Neighbors (66.67%). Segmentation did have an effect for the decision tree. While the accuracy for males was 100% the accuracy for females was only 89%. For the Decision Tree model the combined data set resulted in an accuracy for the combined dataset, but the model accuracy on the female data set only resulted in an accuracy of 89%, while male accuracy was at 100%. For the Random Forest model, the male and female segmentation appears to have made no difference, as both resulted in a 100% accuracy for each respective dataset. When the Naive Bayes Model was implemented on the male and female datasets, the male dataset was 14.66% less accurate than the female dataset. When the SVM model was applied to the male and female datasets, the male dataset model was 10.34% more accurate than the female dataset. After applying the Logistic Regression Model on both the male and female dataset, the accuracy of the model on the female data set appeared to be 11% greater than the model accuracy on the male dataset. The final model that was applied on the data was the K-Neighbors model. The K-Neighbors model was 38.03% more accurate when applied on the female dataset, rather than the male dataset. Segmenting by gender did make a difference (d’Emden, 2014).

There are several reasons that there are differences in the levels of accuracy for models implemented on male and female datasets. One of the major challenges in machine learning is bias from past data sources and the inability to generalize to future datasets (d’Emden, 2014). The discrepancies between the accuracy of the ML models for male and female datasets encourage the question whether data should be segmented. While model performance, for both segmented and unsegmented datasets, was similarly accurate, 100%, for both Decision Tree and Random Forest models, caution should be given to these results. For instance, a recent report discusses the overfitting of data during Random Forest implementation (https://diagnprognres.biomedcentral.com/articles/10.1186/s41512-024-00177-1) . While Decision Trees relies on different model architectures, similar results were reached, which encourages further research on this topic. Gender-specific models will most likely increase the ML ecosystem within the healthcare industry. Instead of running general models on entire datasets, specific ML models can be implemented and tested on specific data types, e.g., based on sex.

The Support Vector Machine Model is an effective model to implement on a general dataset. SVM Models are very robust and are capable of working in datasets with large numbers of features (Guido et al., 2024). SVM Models are also good at generalizing based on small amounts of data. In a field such as medical diagnostics, where data is sometimes scarce, this method of data analysis may tend to be more effective (Cham, 2023). While these are all positive features, there are also additional considerations. SVM Models are relatively computationally complex, so for larger datasets the models will require a lot of computational power (Guido et al., 2024). Another consideration that comes with SVM Models is the prevalence of overfitting as a result of Unbalanced datasets (which are commonplace in healthcare) (Cham, 2023).

Regardless of these advantages/disadvantages, the project highlighted the SVM Model as the most effective for general use on medical patient datasets for drug classification tasks.

The machine learning space and its applications on the medical sphere still remain relatively undiscovered, and most new application models are still novel. Machine Learning uses patterns in data to predict a new patient's possible ailments (Habehh & Gohel, 2021). This process also occurs in the real world, when a doctor will use past diagnoses to determine what a new patient's issues may be and prescribe the correct drug. Unlike a human doctor, however, machine learning can use only mathematical probabilities to determine the correct drug to prescribe to a patient. To extend this project, one should consider using more ML models on similar datasets with/without segmentation, and creating more insights and comparisons between models. Models such as gradient boosting machines, neural networks, and k-mean clustering should be considered.

Figure Legend

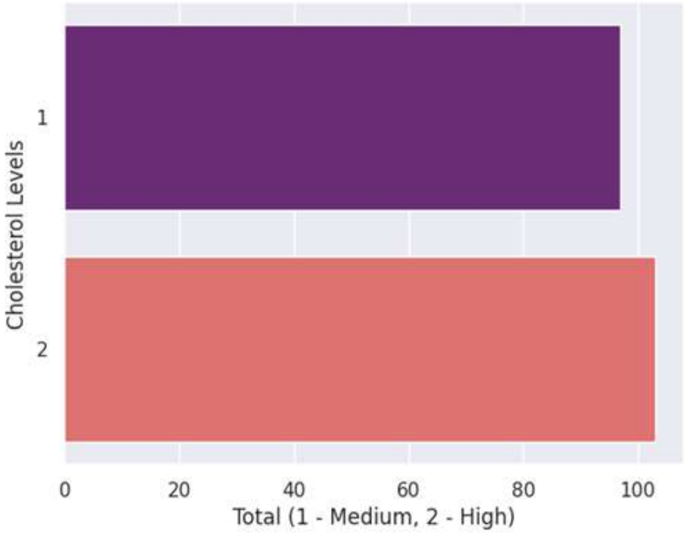


Fig. 1. Graph showing cholesterol values for the UCI ML Dataset.

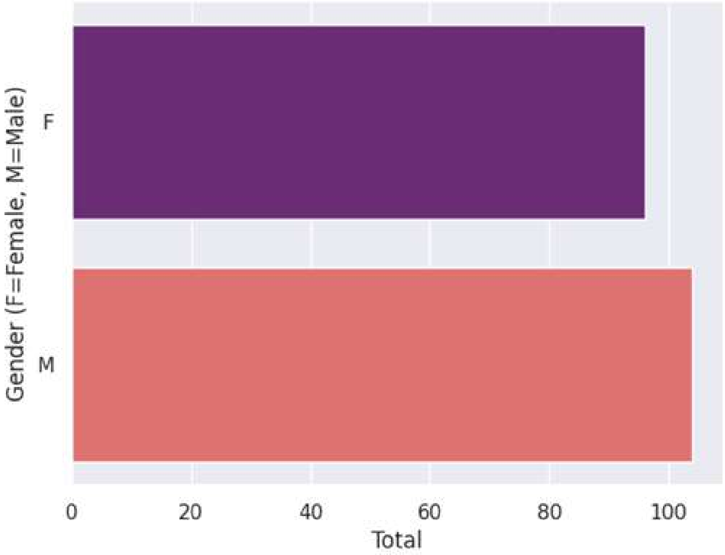


Fig. 2. Graph showing number of males and females in the UCI ML Dataset.

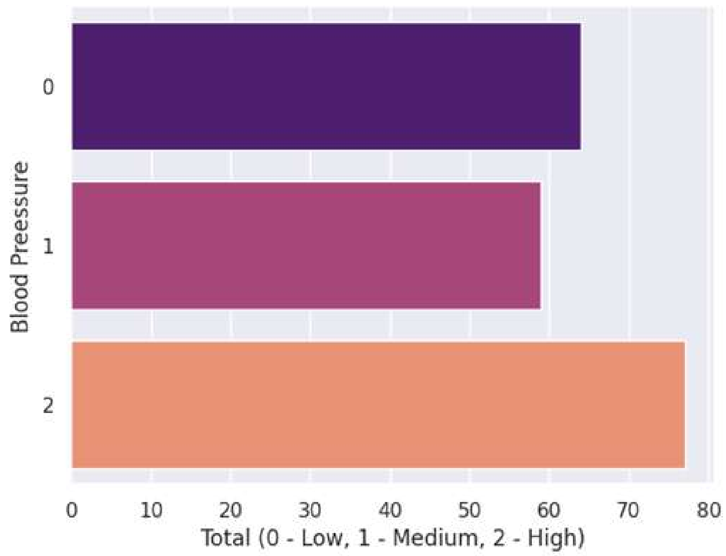


Fig. 3. Graph showing number of patients with Low, Medium and High blood pressure in UCI ML Dataset.

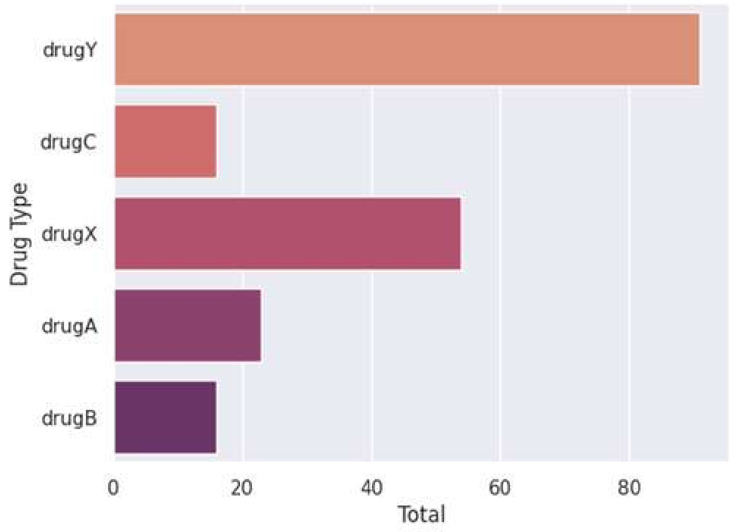


Fig. 4. Graph showing number of patients with each prescribed drug in the UCI ML Dataset.

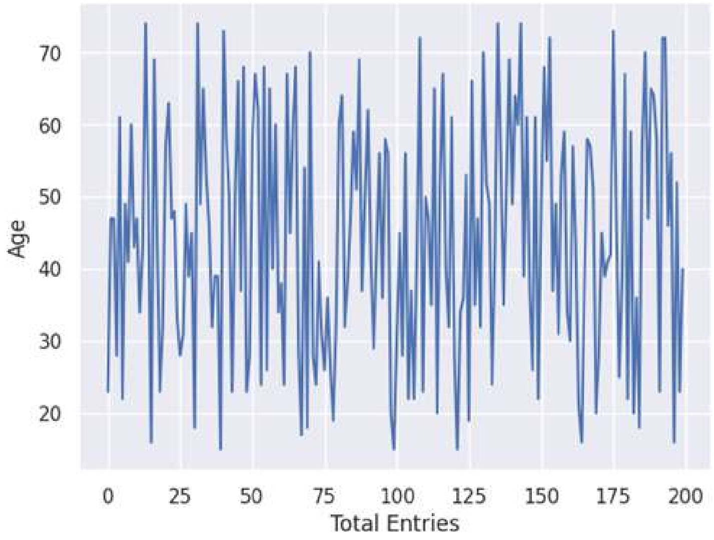


Fig. 5. Graph of all patient’s ages in UCI ML Dataset.

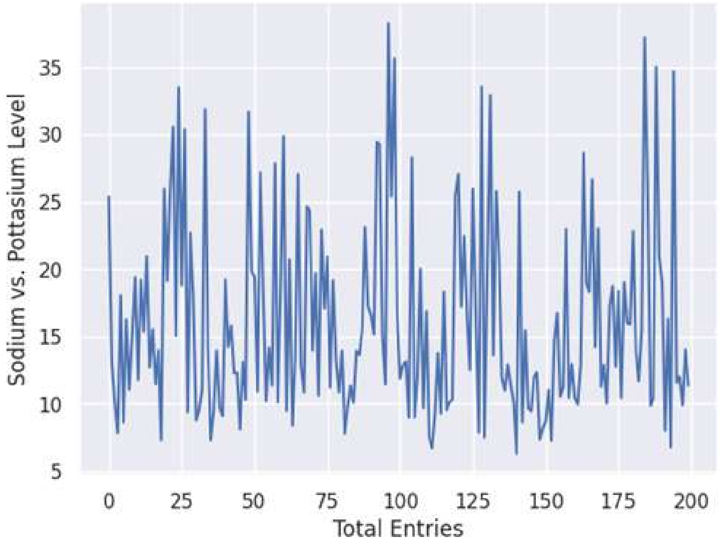
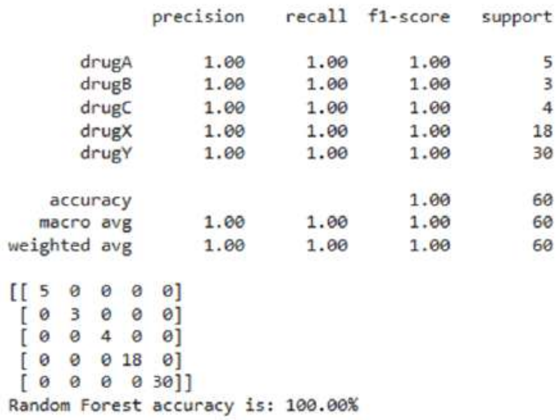


Fig. 6. Sodium vs. Potassium ratios of each patient in the UCI ML Dataset.



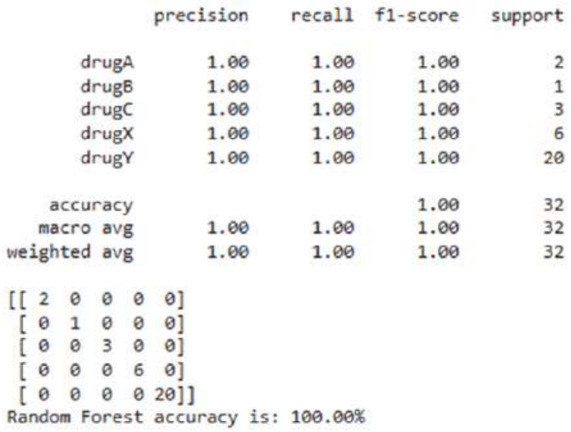
```
              precision    recall  f1-score   support

       drugA       1.00      1.00      1.00         5
       drugB       1.00      1.00      1.00         3
       drugC       1.00      1.00      1.00         4
       drugX       1.00      1.00      1.00        18
       drugY       1.00      1.00      1.00        30

    accuracy                           1.00        60
   macro avg       1.00      1.00      1.00        60
weighted avg       1.00      1.00      1.00        60

[[ 5  0  0  0  0]
 [ 0  3  0  0  0]
 [ 0  0  4  0  0]
 [ 0  0  0 18  0]
 [ 0  0  0  0 30]]
Random Forest accuracy is: 100.00%
```

Fig. 7. Data output from Random Forest model Implementation on all data.



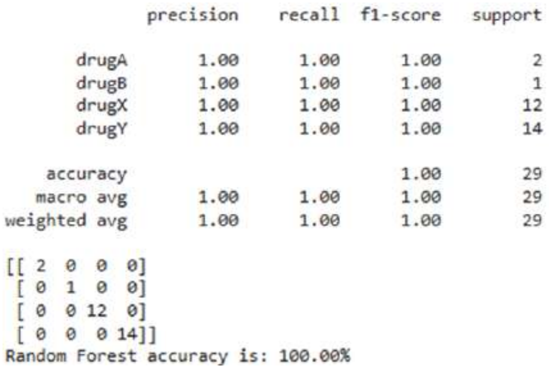
```
              precision    recall  f1-score   support

       drugA       1.00      1.00      1.00         2
       drugB       1.00      1.00      1.00         1
       drugC       1.00      1.00      1.00         3
       drugX       1.00      1.00      1.00         6
       drugY       1.00      1.00      1.00        20

    accuracy                           1.00        32
   macro avg       1.00      1.00      1.00        32
weighted avg       1.00      1.00      1.00        32

[[ 2  0  0  0  0]
 [ 0  1  0  0  0]
 [ 0  0  3  0  0]
 [ 0  0  0  6  0]
 [ 0  0  0  0 20]]
Random Forest accuracy is: 100.00%
```

Fig. 8. Data output from Random Forest model Implementation on just male data.



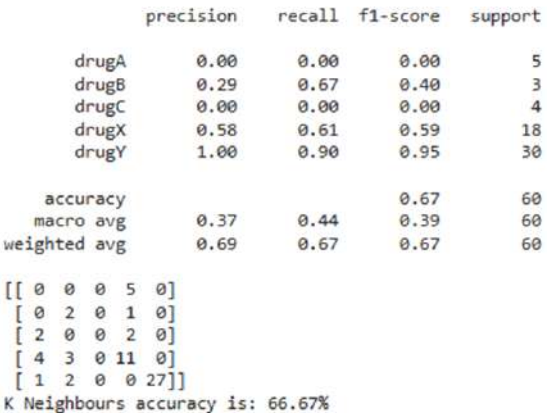
```
              precision    recall  f1-score   support

       drugA       1.00      1.00      1.00         2
       drugB       1.00      1.00      1.00         1
       drugX       1.00      1.00      1.00        12
       drugY       1.00      1.00      1.00        14

    accuracy                           1.00        29
   macro avg       1.00      1.00      1.00        29
weighted avg       1.00      1.00      1.00        29

[[ 2  0  0  0]
 [ 0  1  0  0]
 [ 0  0 12  0]
 [ 0  0  0 14]]
Random Forest accuracy is: 100.00%
```

Fig. 9. Data output from Random Forest model Implementation on just female data.



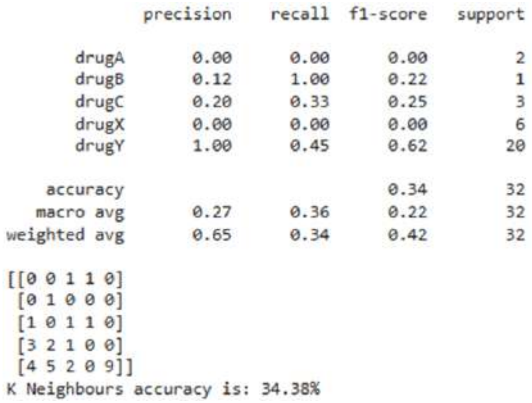
```
              precision    recall  f1-score   support

       drugA       0.00      0.00      0.00         5
       drugB       0.29      0.67      0.40         3
       drugC       0.00      0.00      0.00         4
       drugX       0.58      0.61      0.59        18
       drugY       1.00      0.90      0.95        30

    accuracy                           0.67        60
   macro avg       0.37      0.44      0.39        60
weighted avg       0.69      0.67      0.67        60

[[ 0  0  0  5  0]
 [ 0  2  0  1  0]
 [ 2  0  0  2  0]
 [ 4  3  0 11  0]
 [ 1  2  0  0 27]]
K Neighbours accuracy is: 66.67%
```

Fig. 10. Data output from K Neighbor model Implementation on all data.



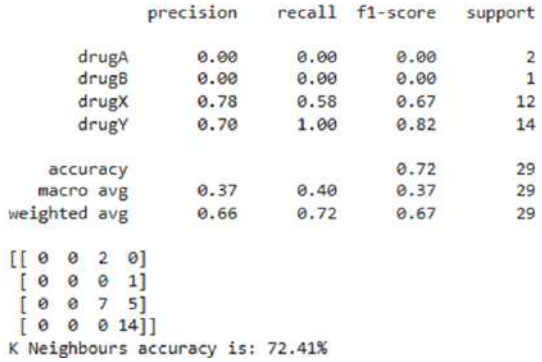
```
              precision    recall  f1-score   support

       drugA       0.00      0.00      0.00         2
       drugB       0.12      1.00      0.22         1
       drugC       0.20      0.33      0.25         3
       drugX       0.00      0.00      0.00         6
       drugY       1.00      0.45      0.62        20

    accuracy                           0.34        32
   macro avg       0.27      0.36      0.22        32
weighted avg       0.65      0.34      0.42        32

[[0 0 1 1 0]
 [0 1 0 0 0]
 [1 0 1 1 0]
 [3 2 1 0 0]
 [4 5 2 0 9]]
K Neighbours accuracy is: 34.38%
```

Fig. 11. Data output from K Neighbor model Implementation on just male data.



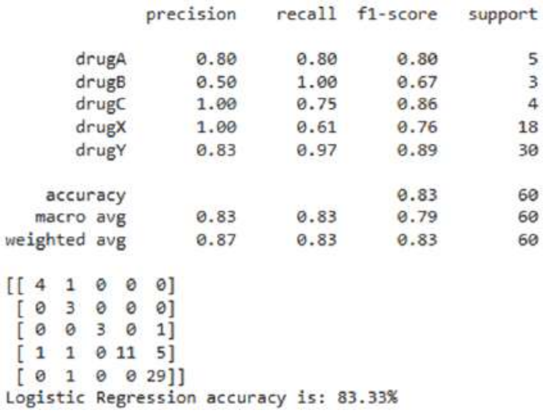
```
              precision    recall  f1-score   support

       drugA       0.00      0.00      0.00         2
       drugB       0.00      0.00      0.00         1
       drugX       0.78      0.58      0.67        12
       drugY       0.70      1.00      0.82        14

    accuracy                           0.72        29
   macro avg       0.37      0.40      0.37        29
weighted avg       0.66      0.72      0.67        29

[[ 0  0  2  0]
 [ 0  0  0  1]
 [ 0  0  7  5]
 [ 0  0  0 14]]
K Neighbours accuracy is: 72.41%
```

Fig. 12. Data output from K Neighbor model Implementation on just female data.



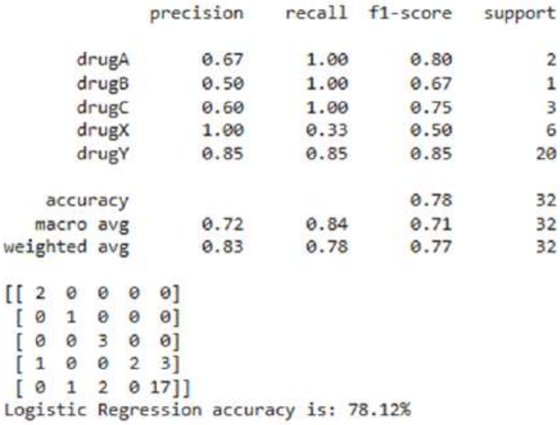
```
              precision    recall  f1-score   support

       drugA       0.80      0.80      0.80         5
       drugB       0.50      1.00      0.67         3
       drugC       1.00      0.75      0.86         4
       drugX       1.00      0.61      0.76        18
       drugY       0.83      0.97      0.89        30

    accuracy                           0.83        60
   macro avg       0.83      0.83      0.79        60
weighted avg       0.87      0.83      0.83        60

[[ 4  1  0  0  0]
 [ 0  3  0  0  0]
 [ 0  0  3  0  1]
 [ 1  1  0 11  5]
 [ 0  1  0  0 29]]
Logistic Regression accuracy is: 83.33%
```

Fig. 13. Data output from Logistic Regression model Implementation on all data.



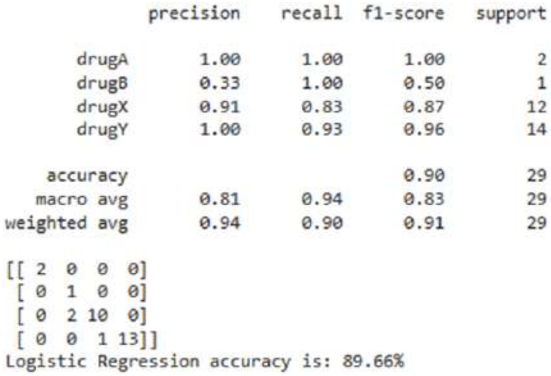
```
              precision    recall  f1-score   support

       drugA       0.67      1.00      0.80         2
       drugB       0.50      1.00      0.67         1
       drugC       0.60      1.00      0.75         3
       drugX       1.00      0.33      0.50         6
       drugY       0.85      0.85      0.85        20

    accuracy                           0.78        32
   macro avg       0.72      0.84      0.71        32
weighted avg       0.83      0.78      0.77        32

[[ 2  0  0  0  0]
 [ 0  1  0  0  0]
 [ 0  0  3  0  0]
 [ 1  0  0  2  3]
 [ 0  1  2  0 17]]
Logistic Regression accuracy is: 78.12%
```

Fig. 14. Data output from Logistic Regression model Implementation on just male data.



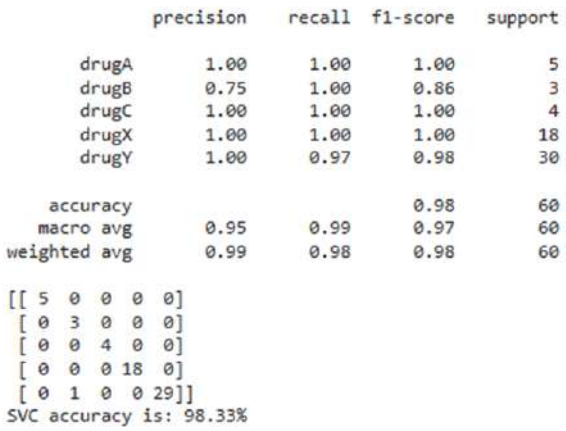
```
              precision    recall  f1-score   support

       drugA       1.00      1.00      1.00         2
       drugB       0.33      1.00      0.50         1
       drugX       0.91      0.83      0.87        12
       drugY       1.00      0.93      0.96        14

    accuracy                           0.90        29
   macro avg       0.81      0.94      0.83        29
weighted avg       0.94      0.90      0.91        29

[[ 2  0  0  0]
 [ 0  1  0  0]
 [ 0  2 10  0]
 [ 0  0  1 13]]
Logistic Regression accuracy is: 89.66%
```

Fig. 15. Data output from Logistic Regression model Implementation on just female data.



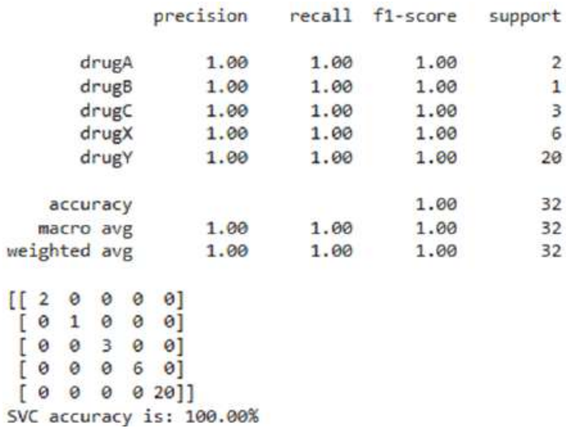
```
              precision    recall  f1-score   support

       drugA       1.00      1.00      1.00         5
       drugB       0.75      1.00      0.86         3
       drugC       1.00      1.00      1.00         4
       drugX       1.00      1.00      1.00        18
       drugY       1.00      0.97      0.98        30

    accuracy                           0.98        60
   macro avg       0.95      0.99      0.97        60
weighted avg       0.99      0.98      0.98        60

[[ 5  0  0  0  0]
 [ 0  3  0  0  0]
 [ 0  0  4  0  0]
 [ 0  0  0 18  0]
 [ 0  1  0  0 29]]
SVC accuracy is: 98.33%
```

Fig. 16. Data output from SVM model Implementation on all data.



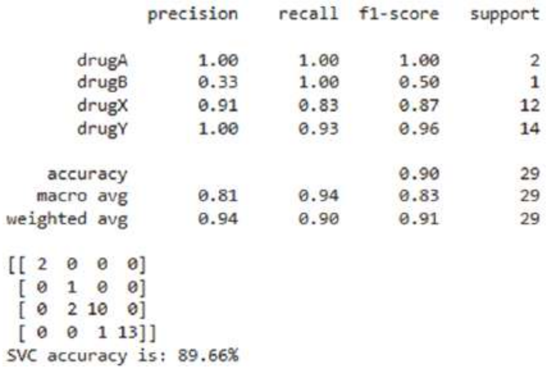
```
              precision    recall  f1-score   support

       drugA       1.00      1.00      1.00         2
       drugB       1.00      1.00      1.00         1
       drugC       1.00      1.00      1.00         3
       drugX       1.00      1.00      1.00         6
       drugY       1.00      1.00      1.00        20

    accuracy                           1.00        32
   macro avg       1.00      1.00      1.00        32
weighted avg       1.00      1.00      1.00        32

[[ 2  0  0  0  0]
 [ 0  1  0  0  0]
 [ 0  0  3  0  0]
 [ 0  0  0  6  0]
 [ 0  0  0  0 20]]
SVC accuracy is: 100.00%
```

Fig. 17. Data output from SVM model Implementation on just male data.



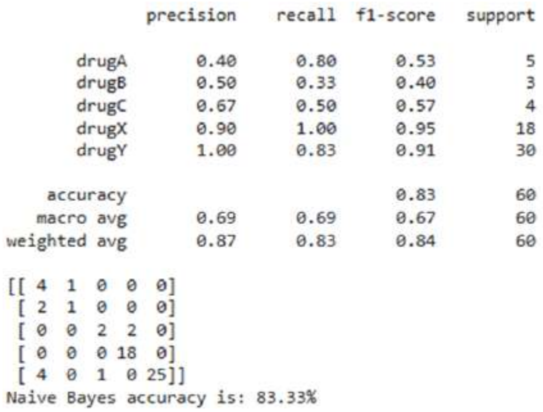
```
              precision    recall  f1-score   support

       drugA       1.00      1.00      1.00         2
       drugB       0.33      1.00      0.50         1
       drugX       0.91      0.83      0.87        12
       drugY       1.00      0.93      0.96        14

    accuracy                           0.90        29
   macro avg       0.81      0.94      0.83        29
weighted avg       0.94      0.90      0.91        29

[[ 2  0  0  0]
 [ 0  1  0  0]
 [ 0  2 10  0]
 [ 0  0  1 13]]
SVC accuracy is: 89.66%
```

Fig. 18. Data output from SVM model Implementation on just female data.



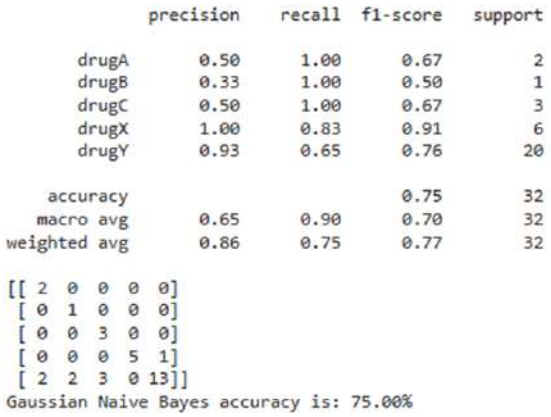
```
              precision    recall  f1-score   support

       drugA       0.40      0.80      0.53         5
       drugB       0.50      0.33      0.40         3
       drugC       0.67      0.50      0.57         4
       drugX       0.90      1.00      0.95        18
       drugY       1.00      0.83      0.91        30

    accuracy                           0.83        60
   macro avg       0.69      0.69      0.67        60
weighted avg       0.87      0.83      0.84        60

[[ 4  1  0  0  0]
 [ 2  1  0  0  0]
 [ 0  0  2  2  0]
 [ 0  0  0 18  0]
 [ 4  0  1  0 25]]
Naive Bayes accuracy is: 83.33%
```

Fig. 19. Data output from Naïve Bayes model Implementation on all data.



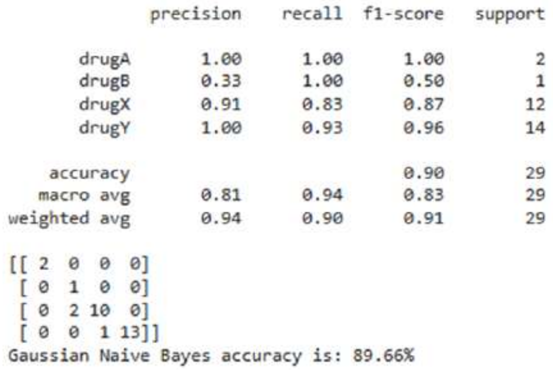
```
              precision    recall  f1-score   support

       drugA       0.50      1.00      0.67         2
       drugB       0.33      1.00      0.50         1
       drugC       0.50      1.00      0.67         3
       drugX       1.00      0.83      0.91         6
       drugY       0.93      0.65      0.76        20

    accuracy                           0.75        32
   macro avg       0.65      0.90      0.70        32
weighted avg       0.86      0.75      0.77        32

[[ 2  0  0  0  0]
 [ 0  1  0  0  0]
 [ 0  0  3  0  0]
 [ 0  0  0  5  1]
 [ 2  2  3  0 13]]
Gaussian Naive Bayes accuracy is: 75.00%
```

Fig. 20. Data output from Naïve Bayes model Implementation on just male data.



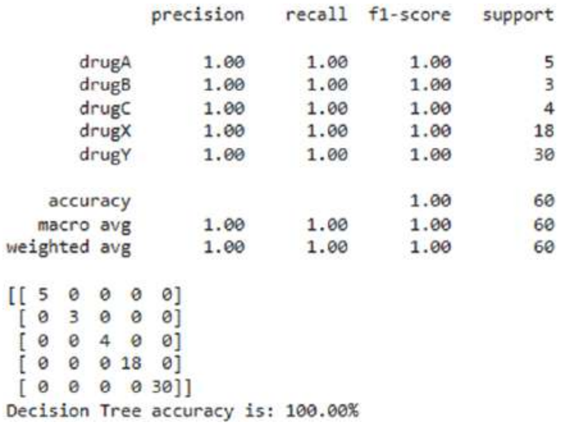
```
              precision    recall  f1-score   support

       drugA       1.00      1.00      1.00         2
       drugB       0.33      1.00      0.50         1
       drugX       0.91      0.83      0.87        12
       drugY       1.00      0.93      0.96        14

    accuracy                           0.90        29
   macro avg       0.81      0.94      0.83        29
weighted avg       0.94      0.90      0.91        29

[[ 2  0  0  0]
 [ 0  1  0  0]
 [ 0  2 10  0]
 [ 0  0  1 13]]
Gaussian Naive Bayes accuracy is: 89.66%
```

Fig. 21. Data output from Naïve Bayes model Implementation on just female data.



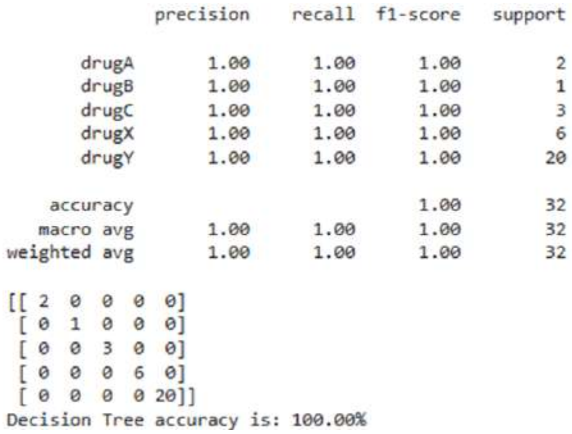
```
              precision    recall  f1-score   support

       drugA       1.00      1.00      1.00         5
       drugB       1.00      1.00      1.00         3
       drugC       1.00      1.00      1.00         4
       drugX       1.00      1.00      1.00        18
       drugY       1.00      1.00      1.00        30

    accuracy                           1.00        60
   macro avg       1.00      1.00      1.00        60
weighted avg       1.00      1.00      1.00        60

[[ 5  0  0  0  0]
 [ 0  3  0  0  0]
 [ 0  0  4  0  0]
 [ 0  0  0 18  0]
 [ 0  0  0  0 30]]
Decision Tree accuracy is: 100.00%
```

Fig. 22. Data output from Naïve Bayes model Implementation on all data.



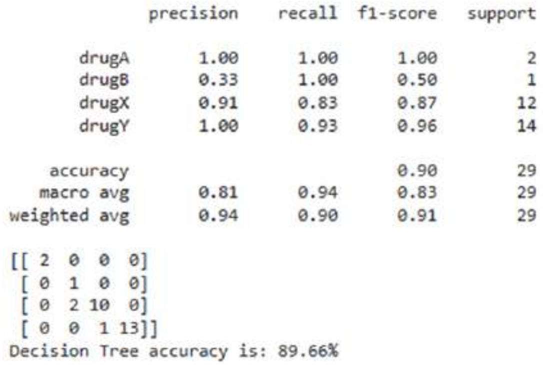
```
              precision    recall  f1-score   support

       drugA       1.00      1.00      1.00         2
       drugB       1.00      1.00      1.00         1
       drugC       1.00      1.00      1.00         3
       drugX       1.00      1.00      1.00         6
       drugY       1.00      1.00      1.00        20

    accuracy                           1.00        32
   macro avg       1.00      1.00      1.00        32
weighted avg       1.00      1.00      1.00        32

[[ 2  0  0  0  0]
 [ 0  1  0  0  0]
 [ 0  0  3  0  0]
 [ 0  0  0  6  0]
 [ 0  0  0  0 20]]
Decision Tree accuracy is: 100.00%
```

Fig. 23. Data output from Naïve Bayes model Implementation on just male data.



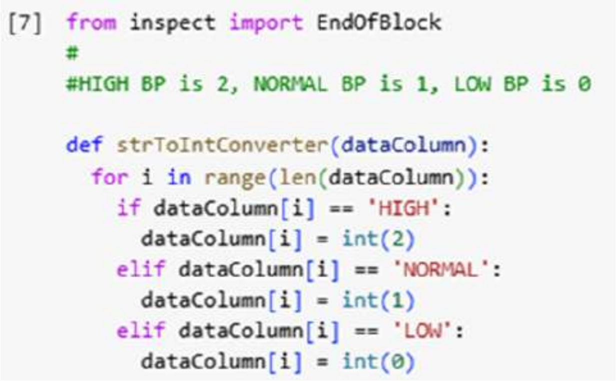
```
              precision    recall  f1-score   support

       drugA       1.00      1.00      1.00         2
       drugB       0.33      1.00      0.50         1
       drugX       0.91      0.83      0.87        12
       drugY       1.00      0.93      0.96        14

    accuracy                           0.90        29
   macro avg       0.81      0.94      0.83        29
weighted avg       0.94      0.90      0.91        29

[[ 2  0  0  0]
 [ 0  1  0  0]
 [ 0  2 10  0]
 [ 0  0  1 13]]
Decision Tree accuracy is: 89.66%
```

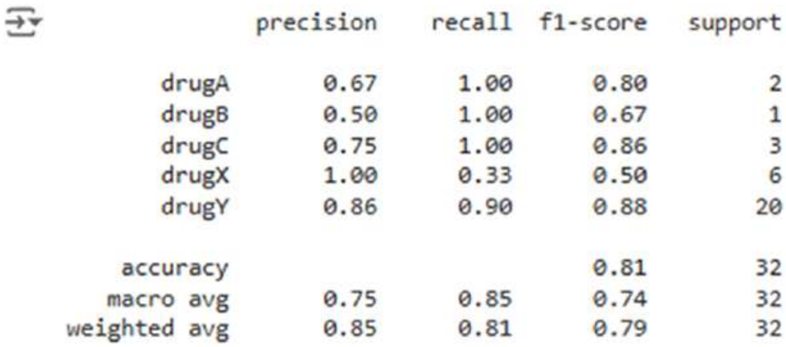
Fig. 24. Data output from Naïve Bayes model Implementation on just female data.



```
[7] from inspect import EndOfBlock
    #
    #HIGH BP is 2, NORMAL BP is 1, LOW BP is 0

    def strToIntConverter(dataColumn):
      for i in range(len(dataColumn)):
        if dataColumn[i] == 'HIGH':
          dataColumn[i] = int(2)
        elif dataColumn[i] == 'NORMAL':
          dataColumn[i] = int(1)
        elif dataColumn[i] == 'LOW':
          dataColumn[i] = int(0)
```

Fig. 25. The segment of code that allows the user to overwrite string values in a list with integer values. The name of the function is called “strToIntConverter”.

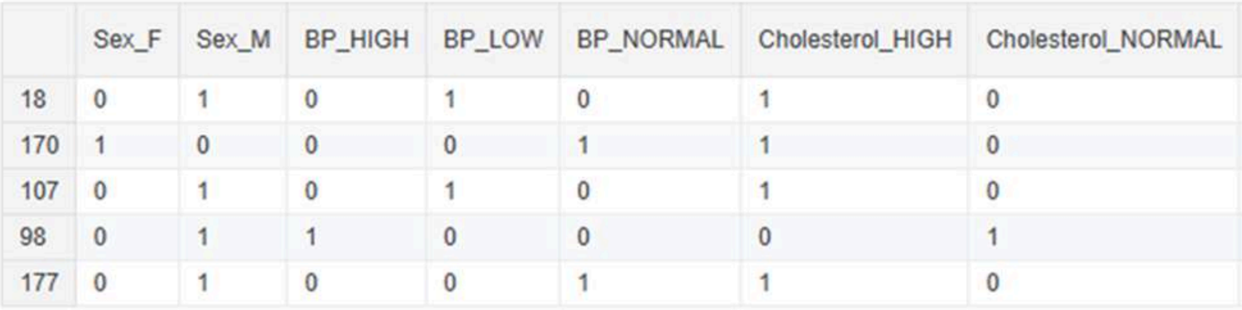


```
              precision    recall  f1-score   support

       drugA       0.67      1.00      0.80         2
       drugB       0.50      1.00      0.67         1
       drugC       0.75      1.00      0.86         3
       drugX       1.00      0.33      0.50         6
       drugY       0.86      0.90      0.88        20

    accuracy                           0.81        32
   macro avg       0.75      0.85      0.74        32
weighted avg       0.85      0.81      0.79        32
```

Fig. 26. Example of an output from a ML model.



	Sex_F	Sex_M	BP_HIGH	BP_LOW	BP_NORMAL	Cholesterol_HIGH	Cholesterol_NORMAL
18	0	1	0	1	0	1	0
170	1	0	0	0	1	1	0
107	0	1	0	1	0	1	0
98	0	1	1	0	0	0	1
177	0	1	0	0	1	1	0

Fig. 27. A table showing the output of a proposed way to sort the strings into integer values.

References

1. Medlinskiene, Kristina, et al. "Barriers and facilitators to the uptake of new medicines into clinical practice: a systematic review." *BMC health services research* 21.1 (2021): 1198.
2. Sae-Ang, Apichat, et al. "Drug recommendation from diagnosis codes: Classification vs. Collaborative filtering approaches." *International Journal of Environmental Research and Public Health* 20.1 (2022): 309.
3. Kumar, Yogesh, et al. "Artificial intelligence in disease diagnosis: a systematic literature review, synthesizing framework and future research agenda." *Journal of ambient intelligence and humanized computing* 14.7 (2023): 8459-8486.
4. Karalis, Vangelis D. "The integration of artificial intelligence into clinical practice." *Applied Biosciences* 3.1 (2024): 14-44.
5. Arnold, Arthur P., et al. "Male–female comparisons are powerful in biomedical research—don't abandon them." *Nature* 629.8010 (2024): 37-40.
6. d'Emden, Michael C., et al. "Favourable effects of fenofibrate on lipids and cardiovascular disease in women with type 2 diabetes: results from the Fenofibrate Intervention and Event Lowering in Diabetes (FIELD) study." *Diabetologia* 57 (2014): 2296-2303.
7. Guido, Rosita, et al. "An Overview on the Advancements of Support Vector Machine Models in Healthcare Applications: A Review." *Information* 15.4 (2024): 235.
8. Sarang, Poornachandra. "Support Vector Machines: A Supervised Learning Algorithm for Classification and Regression." *Thinking Data Science: A Data Science Practitioner's Guide*. Cham: Springer International Publishing, 2023. 153-165.
9. Habehh, Hafsa, and Suril Gohel. "Machine Learning in Healthcare." *Current genomics* vol. 22,4 (2021): 291-300. doi:10.2174/1389202922666210705124359