Machine Learning Final Project: Comparing Performance of Different Algorithms with Different Datasets

CS-620-50

Dr. Ling Zheng

Presented By Jason McCann

Graduate Student at Monmouth University

Introduction

- Many algorithms are used to attempt to improve performance when evaluating a dataset
- Ex. Logistic regression and K-nearest neighbor
- Goal of the project was to compare the performance of four algorithms discussed in class, with an additional undiscussed algorithm, when evaluated against five different datasets

Algorithms Used

- Logistic Regression
- K-Nearest Neighbor
- Decision Tree
- Random Forest
- Bernoulli Naïve Bayes

Bernoulli Naïve Bayes Algorithm

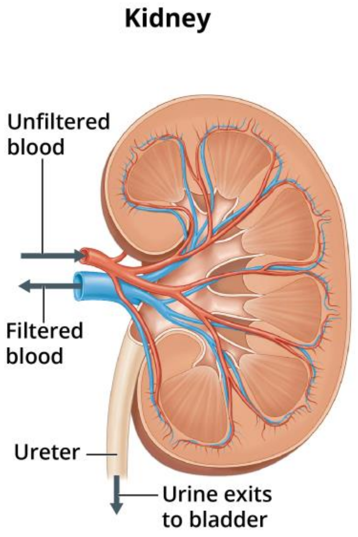
- Derivative of the Naïve Bayes Algorithm
  - Based on Bayes theorem, a mathematical formula which predicts an events probability by utilizing prior knowledge from similar conditions to that of the event

$$P(A|B) = \frac{P(B|A)\ P(A)}{P(B)}$$

  - When given input, probability is predicted for the input being classified for all classes
  - Known as naïve as it treats all attributes as independent of each other and of equal importance
- Great for evaluating datasets with discrete data that is in binary form

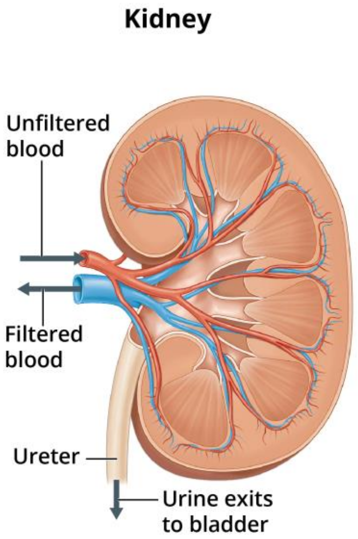
Dataset 1: Kidney Disease Prediction

Feature	Description
Bp	Patient’s blood pressure
Sg	Patient’s specific gravity
Al	Patient’s albumin level
Su	Patient's sugar reading
Rbc	Red blood cell (0 for no and 1 for yes)
Bu	Patient’s blood urea level
Sc	Patient’s serum creatinine level



Dataset 1 Continued

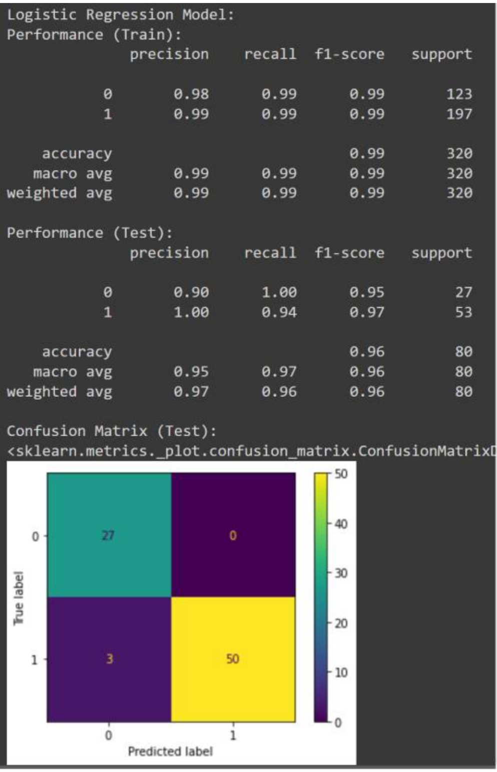
Features	Description
Sod	Patient's sodium level
Pot	Patient's potassium level
Hemo	Patient's hemoglobin level
Wbcc	Patient's white blood cell count
Rbcc	Patient's red blood cell count
Htn	If the patient has hypertension or not (0 for no and 1 for yes)
Class	Whether a patient has chronic kidney disease or not (0 being no and 1 being yes)



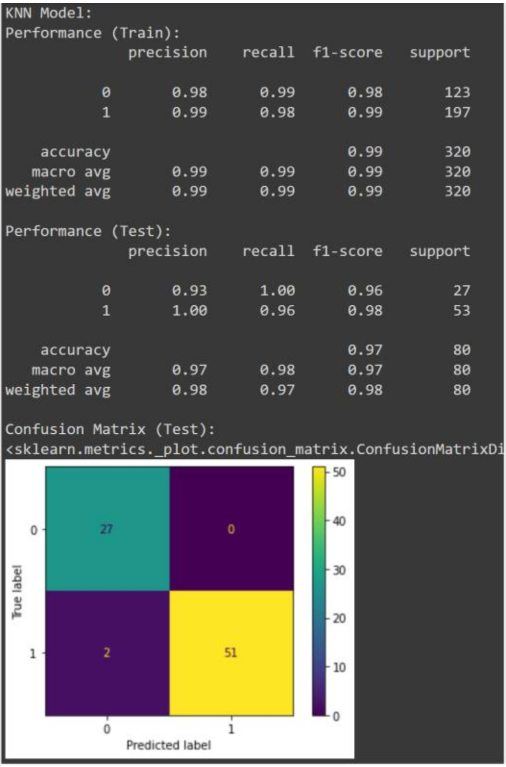
Colab Demo

- Look at dataset and code on Google Colab

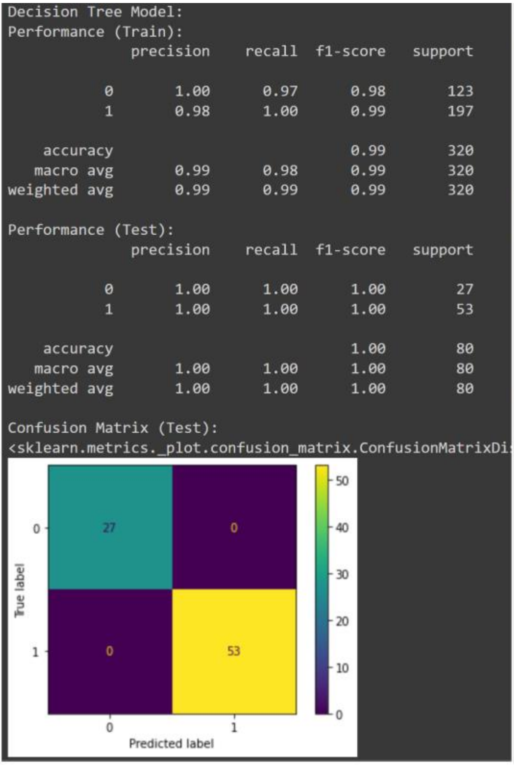
Results



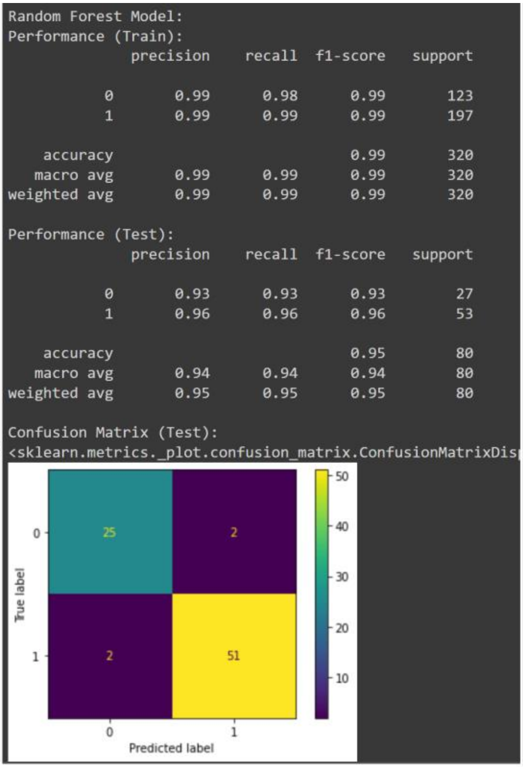
```
Logistic Regression Model:
Performance (Train):
              precision    recall  f1-score   support

           0       0.98      0.99      0.99       123
           1       0.99      0.99      0.99       197

    accuracy                           0.99       320
   macro avg       0.99      0.99      0.99       320
weighted avg       0.99      0.99      0.99       320

Performance (Test):
              precision    recall  f1-score   support

           0       0.90      1.00      0.95        27
           1       1.00      0.94      0.97        53

    accuracy                           0.96        80
   macro avg       0.95      0.97      0.96        80
weighted avg       0.97      0.96      0.96        80

Confusion Matrix (Test):
<sklearn.metrics._plot.confusion_matrix.ConfusionMatrixD
```



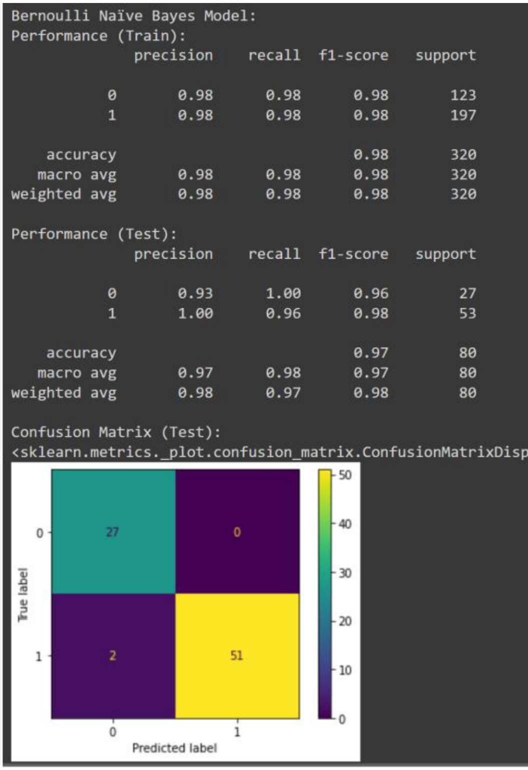
```
KNN Model:
Performance (Train):
              precision    recall  f1-score   support

           0       0.98      0.99      0.98       123
           1       0.99      0.98      0.99       197

    accuracy                           0.99       320
   macro avg       0.99      0.99      0.99       320
weighted avg       0.99      0.99      0.99       320

Performance (Test):
              precision    recall  f1-score   support

           0       0.93      1.00      0.96        27
           1       1.00      0.96      0.98        53

    accuracy                           0.97        80
   macro avg       0.97      0.98      0.97        80
weighted avg       0.98      0.97      0.98        80

Confusion Matrix (Test):
<sklearn.metrics._plot.confusion_matrix.ConfusionMatrixDi
```



```
Decision Tree Model:
Performance (Train):
              precision    recall  f1-score   support

           0       1.00      0.97      0.98       123
           1       0.98      1.00      0.99       197

    accuracy                           0.99       320
   macro avg       0.99      0.98      0.99       320
weighted avg       0.99      0.99      0.99       320

Performance (Test):
              precision    recall  f1-score   support

           0       1.00      1.00      1.00        27
           1       1.00      1.00      1.00        53

    accuracy                           1.00        80
   macro avg       1.00      1.00      1.00        80
weighted avg       1.00      1.00      1.00        80

Confusion Matrix (Test):
<sklearn.metrics._plot.confusion_matrix.ConfusionMatrixDi
```



```
Random Forest Model:
Performance (Train):
              precision    recall  f1-score   support

           0       0.99      0.98      0.99       123
           1       0.99      0.99      0.99       197

    accuracy                           0.99       320
   macro avg       0.99      0.99      0.99       320
weighted avg       0.99      0.99      0.99       320

Performance (Test):
              precision    recall  f1-score   support

           0       0.93      0.93      0.93        27
           1       0.96      0.96      0.96        53

    accuracy                           0.95        80
   macro avg       0.94      0.94      0.94        80
weighted avg       0.95      0.95      0.95        80

Confusion Matrix (Test):
<sklearn.metrics._plot.confusion_matrix.ConfusionMatrixDis
```



```
Bernoulli Naïve Bayes Model:
Performance (Train):
              precision    recall  f1-score   support

           0       0.98      0.98      0.98       123
           1       0.98      0.98      0.98       197

    accuracy                           0.98       320
   macro avg       0.98      0.98      0.98       320
weighted avg       0.98      0.98      0.98       320

Performance (Test):
              precision    recall  f1-score   support

           0       0.93      1.00      0.96        27
           1       1.00      0.96      0.98        53

    accuracy                           0.97        80
   macro avg       0.97      0.98      0.97        80
weighted avg       0.98      0.97      0.98        80

Confusion Matrix (Test):
<sklearn.metrics._plot.confusion_matrix.ConfusionMatrixDisp
```

Dataset 2: Employee's Future Prediction

Feature	Description
Educations	The employee’s education level
JoiningYear	The year the employee joined the company
City	The city of where the office the employee works at is located
PaymentTier	Numerical categorical placement of what payment tier the employee is in (1: Highest, 2: Mid-Level, 3: Lowest)
Age	Current age of the employee
Gender	Employee’s gender
EverBenched	Yes or no to if the employee has ever been kept out of a project for a month or more

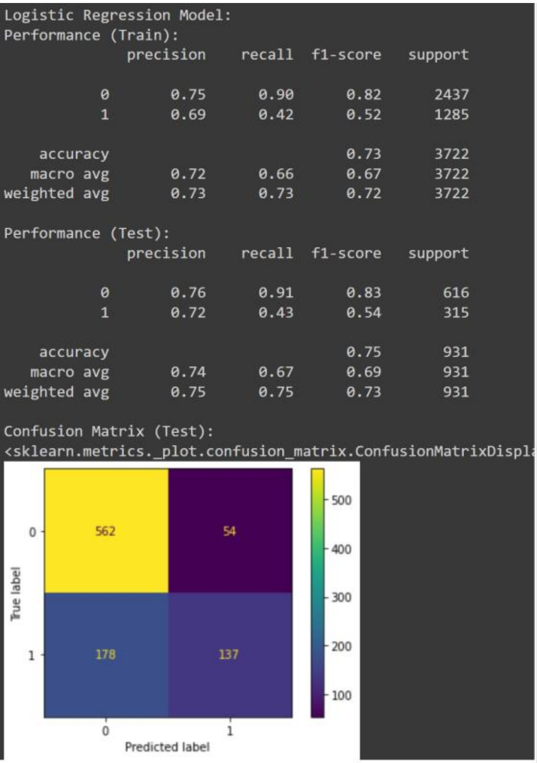
Dataset 2 Continued

Feature	Description
ExperienceInCurrentDomain	Employee’s experience in their current field
LeaveOrNot	Describes whether an employee will leave the company in the next two years or stay (0 being no and 1 being yes)

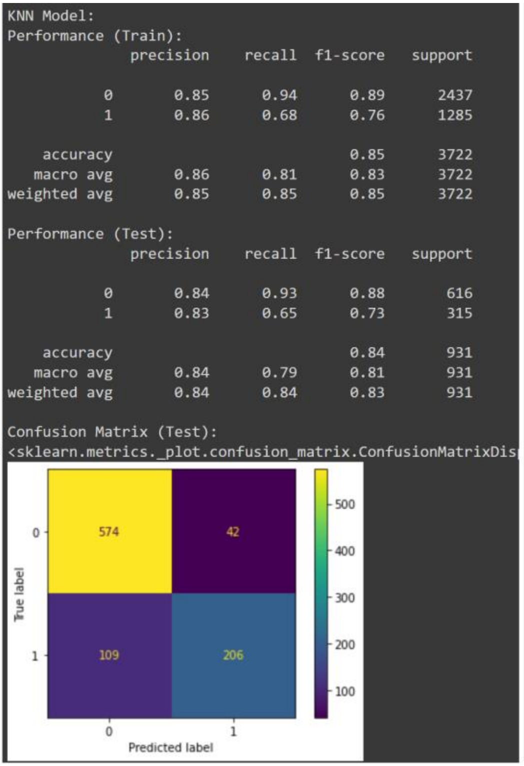
Colab Demo

- Look at dataset and code on Google Colab

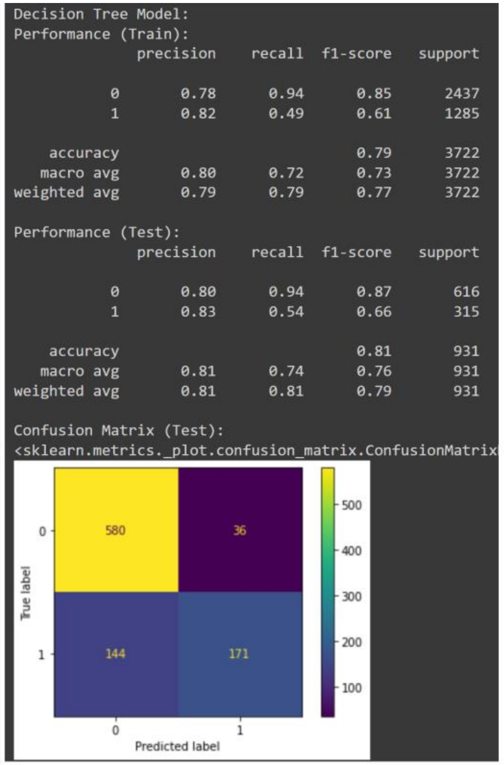
Results



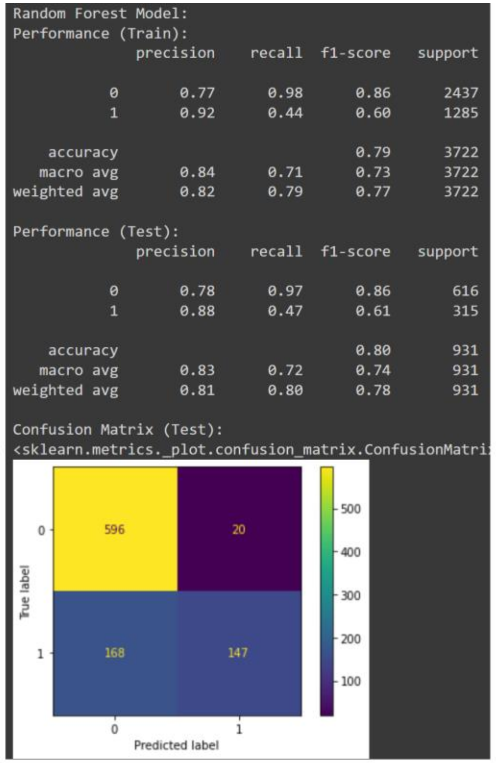
```
Logistic Regression Model:
Performance (Train):
              precision    recall  f1-score   support

           0       0.75      0.90      0.82      2437
           1       0.69      0.42      0.52      1285

    accuracy                           0.73      3722
   macro avg       0.72      0.66      0.67      3722
weighted avg       0.73      0.73      0.72      3722

Performance (Test):
              precision    recall  f1-score   support

           0       0.76      0.91      0.83       616
           1       0.72      0.43      0.54       315

    accuracy                           0.75       931
   macro avg       0.74      0.67      0.69       931
weighted avg       0.75      0.75      0.73       931

Confusion Matrix (Test):
<sklearn.metrics._plot.confusion_matrix.ConfusionMatrixDispla
```



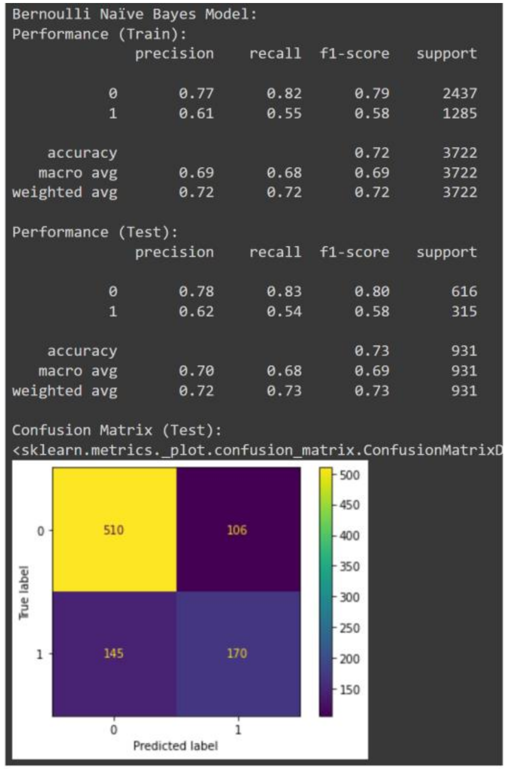
```
KNN Model:
Performance (Train):
              precision    recall  f1-score   support

           0       0.85      0.94      0.89      2437
           1       0.86      0.68      0.76      1285

    accuracy                           0.85      3722
   macro avg       0.86      0.81      0.83      3722
weighted avg       0.85      0.85      0.85      3722

Performance (Test):
              precision    recall  f1-score   support

           0       0.84      0.93      0.88       616
           1       0.83      0.65      0.73       315

    accuracy                           0.84       931
   macro avg       0.84      0.79      0.81       931
weighted avg       0.84      0.84      0.83       931

Confusion Matrix (Test):
<sklearn.metrics._plot.confusion_matrix.ConfusionMatrixDis
```



```
Decision Tree Model:
Performance (Train):
              precision    recall  f1-score   support

           0       0.78      0.94      0.85      2437
           1       0.82      0.49      0.61      1285

    accuracy                           0.79      3722
   macro avg       0.80      0.72      0.73      3722
weighted avg       0.79      0.79      0.77      3722

Performance (Test):
              precision    recall  f1-score   support

           0       0.80      0.94      0.87       616
           1       0.83      0.54      0.66       315

    accuracy                           0.81       931
   macro avg       0.81      0.74      0.76       931
weighted avg       0.81      0.81      0.79       931

Confusion Matrix (Test):
<sklearn.metrics._plot.confusion_matrix.ConfusionMatrix
```



```
Random Forest Model:
Performance (Train):
              precision    recall  f1-score   support

           0       0.77      0.98      0.86      2437
           1       0.92      0.44      0.60      1285

    accuracy                           0.79      3722
   macro avg       0.84      0.71      0.73      3722
weighted avg       0.82      0.79      0.77      3722

Performance (Test):
              precision    recall  f1-score   support

           0       0.78      0.97      0.86       616
           1       0.88      0.47      0.61       315

    accuracy                           0.80       931
   macro avg       0.83      0.72      0.74       931
weighted avg       0.81      0.80      0.78       931

Confusion Matrix (Test):
<sklearn.metrics._plot.confusion_matrix.ConfusionMatri
```



```
Bernoulli Naïve Bayes Model:
Performance (Train):
              precision    recall  f1-score   support

           0       0.77      0.82      0.79      2437
           1       0.61      0.55      0.58      1285

    accuracy                           0.72      3722
   macro avg       0.69      0.68      0.69      3722
weighted avg       0.72      0.72      0.72      3722

Performance (Test):
              precision    recall  f1-score   support

           0       0.78      0.83      0.80       616
           1       0.62      0.54      0.58       315

    accuracy                           0.73       931
   macro avg       0.70      0.68      0.69       931
weighted avg       0.72      0.73      0.73       931

Confusion Matrix (Test):
<sklearn.metrics._plot.confusion_matrix.ConfusionMatrixD
```

Dataset 3: Diabetes Prediction

Feature	Description
Pregnancies	The number of pregnancies the patient has had
Glucose	The glucose level in the blood of the patient
BloodPressure	The blood pressure measurement of the patient
SkinThickness	The thickness of the patients skin
Insulin	The insulin level in the blood of the patient
BMI	The body mass index of the patient
DiabetesPedigreeFunction	The patients Diabetes percentage

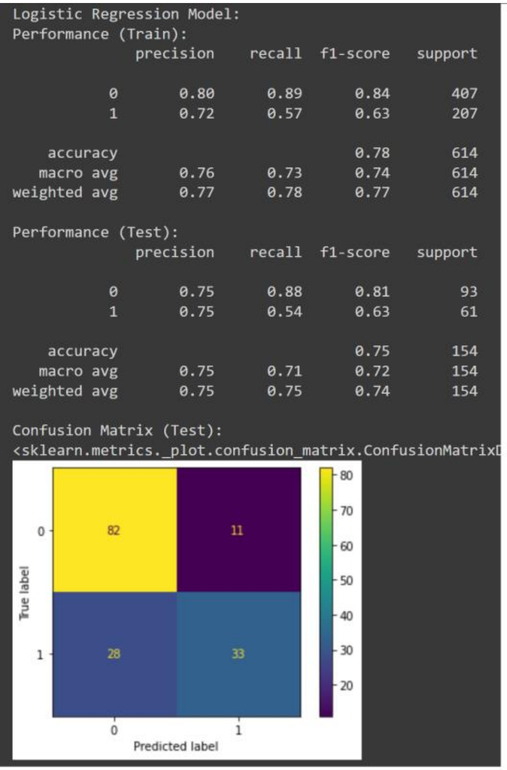
Dataset 3 Continued

Feature	Description
Age	The age of the patient
Outcome	Describes if the patient has diabetes or not (0 being no and 1 being yes)

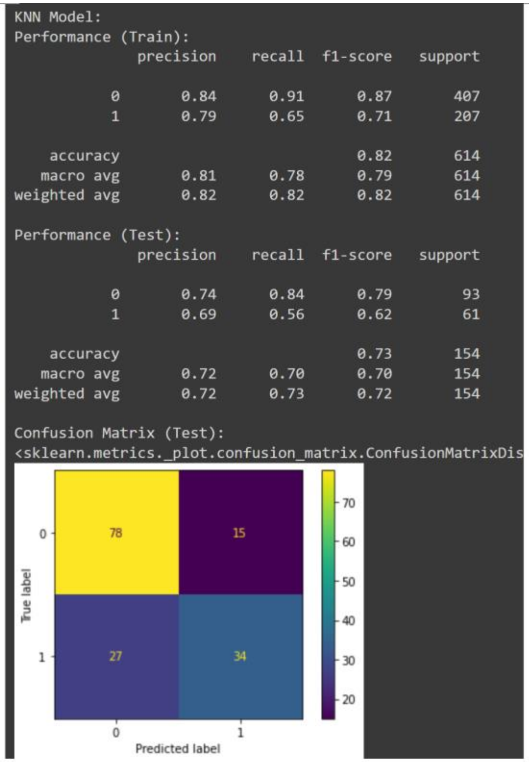
Colab Demo

- Look at dataset and code on Google Colab

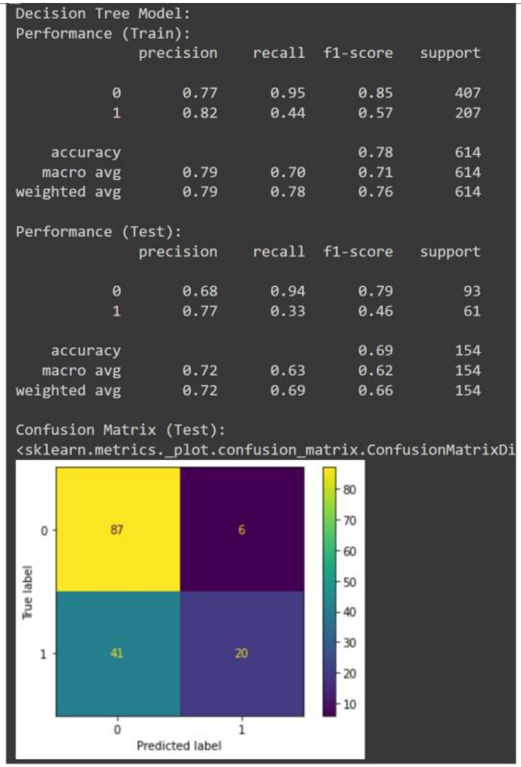
Results



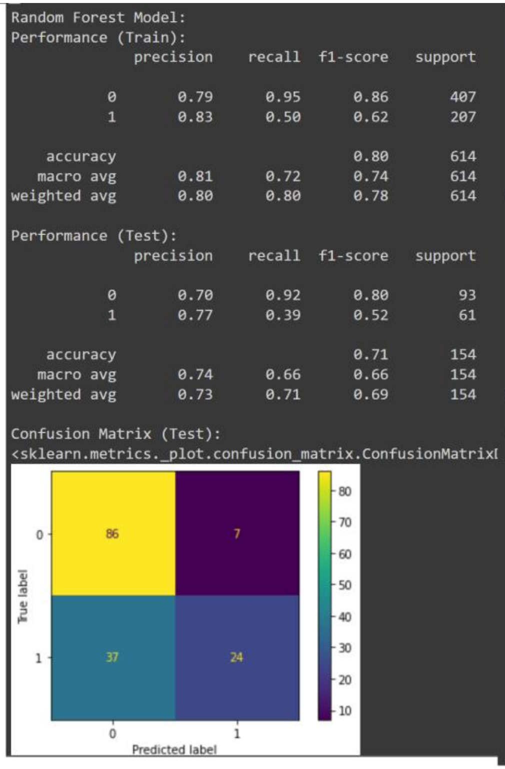
```
Logistic Regression Model:
Performance (Train):
              precision    recall  f1-score   support

           0       0.80      0.89      0.84       407
           1       0.72      0.57      0.63       207

    accuracy                           0.78       614
   macro avg       0.76      0.73      0.74       614
weighted avg       0.77      0.78      0.77       614

Performance (Test):
              precision    recall  f1-score   support

           0       0.75      0.88      0.81        93
           1       0.75      0.54      0.63        61

    accuracy                           0.75       154
   macro avg       0.75      0.71      0.72       154
weighted avg       0.75      0.75      0.74       154

Confusion Matrix (Test):
<sklearn.metrics._plot.confusion_matrix.ConfusionMatrixD
```



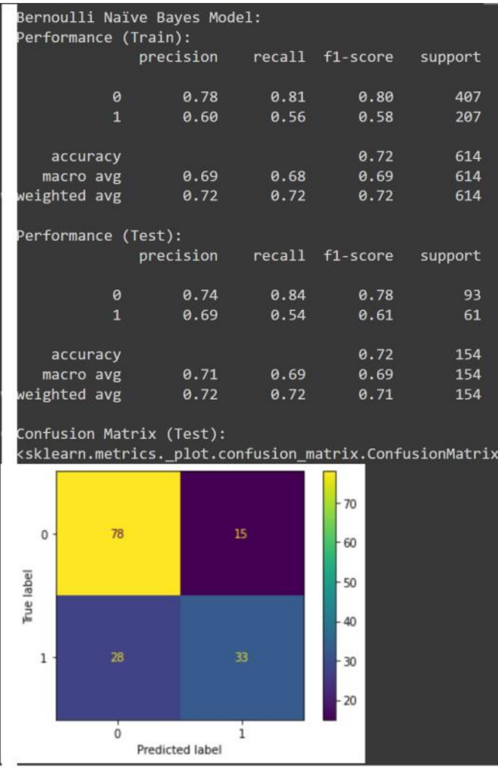
```
KNN Model:
Performance (Train):
              precision    recall  f1-score   support

           0       0.84      0.91      0.87       407
           1       0.79      0.65      0.71       207

    accuracy                           0.82       614
   macro avg       0.81      0.78      0.79       614
weighted avg       0.82      0.82      0.82       614

Performance (Test):
              precision    recall  f1-score   support

           0       0.74      0.84      0.79        93
           1       0.69      0.56      0.62        61

    accuracy                           0.73       154
   macro avg       0.72      0.70      0.70       154
weighted avg       0.72      0.73      0.72       154

Confusion Matrix (Test):
<sklearn.metrics._plot.confusion_matrix.ConfusionMatrixDis
```



```
Decision Tree Model:
Performance (Train):
              precision    recall  f1-score   support

           0       0.77      0.95      0.85       407
           1       0.82      0.44      0.57       207

    accuracy                           0.78       614
   macro avg       0.79      0.70      0.71       614
weighted avg       0.79      0.78      0.76       614

Performance (Test):
              precision    recall  f1-score   support

           0       0.68      0.94      0.79        93
           1       0.77      0.33      0.46        61

    accuracy                           0.69       154
   macro avg       0.72      0.63      0.62       154
weighted avg       0.72      0.69      0.66       154

Confusion Matrix (Test):
<sklearn.metrics._plot.confusion_matrix.ConfusionMatrixDi
```



```
Random Forest Model:
Performance (Train):
              precision    recall  f1-score   support

           0       0.79      0.95      0.86       407
           1       0.83      0.50      0.62       207

    accuracy                           0.80       614
   macro avg       0.81      0.72      0.74       614
weighted avg       0.80      0.80      0.78       614

Performance (Test):
              precision    recall  f1-score   support

           0       0.70      0.92      0.80        93
           1       0.77      0.39      0.52        61

    accuracy                           0.71       154
   macro avg       0.74      0.66      0.66       154
weighted avg       0.73      0.71      0.69       154

Confusion Matrix (Test):
<sklearn.metrics._plot.confusion_matrix.ConfusionMatrix[
```



```
Bernoulli Naïve Bayes Model:
Performance (Train):
              precision    recall  f1-score   support

           0       0.78      0.81      0.80       407
           1       0.60      0.56      0.58       207

    accuracy                           0.72       614
   macro avg       0.69      0.68      0.69       614
weighted avg       0.72      0.72      0.72       614

Performance (Test):
              precision    recall  f1-score   support

           0       0.74      0.84      0.78        93
           1       0.69      0.54      0.61        61

    accuracy                           0.72       154
   macro avg       0.71      0.69      0.69       154
weighted avg       0.72      0.72      0.71       154

Confusion Matrix (Test):
<sklearn.metrics._plot.confusion_matrix.ConfusionMatrix
```

Dataset 4: Students Going to College Prediction

Feature	Description
type_school	Whether the student currently goes to a academic or vocational school
school_accreditation	Quality of the school the student is attending, a grade of A is better than B
gender	The students gender
interest	The interest level the student has in college
residence	If the student lives in a rural or urban type community
parent_age	Age of the student's parents
parent_salary	The per month salary of the student's parents (in IDR/Rupiah)

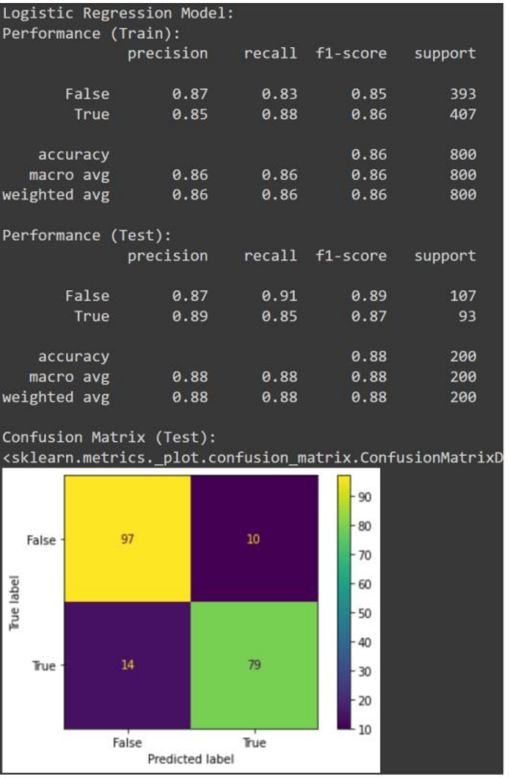
Dataset 4 Continued

Feature	Description
house_area	The area of the student's parent’s house in meters squared
average_grades	Student’s average grade on the 0-100 scale
parent_was_in_college	If the student’s parent attended college (true or false)
will_go_to_college	If the student will go to college or not (False being no and True being yes)

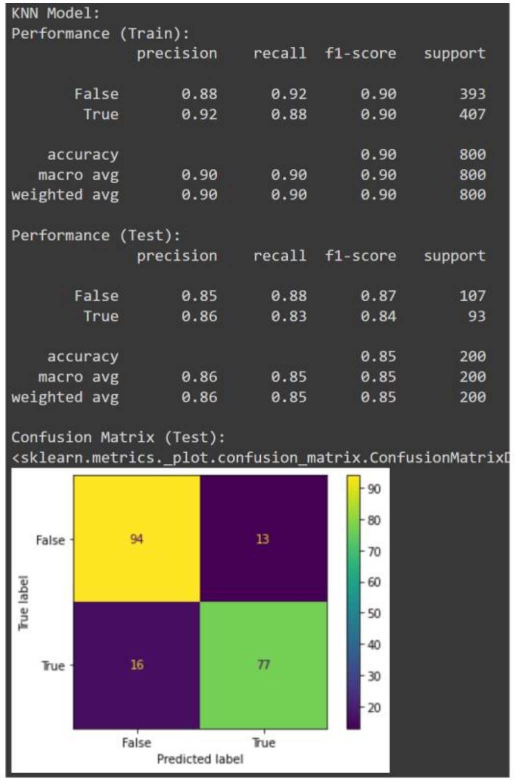
Colab Demo

- Look at dataset and code on Google Colab

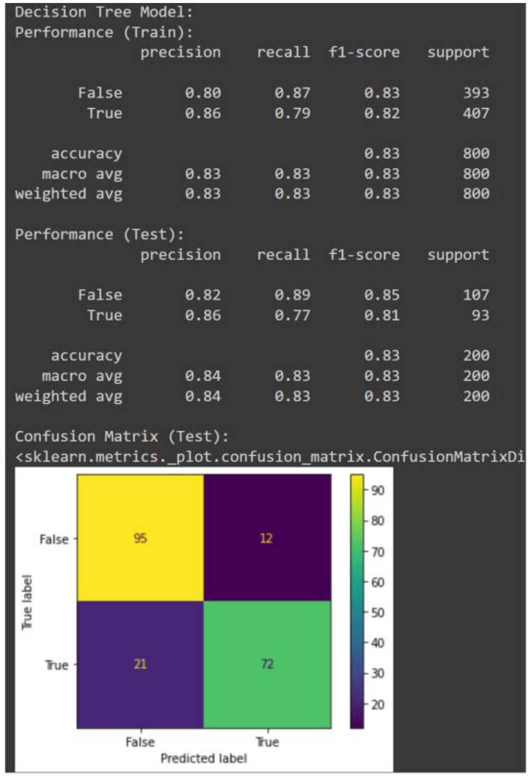
Results



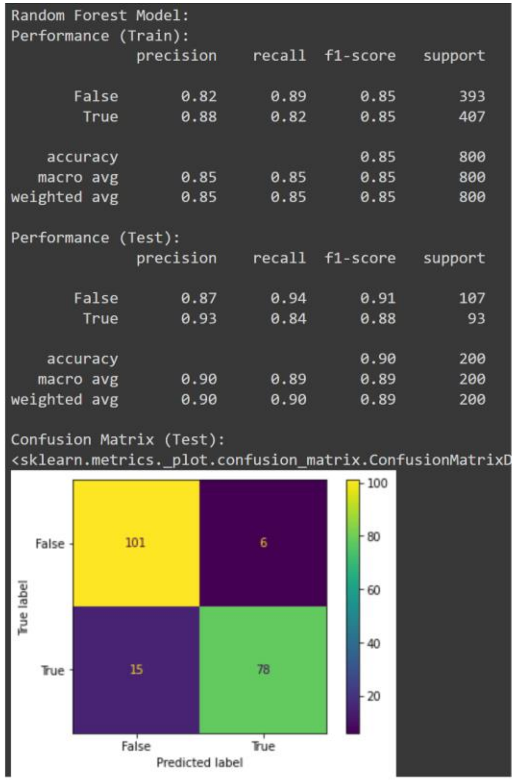
```
Logistic Regression Model:
Performance (Train):
              precision    recall  f1-score   support

       False       0.87      0.83      0.85       393
        True       0.85      0.88      0.86       407

    accuracy                           0.86       800
   macro avg       0.86      0.86      0.86       800
weighted avg       0.86      0.86      0.86       800

Performance (Test):
              precision    recall  f1-score   support

       False       0.87      0.91      0.89       107
        True       0.89      0.85      0.87        93

    accuracy                           0.88       200
   macro avg       0.88      0.88      0.88       200
weighted avg       0.88      0.88      0.88       200

Confusion Matrix (Test):
<sklearn.metrics._plot.confusion_matrix.ConfusionMatrixD
```



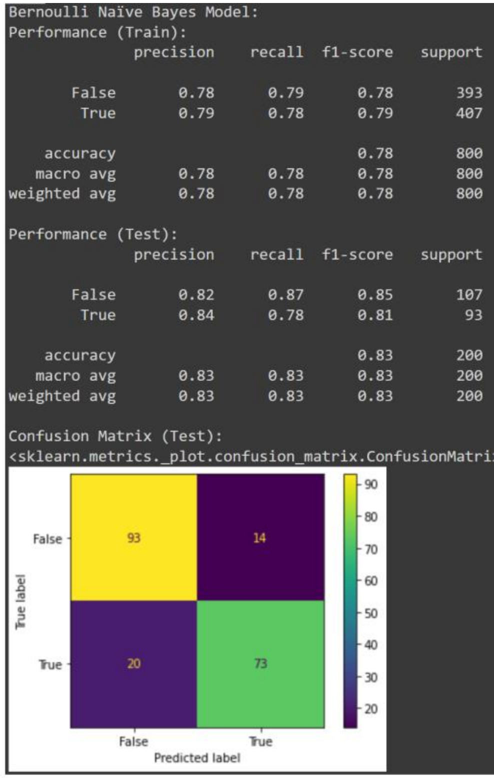
```
KNN Model:
Performance (Train):
              precision    recall  f1-score   support

       False       0.88      0.92      0.90       393
        True       0.92      0.88      0.90       407

    accuracy                           0.90       800
   macro avg       0.90      0.90      0.90       800
weighted avg       0.90      0.90      0.90       800

Performance (Test):
              precision    recall  f1-score   support

       False       0.85      0.88      0.87       107
        True       0.86      0.83      0.84        93

    accuracy                           0.85       200
   macro avg       0.86      0.85      0.85       200
weighted avg       0.86      0.85      0.85       200

Confusion Matrix (Test):
<sklearn.metrics._plot.confusion_matrix.ConfusionMatrixD
```



```
Decision Tree Model:
Performance (Train):
              precision    recall  f1-score   support

       False       0.80      0.87      0.83       393
        True       0.86      0.79      0.82       407

    accuracy                           0.83       800
   macro avg       0.83      0.83      0.83       800
weighted avg       0.83      0.83      0.83       800

Performance (Test):
              precision    recall  f1-score   support

       False       0.82      0.89      0.85       107
        True       0.86      0.77      0.81        93

    accuracy                           0.83       200
   macro avg       0.84      0.83      0.83       200
weighted avg       0.84      0.83      0.83       200

Confusion Matrix (Test):
<sklearn.metrics._plot.confusion_matrix.ConfusionMatrixDi
```



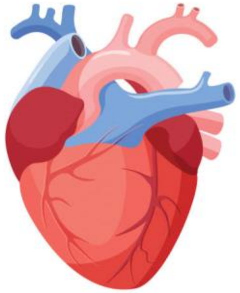
```
Random Forest Model:
Performance (Train):
              precision    recall  f1-score   support

       False       0.82      0.89      0.85       393
        True       0.88      0.82      0.85       407

    accuracy                           0.85       800
   macro avg       0.85      0.85      0.85       800
weighted avg       0.85      0.85      0.85       800

Performance (Test):
              precision    recall  f1-score   support

       False       0.87      0.94      0.91       107
        True       0.93      0.84      0.88        93

    accuracy                           0.90       200
   macro avg       0.90      0.89      0.89       200
weighted avg       0.90      0.90      0.89       200

Confusion Matrix (Test):
<sklearn.metrics._plot.confusion_matrix.ConfusionMatrixD
```



```
Bernoulli Naïve Bayes Model:
Performance (Train):
              precision    recall  f1-score   support

       False       0.78      0.79      0.78       393
        True       0.79      0.78      0.79       407

    accuracy                           0.78       800
   macro avg       0.78      0.78      0.78       800
weighted avg       0.78      0.78      0.78       800

Performance (Test):
              precision    recall  f1-score   support

       False       0.82      0.87      0.85       107
        True       0.84      0.78      0.81        93

    accuracy                           0.83       200
   macro avg       0.83      0.83      0.83       200
weighted avg       0.83      0.83      0.83       200

Confusion Matrix (Test):
<sklearn.metrics._plot.confusion_matrix.ConfusionMatri
```

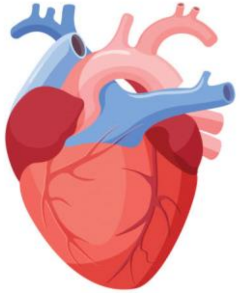
Dataset 5: Heart Attack Prediction

Feature	Description
Age	Age of the patient
Sex	The sex of the patient
exang	If exercise induced angina in the patient(0 for no and 1 for yes
ca	Number of major vessels (0-3)
cp	Patient’s chest pain level (1 for typical angina, 2 for atypical angina, 3 for non-angina pain, and 4 for asymptomatic)
trtbps	Patient’s resting blood pressure (in mm Hg)
chol	Cholesterol level in the patient (mg/dl)



Dataset 5 Continued

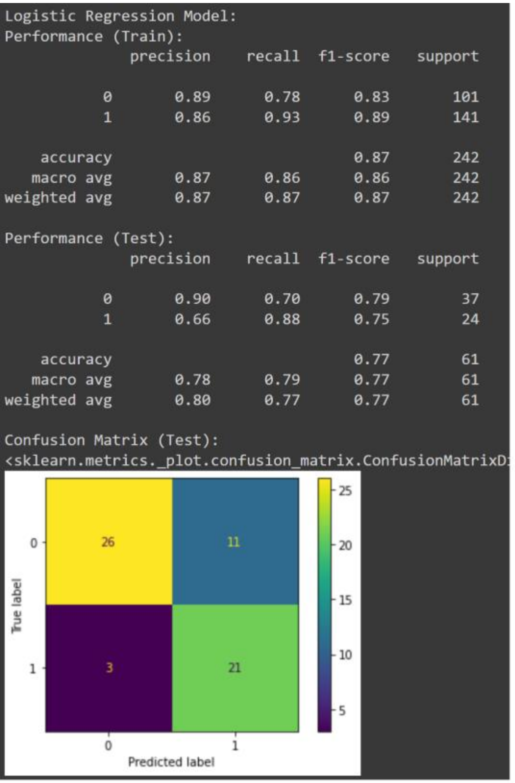
Feature	Description
fbs	Fasting blood sugar > 120 mg/dl (0 is false and 1 is true)
rest_ecg	Resting electrocardiographic results in the patient (0 is normal, 1 is having ST-T wave abnormality, 2 is showing probable or definite left hypertrophy)
thalach	The patients maximum heart rate achieved
target	Predicts if the patient will have greater chance at a heart attack or a lesser chance (0 being a less chance and 1 being a greater chance)



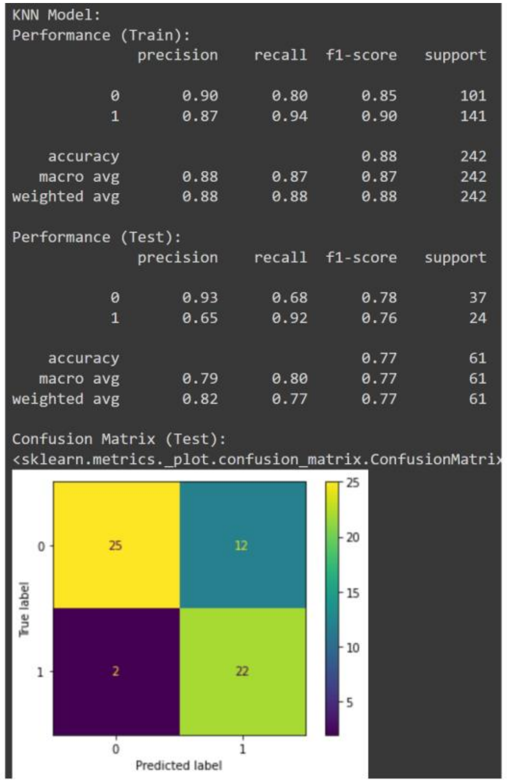
Colab Demo

- Look at dataset and code on Google Colab

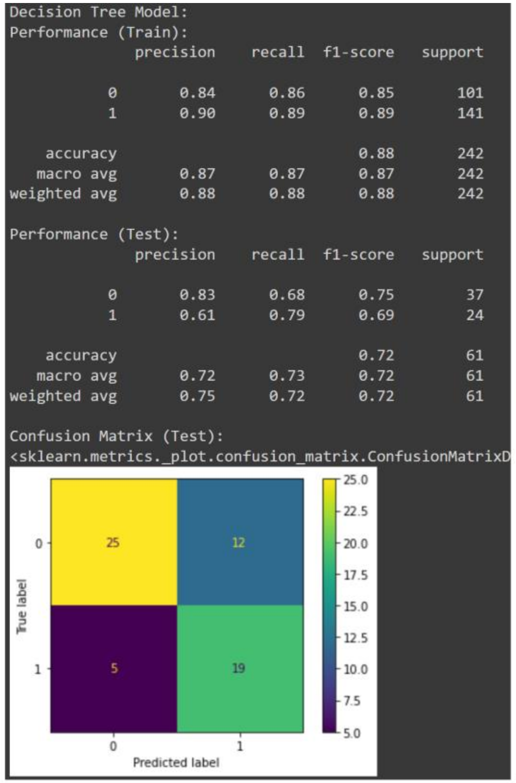
Results



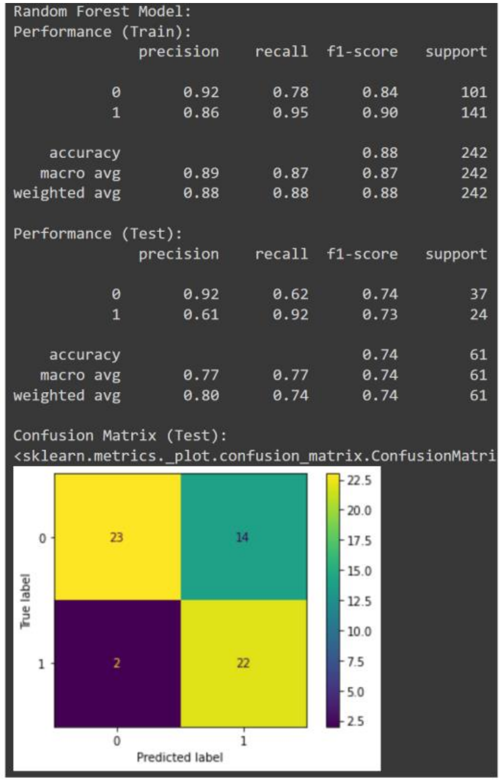
```
Logistic Regression Model:
Performance (Train):
              precision    recall  f1-score   support

           0       0.89      0.78      0.83       101
           1       0.86      0.93      0.89       141

    accuracy                           0.87       242
   macro avg       0.87      0.86      0.86       242
weighted avg       0.87      0.87      0.87       242

Performance (Test):
              precision    recall  f1-score   support

           0       0.90      0.70      0.79        37
           1       0.66      0.88      0.75        24

    accuracy                           0.77        61
   macro avg       0.78      0.79      0.77        61
weighted avg       0.80      0.77      0.77        61

Confusion Matrix (Test):
<sklearn.metrics._plot.confusion_matrix.ConfusionMatrixD
```



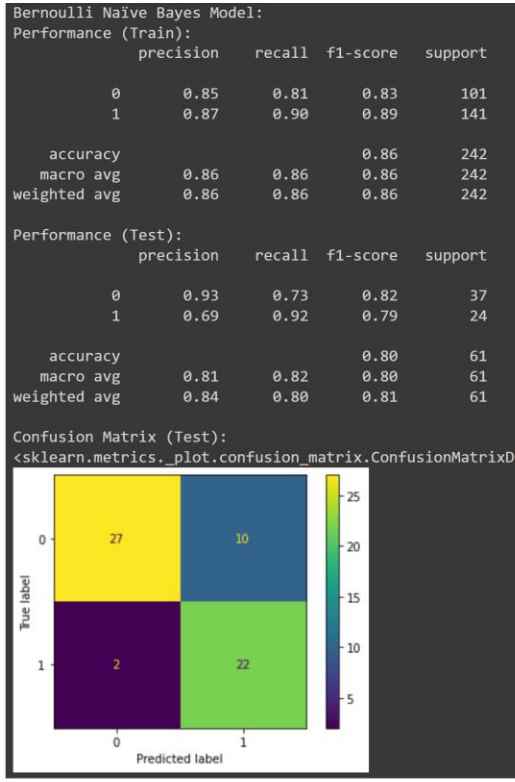
```
KNN Model:
Performance (Train):
              precision    recall  f1-score   support

           0       0.90      0.80      0.85       101
           1       0.87      0.94      0.90       141

    accuracy                           0.88       242
   macro avg       0.88      0.87      0.87       242
weighted avg       0.88      0.88      0.88       242

Performance (Test):
              precision    recall  f1-score   support

           0       0.93      0.68      0.78        37
           1       0.65      0.92      0.76        24

    accuracy                           0.77        61
   macro avg       0.79      0.80      0.77        61
weighted avg       0.82      0.77      0.77        61

Confusion Matrix (Test):
<sklearn.metrics._plot.confusion_matrix.ConfusionMatrix
```



```
Decision Tree Model:
Performance (Train):
              precision    recall  f1-score   support

           0       0.84      0.86      0.85       101
           1       0.90      0.89      0.89       141

    accuracy                           0.88       242
   macro avg       0.87      0.87      0.87       242
weighted avg       0.88      0.88      0.88       242

Performance (Test):
              precision    recall  f1-score   support

           0       0.83      0.68      0.75        37
           1       0.61      0.79      0.69        24

    accuracy                           0.72        61
   macro avg       0.72      0.73      0.72        61
weighted avg       0.75      0.72      0.72        61

Confusion Matrix (Test):
<sklearn.metrics._plot.confusion_matrix.ConfusionMatrixD
```



```
Random Forest Model:
Performance (Train):
              precision    recall  f1-score   support

           0       0.92      0.78      0.84       101
           1       0.86      0.95      0.90       141

    accuracy                           0.88       242
   macro avg       0.89      0.87      0.87       242
weighted avg       0.88      0.88      0.88       242

Performance (Test):
              precision    recall  f1-score   support

           0       0.92      0.62      0.74        37
           1       0.61      0.92      0.73        24

    accuracy                           0.74        61
   macro avg       0.77      0.77      0.74        61
weighted avg       0.80      0.74      0.74        61

Confusion Matrix (Test):
<sklearn.metrics._plot.confusion_matrix.ConfusionMatri
```



```
Bernoulli Naïve Bayes Model:
Performance (Train):
              precision    recall  f1-score   support

           0       0.85      0.81      0.83       101
           1       0.87      0.90      0.89       141

    accuracy                           0.86       242
   macro avg       0.86      0.86      0.86       242
weighted avg       0.86      0.86      0.86       242

Performance (Test):
              precision    recall  f1-score   support

           0       0.93      0.73      0.82        37
           1       0.69      0.92      0.79        24

    accuracy                           0.80        61
   macro avg       0.81      0.82      0.80        61
weighted avg       0.84      0.80      0.81        61

Confusion Matrix (Test):
<sklearn.metrics._plot.confusion_matrix.ConfusionMatrixD
```

Conclusions

- No one particular model always had the highest or lowest accuracy scores
- K-Nearest Neighbor model consistently finished within the top 3 highest accuracy scores
- The highest overall accuracy scores for Bernoulli Naïve Bayes model seemed to come from evaluating datasets with more binary based features

Future Work

- Repeat the current experiment after swapping the Bernoulli Naïve Bayes algorithm with a different algorithm
- Evaluate the same algorithm against a different set of datasets which contain more binary type features

Project Resources

- Dataset Resources
  - https://www.kaggle.com/datasets/abhia1999/chronic-kidney-disease
  - https://www.kaggle.com/datasets/rashikrahmanpritom/heart-attack-analysis-prediction-dataset
  - https://www.kaggle.com/datasets/saddamazyazy/go-to-college-dataset
  - https://www.kaggle.com/datasets/tejashvi14/employee-future-prediction
  - https://www.kaggle.com/datasets/whenamancodes/predict-diabities
- Informational Resources
  - https://corporatefinanceinstitute.com/resources/data-science/bayes-theorem/
  - https://iq.opengenus.org/bernoulli-naive-bayes/
  - https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.BernoulliNB.html
  - https://thecleverprogrammer.com/2021/07/27/bernoulli-naive-bayes-in-machine-learning/
  - https://thecleverprogrammer.com/2021/02/07/naive-bayes-algorithm-in-machine-learning/