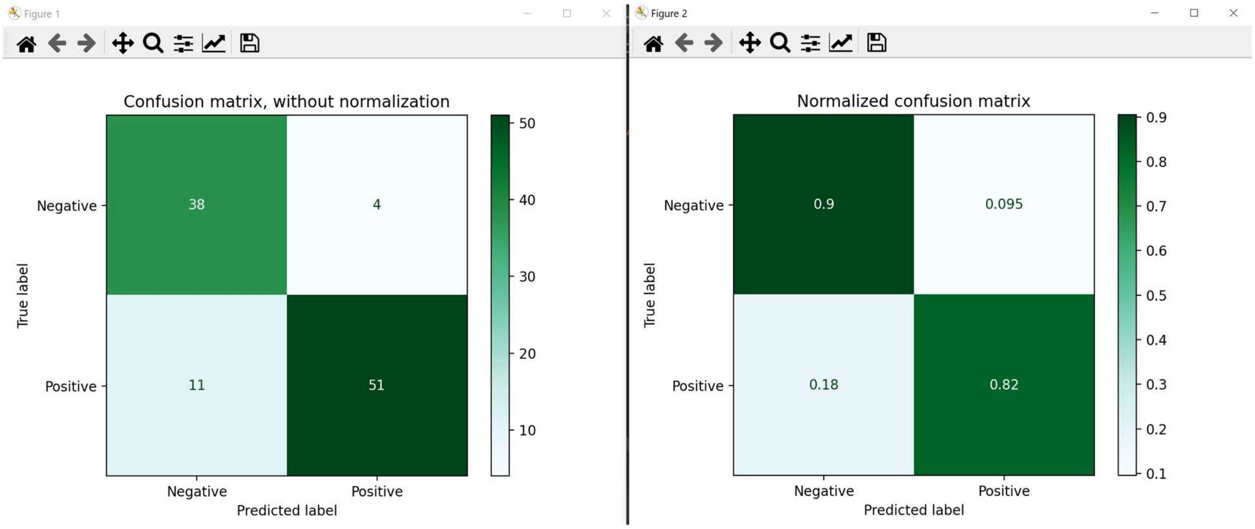
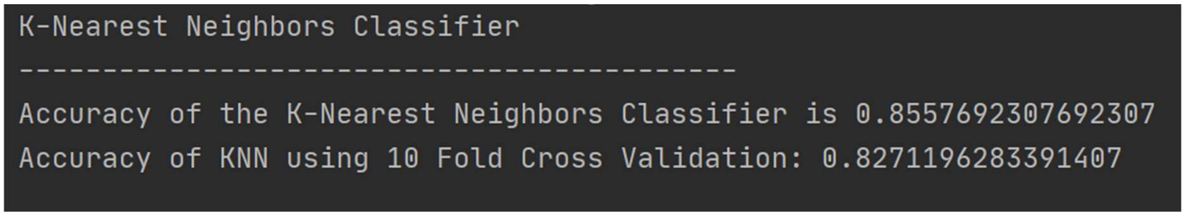
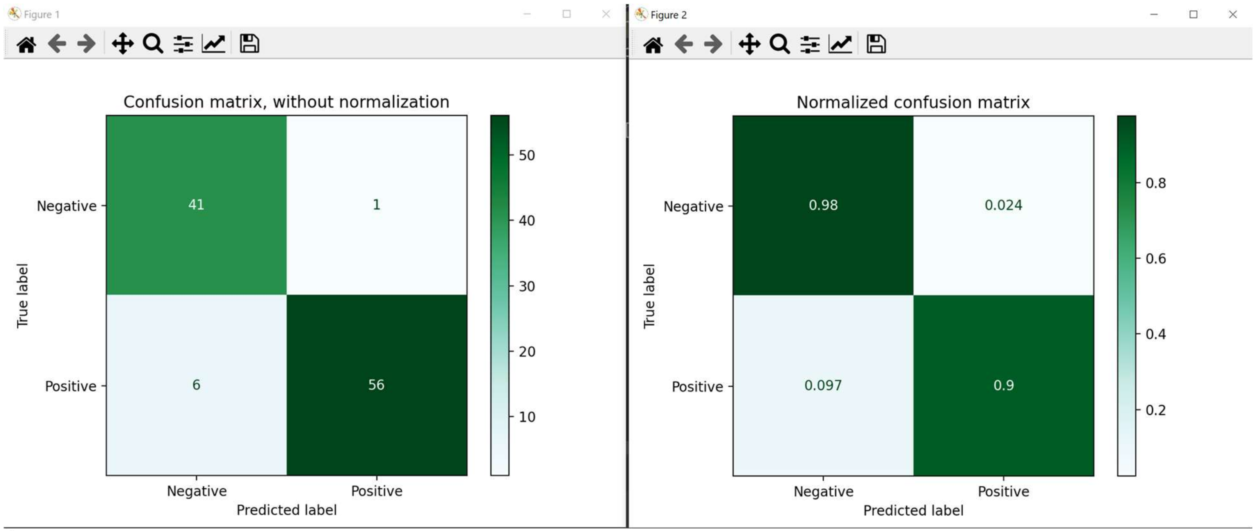
K Nearest Neighbors (KNN) Confusion Matrices

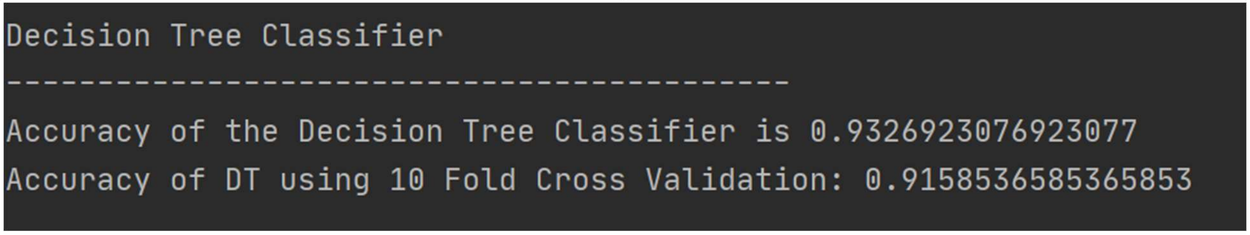




```
K-Nearest Neighbors Classifier
-------------------------------------------
Accuracy of the K-Nearest Neighbors Classifier is 0.8557692307692307
Accuracy of KNN using 10 Fold Cross Validation: 0.8271196283391407
```

Decision Tree (DT) Confusion Matrices





```
Decision Tree Classifier
-------------------------------------------
Accuracy of the Decision Tree Classifier is 0.9326923076923077
Accuracy of DT using 10 Fold Cross Validation: 0.9158536585365853
```

Discussion

The dataset used for this investigation of machine learning was the “**Early stage diabetes risk prediction dataset.”** From the UCI Machine learning repository.

https://archive.ics.uci.edu/ml/datasets/Early+stage+diabetes+risk+prediction+dataset.

The dataset contained predictors of diabetes (both nominal and categorical)

Attributes used in my code were: Age, Gender, Polyuria, Weakness, Obesity (these are some of the most common signs of diabetes).

Class labels were either positive or negative for diabetes.

Approaches:

The two supervised machine learning approaches I chose to use were ‘**K Nearest Neighbors (KNN)**’ and ‘**Decision Trees (DT)’**.

As the dataset had both categorical and nominal data, in order to use the K Nearest Neighbors algorithm I had to convert the categorical data to nominal data. This was done for Gender, Polyuria, Weakness and Obesity using the OrdinalEncoder() function from sklearn. This changed the “yes/ no” or “male/female” for each of the attributes to 0.0 or 1.0.

While technically decision trees could interpret both categorical and nominal attributes, the same preparation of data was used for simplicity and for equal comparison.

Results and Comparing the Approaches:

In order to test the efficiency and accuracy of both machine learning approaches I chose to look at a few metrics. Sklearn.metrics accuracy_score was used to get an overall idea of the prediction accuracy of both classifiers. **DT outperformed KNN in overall accuracy (93% vs 86%)** for one test as shown by the first line of the screenshotted outputs. To get a more representative sample of accuracy I used K fold cross validation (where k = 10) showing multiple trials of accuracy and compared the means. In this case **DT still outperformed KNN (92% vs 83%)**.

More trends can be interpreted from the two approaches using their respective confusion matrices. **Both had better accuracy predicting negative cases** of diabetes with 90% in KNN and 98% in DT. Notably this was the smaller class in the test_data as there were only 42 negative cases vs 56 positive cases.

While I didn’t necessarily notice this anecdotally, from some basic research on the two approaches DT should be faster than KNN as KNN has expensive real time execution*.

*Source (https://towardsdatascience.com/comparative-study-on-classic-machine-learning-algorithms-24f9ff6ab222)

Which is “better “ ?

As discussed above, Decision Trees show increased accuracy and speed as compared to K Nearest Neighbors overall. Another benefit, as mentioned earlier, is that Decision Trees can take both categorical and nominal attributes in training.

For these reasons I would say **Decision Trees are the better supervised machine learning approach** as compared to K Nearest Neighbors.