Data Mining Final Project

Option 1: Supervised Data Mining (Classification)

By: Rushikesh Kshirsagar

**Algorithms:**

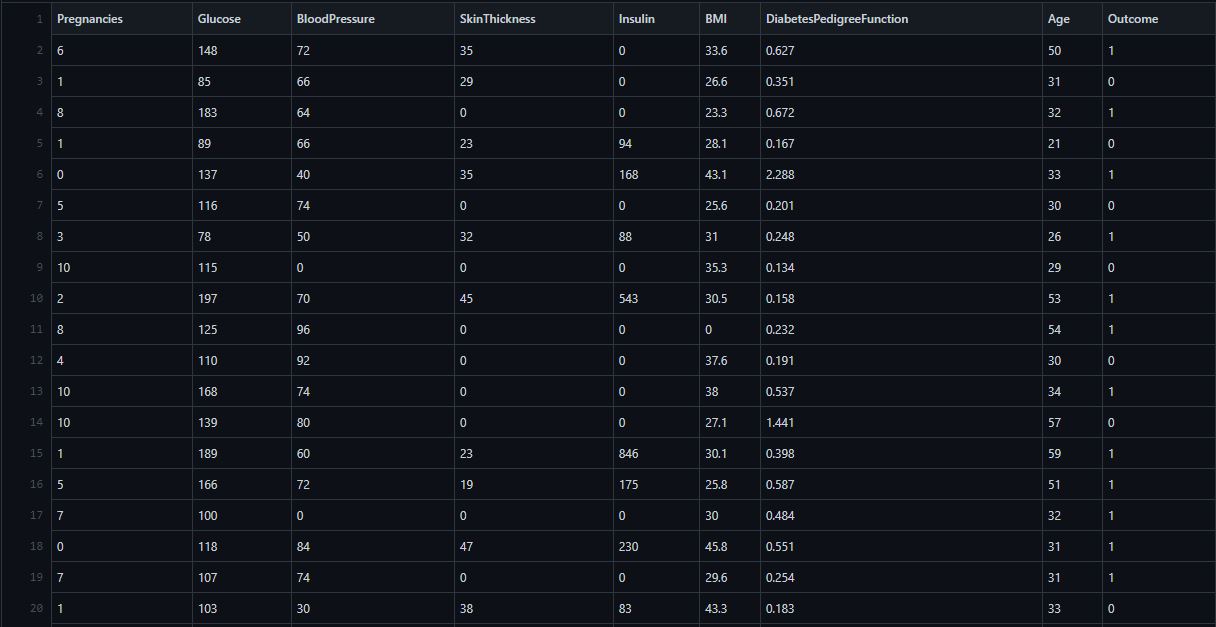
1. Random Forest
2. Decision Tree
3. Long Short-Term Memory (LSTM)

**Dataset:**

* <https://github.com/Rushi-Repo/Final-Project-Data-Mining/blob/main/diabetes.csv>

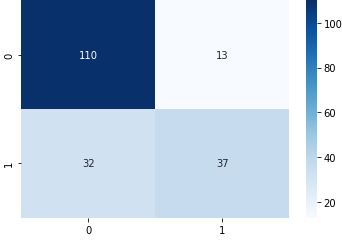
**Environment:**

* Google Collab
* <https://colab.research.google.com/drive/165qlHYlIHAbM_p3HYtXshdFBGWFcqPyY?usp=sharing>
* Jupyter Notebook attached in the root folder.
* Github Repository – <https://github.com/Rushi-Repo/Final-Project-Data-Mining>
* **Evaluation:**
* The dataset consists of 768 rows and 9 columns.
* Out of the 9 columns, the last one (‘Outcome’) is predicted by the model using all the other 8 features.
* The dataset is split 75-25 for training and testing respectively.
* The first 20 entries of the dataset are as shown below.



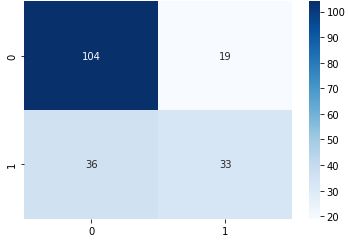
1. Random Forest:

* All the algorithms used in this project are imported from the sklearn’s library.
* The accuracy of the Random Forest algorithm after 10-fold cross validation is 76.6
* Confusion Matrix:



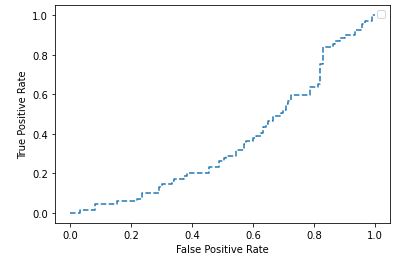
1. Decision Tree:

* The accuracy of the Decision Tree algorithm after 10-fold cross validation is 71.4
* Confusion Matrix:



1. LSTM:

* The accuracy of the LSTM algorithm after 10-fold cross validation is 73.81
* ROC for LSTM:



TRP for LSTM:

Array ([0, 0, 0, 0.01449275, 0.01449275,

0.04347826, 0.04347826, 0.05797101, 0.05797101, 0.07246377,

0.07246377, 0.10144928, 0.10144928, 0.13043478, 0.13043478,

0.14492754, 0.14492754, 0.15942029, 0.15942029, 0.17391304,

0.17391304, 0.1884058, 0.1884058, 0.20289855, 0.20289855,

0.23188406, 0.23188406, 0.26086957, 0.26086957, 0.27536232,

0.27536232, 0.28985507, 0.28985507, 0.31884058, 0.31884058,

0.34782609, 0.34782609, 0.36231884, 0.36231884, 0.37681159,

0.37681159, 0.39130435, 0.39130435, 0.4057971, 0.4057971,

0.43478261, 0.43478261, 0.44927536, 0.44927536, 0.46376812,

0.46376812, 0.49275362, 0.49275362, 0.50724638, 0.50724638,

0.52173913, 0.52173913, 0.55072464, 0.55072464, 0.56521739,

0.56521739, 0.5942029, 0.5942029, 0.63768116, 0.63768116,

0.65217391, 0.65217391, 0.75362319, 0.75362319, 0.84057971,

0.84057971, 0.85507246, 0.85507246, 0.86956522, 0.86956522,

0.88405797, 0.88405797, 0.89855072, 0.89855072, 0.92753623,

0.92753623, 0.95652174, 0.95652174, 0.97101449, 0.97101449,

1, 1])

FRP for LSTM:

Array ([0, 0.00813008, 0.03252033, 0.03252033, 0.08130081,

0.08130081, 0.15447154, 0.15447154, 0.2195122, 0.2195122,

0.23577236, 0.23577236, 0.29268293, 0.29268293, 0.30081301,

0.30081301, 0.33333333, 0.33333333, 0.34146341, 0.34146341,

0.37398374, 0.37398374, 0.38211382, 0.38211382, 0.45528455,

0.45528455, 0.48780488, 0.48780488, 0.50406504, 0.50406504,

0.51219512, 0.51219512, 0.54471545, 0.54471545, 0.56910569,

0.56910569, 0.57723577, 0.57723577, 0.60162602, 0.60162602,

0.6097561, 0.6097561, 0.62601626, 0.62601626, 0.63414634,

0.63414634, 0.64227642, 0.64227642, 0.6504065, 0.6504065,

0.66666667, 0.66666667, 0.69105691, 0.69105691, 0.69918699,

0.69918699, 0.70731707, 0.70731707, 0.71544715, 0.71544715,

0.72357724, 0.72357724, 0.78861789, 0.78861789, 0.81300813,

0.81300813, 0.82113821, 0.82113821, 0.82926829, 0.82926829,

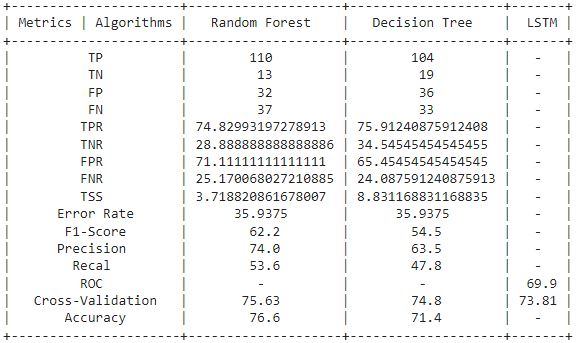
0.86178862, 0.86178862, 0.8699187, 0.8699187, 0.88617886,

0.88617886, 0.90243902, 0.90243902, 0.93495935, 0.93495935,

0.95934959, 0.95934959, 0.96747967, 0.96747967, 0.99186992,

0.99186992, 1])

**Summary:**



**Results:**

* From the experimental result, we can conclude that Random Forest Classifier performs the best amongst the three in terms accuracy and 10-fold cross validations.
* Random forest leverages the power of multiple decision trees. It does not rely on the feature importance given by a single decision tree.
* Random forest chooses features randomly during the training process. Therefore, it does not depend highly on any specific set of features.
* This is a special characteristic of random forest over bagging trees.
* Therefore, the random forest can generalize over the data in a better way.
* This randomized feature selection makes random forest much more accurate than a decision tree.