CSC 859 Fall 2023 09/19 Team 4

Team members:

- Hruthika Saripalli
- Richard Lenhart
- Guillermo Villar

Team lead:

- Guillermo Villar (923696602@sfsu.edu)

Database and Application:

Our chosen data is about diabetes prediction using Machine Learning. Our source is an article (https://thecleverprogrammer.com/2020/07/13/predict-diabetes-with-machine-learning/) that applies different Machine Learning techniques to the database to try to predict based on some features measured if patients have diabetes or not.

This source applies KNN and Decision Tree Classifier for this goal, so we can compare the outcomes with our application of Random Forest, and we can in this way mimic the process of auditing.

Number of samples: Our dataset has a total of 769 samples

Features: The dataset contains 8 features, all of which are numeric (number of pregnancies, glucose, blood pressure, skin thickness, insulin, bmi, Diabetes pedigree function, and Age)

Missing values: We noticed a few missing values which were 0-filled by previous source for their use. So, we will have to analyze how it will affect our outcomes and discuss possible techniques for managing these missing value.

Other information: Ground truth is known and given for all samples. The dataset is in a downloadable csv format.

Chosen tool for development:

After looking into the differences between both recommended tools and checking the pretty much technical equivalence between them, we have chosen to go with Python Scikit, due to it having an easier learning curve specially for people already familiar with python or other programming languages, as well as the versatility it offers.

There are some other factors that would make R a more attractive alternative, like more clear variable namings and a more narrow approach that might be best for a simple application like ours, but the incredible amount of information on the web as well as the development of gen-AI´s that can help us with technical issues have led us to choose Scikit over R.