Victoria De Cun
Asif Abbas
Moein Afrouzmehr
Fidel Gutierrez
Indronil Bhattacharjee

CS508 - Intro to Data Mining
Midterm Report

22 October 2023

Diabetes Predictive Modeling: Leveraging Machine Learning for Early Detection

Dataset

- **Source** - The dataset was obtained from Kaggle.
- **Category** - This dataset originates from the healthcare domain. The class label is a binary attribute indicating whether a patient has diabetes. The attributes encompass various patient-specific details, including age, BMI, glucose levels, cholesterol levels, among others."
- **Features and size** - The dataset comprises 18 numerical features, and 4304 instances. All inter-feature correlations are below 0.35. Consequently, no attributes are excluded from the analysis.
- **Data preprocessing and statistical analysis** - The statistical analysis and the data preprocessing is explained in detail in an accompanying document.
- **Data mining task** - Classification

Algorithms used

- Perceptron
- SVM - Linear and RBF
- Decision Tree
- kNN
- Logistic Regression
- Gradient Boosting
- ANN

For more results on the timing and accuracy, please see the attached .ipynb file.

Future plans

At this stage, we are content with the classifiers used and the results obtained. The next course of action is to prepare the Final Report and practice for the Final Presentation.