Determining Trends in Diabetes Related Health Markers Using Machine Language

Joel Thomas

Week 7 – Rough Draft Template

DSC 580

**Data Preparation:**

This dataset was taken publically from Kaggle and was already in a healthy state for other data scientists to utilize data science prediction techniques and it was already split into a 80/20 split for testing and training amongst Machine learning Models. Below in my work, we do some minor fixes to this dataset that include the following:

* Dropping and targeting the Diabetes\_012 column for ML Practices.
* Splitting the dataset into a 80/20 set.

Outside of this, the EDA process is complete and is ready to create models to predict Diabetes from related health markers using Logistic Regression and Random Forest ML Models.

Below is a brief look at how the dataset looks like:  
A screenshot of a computer

Description automatically generated

Now lets take this data, split it and test/training the Logistic Regression and Random Forest Models and provide charts and graphs to show the strength of our predictions.

A screenshot of a computer

Description automatically generated

A screenshot of a computer code

Description automatically generatedA screenshot of a computer program

Description automatically generated

Now that our models are tested, Lets show Confusion matrix’ to evaluate model effectiveness.

A screenshot of a computer program

Description automatically generated

A graph with numbers and a blue square

Description automatically generatedA graph with numbers and a blue square

Description automatically generated

A screenshot of a computer program

Description automatically generated

**Data Interpretation and Analysis –**

**Accuracy**: Both classifiers achieve relatively high accuracy scores, with the Logistic Regression slightly outperforming the Random Forest classifier. This indicates that both models are able to correctly classify a significant portion of the instances in the dataset.

**Precision**: Precision measures the proportion of true positive predictions among all positive predictions made by the model. Both classifiers have low precision scores, indicating that they have a relatively high rate of false positive predictions. In the context of identifying diabetes, this means that a considerable number of individuals predicted to have diabetes by the models may not actually have it. Recall:

**Recall**, also known as sensitivity, measures the proportion of true positive predictions among all actual positive instances in the dataset. Both classifiers have moderate recall scores, suggesting that they are able to capture a substantial portion of the actual positive cases. However, given the information about the confusion matrix being heavily biased towards the true negative (0,0) square, it's likely that the recall for the positive class (indicating diabetes) is lower than desired.

**F1 Score**: The F1 score is the harmonic mean of precision and recall and provides a balanced assessment of a classifier's performance. Both classifiers have relatively low F1 scores, indicating that they struggle to achieve both high precision and high recall simultaneously. This suggests that there's room for improvement in the classifiers' performance in identifying diabetes.

**Conclusion** –

In summary, while both classifiers demonstrate decent performance in terms of accuracy, there is room for improvement in terms of precision, recall, and overall F1 score, especially in correctly identifying diabetic cases. Further analysis and possibly model refinement are warranted to enhance the classifiers' performance in identifying diabetes accurately. Additionally, addressing the imbalance in the confusion matrix should be a priority to ensure balanced and reliable predictions for both diabetic and non-diabetic cases.