A logo for college computing

Description automatically generated**Assessment Cover Page**

**Declaration**

By submitting this assessment, I confirm that I have read the CCT policy on academic misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source.

I declare it to be my own work and that all material from third parties has been appropriately referenced.

I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution.

|  |  |
| --- | --- |
| *Student Full Name* Jason Liu Doyle |  |
| *Student Number* 2024200 |  |
| Module Title  Machine Learning |  |
| *Assessment Title*  CA1 |  |
| *Assessment Due Date 28/04/2024* |  |
| *Date of Submission*  *28/04/2024* |  |

Introduction

Diabetes is a chronic illness that detrimentally affects people’s body’s abilities to process blood glucose which then leads to potentially severe health complications like heart disease, kidney failure, blindness and limb amputation to just name a few. It is currently a growing global issue, where incidence rates have been skyrocketing because of factors like rising obesity rates, aging populations, lifestyle factors etc. The international Diabetes Federation estimated that 537 million adults are living with diabetes in 2021 and there would be 643 million by 2030 (International Diabetes Federation, 2021). Early diagnosis and management of diabetes can be crucial to mitigate the severe health outcomes. Machine Learning models are promising tools to help analyse big data in health to predict individual risk quicker and more precisely compared to traditional methods (Kavakiotis et al., 2017). It's these tech advancements that allow earlier intervention which will hopefully significantly improve patient’s outcomes and reduce healthcare costs that are tied into the disease (American Diabetes Association, 2021).

With this information I will propose a project to use data analytics to help aid in early intervention for individuals suspected to have diabetes, through an earlier diagnosis. I will do this with two goals; First, to identify features of strong correlation to the presence of diabetes, and Secondly, to create four prediction models, compare their performance and choosing the best one.

I will be doing this with a dataset of 253680 rows and 22 columns (Teboul, 2022). These columns include a list of important health indicators and identifies such as BMI, Smoker, Physical Activity and Age. The goal being to focus on model recall and positive identification of Diabetes.

(Dataset available at end of references)

# Characterization

I started off by importing all the necessary libraries for the rest of the project and importing the data. The dataset seems pretty clean already, with no null sets, no categorical data and I will choose to leave in the duplicates for the time being, as there is no clear indication to remove them yet. I added prediabetic and diabetic into the same class and conducted some simple EDA with a corr matrix and profile report.

Logistic Regression Model

I set X as my variables and y as my target, scaled them and then I let it run. I used SMOTE to address the class imbalance by oversampling the minority class (1.0/has diabetes) for a second run taking note of both results. I used K fold validation on both but know that data leakage would be an issue with the model which included SMOTE. I tried to use some initiative and research ways around this and experimented with a pipeline as well as conducting validation on each fold with a loop, but I couldn’t get it to work in the end \*(This is the same case for other models) . I got a rank of the most important features before repeating the process with 25% and 30%.

Decision Tree

I repeated the whole process with a random forest model with some exceptions like not scaling the date and using all features instead of a select few. I set up X and y, split them into training and testing, then conducted SMOTE. Created a decision tree classifier model instead of a logistic regression one, ran it, recorded the results and repeated the process another two times with a training split of 25% and 30%. I also created a ranked list based on importance in the decision tree classifier.

Random Forest

I repeated the process except this time with a random forest model. I started with 1000 estimators but changed to 100, as wating for the models to load each time took too long for the given timeframe.

GNB

In a processes similar to the previous model I created a Gaussian Naïve Bayes model. I scaled the data this time and used all features again. I added a cross val score, and I narrowed it down to cv=45 as being the most accurate. I also stacked the array of the true targets with the predicted ones, transposed the results and compared them.

SVM

I finished with my last model which was a Support vector machine. I used the strongly corelated features as in the logistic regression model, trained SMOTE, ran the model and recorded results.

Hyperparameter

The primary purpose of hyperparameter tuning in ML is to improve the performance of your model by using the best combination of hyperparameters (Géron, 2019). Hyperparameters are like settings that control the way a model learns and it’s ability to generalise to new data. Tuning parameters like learning rate, regularisation strength or tree depth are changed to improve the model’s predictive accuracy and generalisation capabilities. Hyperparameter tuning helps in mitigating overfitting, impoves model robustness and maxes the predictive performance across different datasets.

Specific Hyperparameter tuning techniques like GridSearchCV play an important role in improving ML models. GridSearchCV systematically exploers a preset hyperparameter space, and then evaluates a model performance through cross-validation. This technique, which seems to be commonly used, streamline the process of fine tuning model configurations for the best performance across diverse datasets (Pedregosa et al., 2011).

Results

*\*I have provided two tables below for easier viewing; Model results + Influential Features*

My project started with the goal of pinpointing the most influential features for diabetes. To achieve this goal I used a correlation matrix, as well as feature importance rankings in my logistic regression model and decision tree algorithms. Logistic regression showed the hierarchy of importance in the following order; General Health, Age, BMI, High Blood Pressure and High cholesterol. With my decision tree model we saw; High blood Pressure, Health and High Cholesterol. I collected all this data, calculated the mean for each feature and then ranked them. I found that General Health, high blood pressure, BMI, Age, and High Cholesterol had the highest correlation to diabetes, 

The implications of these results would suggest a more focused approach in clinical screenings for diabetes, which would allow healthcare professionals to concentrate on these important factors. At the same time, these factors can give the public a clear list of markers to watch out for in their own prevention of Diabetes.

I kept track of all my model results in an excel spreadsheet which I downloaded a csv and loaded back into my Jupyter notebook for analysis.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generatedA screenshot of a graph

Description automatically generatedA screenshot of a computer screen

Description automatically generated

For my models, each were given two rounds of testing, one with and one without SMOTE (Except for SVM due to errors with loading results). The models with SMOTE show better recall in ‘Diabetes present’, with the highest being 74% in my logistic regression model. But, this came as a trade off to both overall accuracy and precision. Because we are working with healthcare data the presence of flase positives can lead to undue stress and panic in patients, and I considered this carefully in my approach.

A screenshot of a computer

Description automatically generatedA screenshot of a graph

Description automatically generatedA screenshot of a graph

Description automatically generated

A screenshot of a computer

Description automatically generatedA screenshot of a graph

Description automatically generatedA screenshot of a graph

Description automatically generated

The logistic regression model before the SMOTE had a greater accuracy and a pretty good predictive score for non diabetic instances. But, it was really bad when it came to our priority, which was predicting diabetes presence. A similar thing was seen across all the models, with the exception of my SVM model which presented operational challenges after SMOTE that I just couldn’t get right within the time limit for the project.

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generatedA screenshot of a graph

Description automatically generated

The class imbalance was the main problem during this project for me. I tried different strategies to mitigate it, particularly focusing on balancing my priority of detecting diabetes with overall precision.

# Conclusion

In conclusion, while my models have shown a potential in identifying key indicators for Diabetes, they also show the delicate balance needed between the recall and the precision in the medical field. Further work should try and explore different ways to handle the class imbalance, refine the precision of the models while not sacrificing our priority of detecting Diabetes effectively.



 

# Word Count : 1284

# References

American Diabetes Association; 2. Classification and Diagnosis of Diabetes: *Standards of Medical Care in Diabetes—2021*. *Diabetes Care* 1 January 2021; 44 (Supplement\_1): S15–S33.

Géron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems (2nd ed.). O'Reilly Media.

International Diabetes Federation. (2021). IDF Diabetes Atlas, 10th ed. Brussels, Belgium: International Diabetes Federation.

Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. (2017). Machine Learning and Data Mining Methods in Diabetes Research. *Computational and Structural Biotechnology Journal*, 15, 104-116.

Teboul, A. (2022). *Diabetes Health Indicators Dataset*. Available at: <https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset> [26/04/2024].