Project Description

Group 10

Name	Student ID	Email
Saw Yin Qi	20619467	hcyys4@nottingham.edu.my
Joshua Mak	20469457	hfyjm2@nottingham.edu.my
David Leong	20620891	hcydl4@nottingham.edu.my
Abdullah Sami Bin Mamun	20614435	hcyam6@nottingham.edu.my
Mohammed Ahnaf Habib	20511215	hfymh6@nottingham.edu.my

Supervisor: Dr Doreen Sim Ying Ying

Title: Diabetes Prediction and Multiple Advisory Application

Introduction:

Diabetes is one of the fastest growing health issues globally, with an estimated 1 in 10 people affected by the condition. While there are several ways to determine if a person has diabetes or not, such as visiting the doctor which can be time consuming or purchasing a diabetes test kit, which can be costly, these methods are often inconvenient for people with busy schedules or limited income. This is where machine learning can provide a solution.

Machine learning has already proven to be a success in the medical field, from identifying cancer cells to aiding in drug development. We hope to use these capabilities to predict whether an individual has diabetes based on their health inputs, additionally the application will give personalised advice to help users improve their health. By offering a free and efficient method to identify potential diabetes, our project aims to make diabetes diagnosis more accessible and give health advice which benefits any users.

Objective:
Our aim for this project is to develop a software which notifies users if they have suspected diabetes or not. Advice would also be displayed once the suspected results have been shown. This software is made to notify users on their current health status which is important in today's day and age. By identifying early signs of diabetes at an early stage, users can take precautionary measures to prevent their blood sugar increasing to a critical level.

Methodology:
Data source:

- The dataset will be sourced from public datasets from Kaggle.

Data preprocessing:

- Exploratory data analysis will be done to understand the data distribution and relationships between features and to identify potential issues such as class imbalances.
- Both mean, mode and median and k-nearest neighbours will be used for the imputation of missing values, where the performance of the model trained on both preprocessed datasets will be evaluated.
- Outliers analysis can be done using z-score and interquartile range where outliers can be capped, transferred or removed if they are determined to be likely to result from data entry errors.
- Categorical variables will be encoded into numerical values if necessary.
- Features will then be scaled, ensuring that all features contribute proportionally to the model's performance.
- Class imbalances will be handled either by using resampling methods or by using class weights. Resampling could be done by oversampling and or undersampling classes or by using a synthetic data generation such as SMOTE.
- Feature selection and engineering will be done where appropriate.

Model development:

- Multiple machine learning and deep learning algorithms will be developed and implemented to predict diabetes. The models include:
    - Feedforward neural networks
    - Recurrent neural networks
    - Gradient boosting machines
    - Support vector machines
    - Decision tree
    - Logistic regression
    - Random forests
- All models will be implemented using **Python**.
- Ensemble models will then be developed using adaptive boosting, or stacking, voting or averaging.

Model training and evaluation:

- K-fold cross-validation with k value 5 will be used for the training and evaluation of all models to ensure that the model is not overfitting to the dataset trained.
- The models will then be evaluated using a set of performance metrics including:
    - The precision, recall and f1 score for both positive and negative outcomes
    - Macro average and weighted average of the metrics
    - Accuracy and loss
    - Confusion matrix
    - Accuracy and loss over epoch
    - Comparison of accuracy over each fold
    - Receiver operating characteristic curve and its AUC.

Model optimisation:

- The models with the best performance based on its evaluation will then be further optimised by reducing model complexity, optimising neural network training etc. depending on the specific model to improve its predictive accuracy.
- Further preprocessing of the dataset can also be done if appropriate such as using different regularisation techniques, further feature engineering etc.
- The best performing models will then be used to develop different ensemble models where the performance of each model will be evaluated

Model implementation:

- The model with the best performance will then be used as the final model to be used in the diabetes prediction and advisory app.
- The final model will be integrated into the app using flask or django.

App implementation:

- The app will use **Python** for the backend and **HTML, CSS and JavaScript** for the frontend.

Timeline:

Dates	**Description**
Sep 2024-Dec 2024	● Research and dataset collection: research conducted on ML techniques that can be used to efficiently and accurately predict diabetes from existing datasets. Datasets will be used from the Kaggle website and also with an accurate model being discovered the group will proceed with the further proceedings . ● Testing will be done on the datasets to see which dataset gives more accurate predictions for diabetes. ● Initial development: Once the best Machine Learning Model is identified, web development of the webpage and the advisory application will begun and interim group report
Dec 2024 - March 2025	● An early prototype of the diagnosis system will be developed ● Any necessary changes towards the software should be made around this time ● Report writing continues throughout this period
March 2025 - April 2025	● Final testing, debugging, correction of error will be done throughout this period ● Report will be finalised and completed around this time and submitted by 2nd May, 2024 ● The software will be completed and finalised around this time ● Preparation of posters and booth for open day ● Presentation of software

Resources:
Dataset: Dataset will be collected from Kaggle and will be decided later on after research on accuracies.

Python for conducting **ML techniques** on the dataset, **Python** for **back-end** development-**Django and Flask**, **Javascript** for user-interaction, **HTML** and **CSS** which includes the Bootstrap framework for the **Front-end** development.

Conclusion:
The project aims to build a prediction and advisory platform for diabetes- one of the significant health risk factors in the world. Moreover, the project will emphasise accuracy in machine learning for better and more accurate predictions, and allow the user to receive a prediction and further advice for improvement with just a few steps. While emphasising a more user-friendly interface, the application will act as a suitable platform for people of all walks of life to get advice for diabetes prevention and other factors that can cause diabetes in the near future.