MLD Project Phase 1

Topic: Dementia Prediction using ML

By: 1. Vishwashree V Karhadkar [Student ID: 202307962]

2. Gagandeep Singh [Student ID: 202303876]

OVERVIEW

kaggle:https://www.kaggle.com/code/mdismielhossenabir/dementia-health-prediction/input

We have gathered a health dataset consisting of various patients' data, which will predict whether the patient would have dementia or not. The dataset is fetched from Kaggle, the link is attached above.
We will be unitizing multiple columns of the dataset that focus on predicting whether an individual could have dementia based on various dataset features like health, lifestyle, and genetic factors. Key fields include biometric data such as Age, Diabetic status, alcohol level, heart rate, blood oxygen level, body temperature and Weight, Prescription, Dosage in mg etc. Some other columns which might be useful include dominant_Hand, gender, family_History of dementia, smoking_status, and the presence of the APOE_ε4 gene. Additionally, the dataset consists of physical activity, depression status, cognitive test scores, medication history, Nutritional diet, sleep quality, and chronic health conditions, which are also incorporated to enhance prediction accuracy. This comprehensive dataset provides a robust foundation for building a predictive model to assess **Dementia risk.**

Data set variables:

The dataset comprises **1000** rows of patient data and has **24** columns, some of which are represented as numeric values such as heart rate(between 60-100 bpm), blood oxygen level(from 90-100), body temperate(from 36-37), weight, age(60-100) and many more.

There are text values such as prescription (prescribed medicines), educational level, gender(male and female), family history, smoking status (currently smoking or former smoker), physical activity(mild/moderate or passive), and depression status (Yes or No).

Further, the predictive column ‘Diabetic’ has binary values in 0(false) or 1(true).
These attributes and data would be significant in analyzing, training, and predicting our ML model to build the desired prediction model.

Goal and Questions to Predict

1. Our standard task is to predict and diagnose based on the dataset whether the person has dementia or not, with the help of using important features available in the given dataset and using machine learning algorithms. The targeted value will be given in binary classification form i.e. if it is found that the person has dementia, the result would be shown as 1 otherwise 0.

How we will be answering:

-> Initially, we will be choosing the prominent features out of the dataset and then will train the dataset using the df-analyzer and its functions, which would be important for the prediction of dementia. Later and/or during the coding phase, we would add and remove a few features from the dataset for instance age, gender, diabetic level, family history etc to see how the predictions would change accordingly i.e. to see if we are diverting away from the predictions or close to prediction by adding those features. We would also be using most of the features at every iteration to see what difference it would make.

-> Furthermore, we will be calculating some different predictions based on different columns presented in the dataset, such as what gender might have a high risk of getting dementia which age category has a major risk or which medical history they had who got positive results for dementia, do they have history of smoking, depression and does that impact our prediction model or does it make more precise based on the input features.

References and Publications

*Regarding the references part, We have already dropped and messaged the person who has worked on the dataset [Md. Ismael Hossein Abir] whose name has been seen on Kaggle asked about the source of the dataset. We should update the source and findings as we gather more knowledge with this dataset.