# MLD Project Phase 3

## Topic: Dementia Prediction using ML

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

2. Gagandeep Singh [Student ID: 202303876]

### INTRODUCTION

In phase 3 of our project, we have included the feedback from phase 2 by adding the missing redundancy aware feature selection in our df-analyze script and executed it on our main dataset file. Also as per the suggestions on the task to be done from phase 1, we have further divided the dataset into gender-specific groups. These Datasets were also analyzed using df-analyze which provided us valuable insights based on male and female subgroups in terms of predicting dementia. This report presents detailed findings on the methods, prediction accuracy, feature selection, and statistical performance of different models with supporting result tables including a full dataset, and male and female-specific results attached in the report.

## 1. Materials and Methods

### 1.1 Dataset Description:

The dataset in our tabular project has been used to predict if the individual has dementia or not based on various health metrics and lifestyle factors. It includes 24 feature measurements in total. The dataset has been collected from the Kaggle website and as per the description, it was assembled and collected in 2023 from various medical sources, publication sites, and studies to aim at understanding the factors associated with dementia and predicting using ML.[1] The Dataset has a total number of **1000** samples in total. According to that, **485** entries of the dataset are diagnosed with dementia (value=1) which is our group of interest, and the remaining **515** samples are represented without dementia (value=0) (not group of interest).

For the scope of this phase, we are diving our dataset into male and female entries and we would be noticing the findings on how df analyze software presents results on the basis of these 2 main features.

So according to that, we have now 2 different datasets.

Female entries:

Group of interest (Dementia = 1): **244** entries

Not a group of interest (Dementia = 0): **260** entries

Male entries:

Group of interest (Dementia = 1): **241** entries

Not a group of interest (Dementia = 0): **255** entries

### 1.1.2 Target Variable Description:

Our target variable for this dataset selected is dementia, a **Binary Classification** represented in the last column. It can be derived as Dementia =1 and not dementia =0. If the value stands at 1, we can consider that that person is likely to have dementia based on the input features provided.

### 1.1.3 Feature Measurements:

The following two tables given below in this section show feature measurements for continuous as well as categorical variable values with the feature (column name), and its mean, the Standard deviation value for those specific columns and are divided based on a group of interest and not the group of Interest.

- Table 1.1.1 shows continuous values for full dataset samples. [given below]
- Table 1.1.2 shows Categorical Valuesor full dataset samples. [given below]

We have attached the other gender-specific datasets tables for males and females in the Appendix **section [A]** which describes the feature measurement values and their mean and median based on the male dataset and female dataset for their continuous and categorical variables.
These are labeled as follows:

1. Table 1.2.1 shows continuous values for Male dataset samples.
2. Table 1.2.2 shows Categorical Valuesor Male dataset samples.
3. Table 1.3.1 shows continuous values for Female dataset samples.
4. Table 1.3.2 shows Categorical Valuesor Female dataset samples.

**Table 1.1.1. Continues Values for Full Dataset**

| Column | Group of Interest Dementia=1 | | Group of Not-Interest Dementia=0 | |
|---|---|---|---|---|
| | Mean | Std Deviation | Mean | Std Dev |
| Diabetic | 0.536 | 0.499 | 0.491 | 0.500 |
| Alcohol Level | 0.098 | 0.059 | 0.098 | 0.058 |
| Heart Rate | 79.536 | 12.127 | 79.238 | 12.098 |
| BloodOxygen | 95.010 | 2.886 | 95.529 | 2.957 |
| BodyTemp | 36.775 | 0.423 | 36.747 | 0.438 |
| Weight | 73.585 | 15.027 | 75.016 | 13.898 |
| Age | 74.325 | 9.339 | 75.456 | 8.832 |

| | | | | |
|---|---|---|---|---|
| CognitiveTest | 3.620 | 2.310 | 8.984 | 0.808 |
| Dementia | 1 | 0 | 0 | 0 |

**Table 1.1.2. Categorical Values for Full Dataset**

| Column | Values | Group of Interest Dementia=1 | Group of Not-Interest Dementia=0 |
|---|---|---|---|
| **Prescription** | Galantamine | 125 | NA |
| | Donepezil | 113 | NA |
| | Rivastigmine | 119 | NA |
| | Memantine | 128 | NA |
| **Education level** | Primary School | 181 | 208 |
| | Secondary School | 155 | 149 |
| | Diploma/Degree | 51 | 101 |
| | No School | 98 | 57 |
| **Dominanat Hand** | Left | 251 | 268 |
| | Right | 234 | 247 |
| **Gender** | Male | 241 | 255 |
| | Female | 244 | 260 |
| **Family History** | No | 255 | 225 |
| | Yes | 230 | 290 |
| **Smoking Status** | Current smoker | 90 | 90 |
| | Former smoker | 252 | 206 |
| | Never smoker | 233 | 219 |
| **APOE_e4** | Positive | 435 | 259 |

| | | | |
|---|---|---|---|
| | Negative | 50 | 256 |
| **Physical Activity** | Sedentary | 158 | 173 |
| | Moderate Activity | 150 | 160 |
| | Mild activity | 169 | 182 |
| **Depression status** | Yes | 245 | 0 |
| | No | 240 | 516 |
| **Medical history** | Yes | 251 | 263 |
| | No | 234 | 252 |
| **Nutrition diet** | Low Carb | 162 | 168 |
| | Meditarrean | 169 | 169 |
| | Balanced | 154 | 178 |
| **Sleep Quality** | Poor | 264 | 270 |
| | Good | 221 | 245 |
| **Chronic health condition** | Diabetes | 260 | 253 |
| | Heart Disease | 65 | 90 |
| | Hypertension | 72 | 81 |
| | Na | 88 | 0 |

### 1.2 Machine Learning

The df-analyze is a Python library package that offers different and multiple utility functions and switches for processing and analyzing datasets for machine learning and prediction. It is a command line tool and can perform AutoML[3] on different datasets which are small to medium-sized tabular datasets (less than 200000 samples and from 50-10 features. Df-analyze automates and runs different ML algorithms and gives extensive and elaborate reports. It includes many important key functionalities which are feature type inference, feature description (e.g. univariate associations and stats), data cleaning (e.g. NaN handling and imputation), training, validation, and test splitting, feature selection, hyperparameter tuning, model selection, and validation. It runs the ML algorithms to proceed with an output with all important results stored in tabular format.[2]

The df-analyze tool provides a lot of sensitivity and specificity metrics which are valuable for very important and have significance for assessing performance in ML models.[2] By utilizing this functionality, we can measure how effectively the model identifies true positives and true negatives which are essential and critical for healthcare datasets such as our dementia prediction dataset where accuracy predictions have very high importance and decision importance.

In our previous iteration, we hypothesized that we would run df-analyze 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 Df-analyze. With that specific Run, we have separated our dataset into gender-specific data, where one dataset is male and another is female. Based on this, we have run the df-analyze on both of these datasets to find interesting findings and comparisons to notice how the prediction goes.

The interesting features for prediction found after running the df-analyze in our datasets are as follows which have been listed below. The source of these findings is from **[selection>filter>prediction_selection_repot.md]** from each run results folder.

When analyzing the interesting features which are across the full dataset run, male and female datasets, various commonalities have been seen which are as follows:

1. Cognitive Test Scores, Prescription information, Depression Status and APOE ε4 are shared between all datasets, underscoring the importance of these features that might be helpful in the prediction of dementia.
2. Smoking Status appears in both the full and male datasets.
3. Physical Activity and Nutrition Diet are seen in common between only the male and female datasets, suggesting the life factors and is an interesting finding as it's only specific and common in gender-specific datasets.

Following are feature selection reports for all 3 dataset Runs:

### 1. Full dataset

['Cognitive_Test_Scores', 'Dosage_in_mg', 'Weight', 'AlcoholLevel', 'Prescription', 'Depression_Status', 'APOE_ε4', 'Education_Level', 'Smoking_Status', 'Family_History', 'Medication_History']

#### 2. Male dataset

['Cognitive_Test_Scores', 'Dosage_in_mg', 'BloodOxygenLevel', 'Age', 'Prescription', 'Depression_Status', 'APOE_ε4', 'Smoking_Status', 'Family_History', 'Physical_Activity', 'Nutrition_Diet'].

#### 3. Female dataset

['Cognitive_Test_Scores', 'Dosage_in_mg', 'Weight', 'HeartRate', 'Prescription', 'Depression_Status', 'APOE_ε4', 'Education_Level', 'Physical_Activity', 'Nutrition_Diet', 'Diabetic']

## 1.3 Statistical Analysis

The statistical analysis that is used in this project involves using df-analyze machine learning models which have various feature selection techniques that show predictive outcomes. Various models such as Logistic Regression (LR), Stochastic Gradient Descent (SGD), Random Forest (RF), LightGBM (LGBM), and K-Nearest Neighbors (KNN) were used across the full, male and female datasets. Each model was applied with various feature selection strategies, which include association-based (assoc), predictive (pred), embedded methods (embed_linear, embed_lgbm), and wrapper-based (wrap) models.

For this phase of the project, we will be moving forward and focusing on the **overall accuracy** of the predicted features. It is found in the results directory, a table named 5-fold performance on the holdout set in the file named “results_report.md”. We found that multiple models showed the same accuracy as 100% or a value of 1.00. The following table shows the overall result of the Full dataset.

The results for the Male and Female dataset runs are given in the Appendix in **section [B]** where table 3.2 indicates males, and 3.3 indicates denotes female dataset run results.

**Table 3.1: Full Dataset Overall Accuracy**

| Model | Selection | Embed Selector | Acc |
|---|---|---|---|
| lr | assoc | none | 1.000 |
| lgbm | pred | none | 1.000 |
| sgd | none | none | 1.000 |
| sgd | embed_linear | linear | 1.000 |
| sgd | embed_lgbm | lgbm | 1.000 |
| sgd | assoc | none | 1.000 |
| rf | wrap | none | 1.000 |

| | | | |
|---|---|---|---|
| rf | pred | none | 1.000 |
| rf | embed_lgbm | lgbm | 1.000 |
| rf | assoc | none | 1.000 |
| lr | wrap | none | 1.000 |
| lr | pred | none | 1.000 |
| lr | embed_linear | linear | 1.000 |
| lr | embed_lgbm | lgbm | 1.000 |
| lgbm | wrap | none | 1.000 |
| sgd | wrap | none | 1.000 |
| lgbm | embed_lgbm | lgbm | 1.000 |
| lgbm | assoc | none | 1.000 |
| knn | wrap | none | 1.000 |
| lgbm | embed_linear | linear | 1.000 |
| knn | embed_lgbm | lgbm | 1.000 |
| lr | none | none | 0.997 |
| knn | pred | none | 0.960 |

| | | | |
|---|---|---|---|
| sgd | pred | none | 0.957 |
| knn | none | none | 0.828 |
| knn | assoc | none | 0.823 |
| knn | embed_linear | linear | 0.670 |
| dummy | embed_linear | linear | 0.520 |
| lgbm | none | none | 0.515 |
| rf | embed_linear | linear | 0.515 |
| rf | none | none | 0.515 |
| dummy | assoc | none | 0.515 |
| dummy | pred | none | 0.500 |
| dummy | none | none | 0.495 |
| dummy | wrap | none | 0.487 |
| dummy | embed_lgbm | lgbm | 0.482 |

## 2. Results

### 2.1 Summary results with Table.

The following table is the output after running df-analyze which is summarized below. Table 2.1 shows the performance of various models using a 5-fold performance on the holdout set table from the Full Dataset.

Each model was evaluated based on several metrics, including Accuracy (Acc), Area Under the Receiver Operating Characteristic Curve (AUROC), Balanced Accuracy (Bal-Acc), F1 Score (F1), Negative Predictive Value (NPV), Positive Predictive Value (PPV), Sensitivity (Sens), and Specificity (Spec). Overall, several models demonstrated excellent performance, with many achieving near-perfect results.

We have completed 3 different runs, first for the full dataset run, we have displayed the table here and for the Male and Female dataset results table is displayed in the Appendix **section [C] where the result table is 2.2 and 2.3 respectively.**

Table: 2.1

**Table: 2.1. 5-fold performance on holdout set table for Full Dataset**

| Model | Selection | Embed Selector | Acc | AUROC | Bal-Acc | F1 | NPV | PPV | Sens | Spec |
|---|---|---|---|---|---|---|---|---|---|---|
| lr | assoc | none | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| lgbm | pred | none | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| sgd | none | none | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| sgd | embed_linear | linear | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| sgd | embed_lgbm | lgbm | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| sgd | assoc | none | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| rf | wrap | none | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| rf | pred | none | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| rf | embed_lgbm | lgbm | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| rf | assoc | none | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| lr | wrap | none | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| lr | pred | none | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| lr | embed_linear | linear | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| lr | embed_lgbm | lgbm | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| lgbm | wrap | none | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| sgd | wrap | none | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| lgbm | embed_lgbm | lgbm | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| lgbm | assoc | none | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| knn | wrap | none | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| lgbm | embed_linear | linear | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| knn | embed_lgbm | lgbm | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| lr | none | none | 0.997 | 1.000 | 0.997 | 0.997 | 1.000 | 0.995 | 0.997 | 0.995 |
| knn | pred | none | 0.960 | 0.992 | 0.959 | 0.960 | 1.000 | 0.929 | 0.959 | 0.917 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| sgd | pred | none | 0.957 | 0.995 | 0.956 | 0.957 | 0.995 | 0.928 | 0.956 | 0.917 |
| knn | none | none | 0.828 | 0.825 | 0.825 | 0.825 | 0.899 | 0.783 | 0.825 | 0.727 |
| knn | assoc | none | 0.823 | 0.944 | 0.817 | 0.815 | 0.984 | 0.748 | 0.817 | 0.644 |
| knn | embed_linear | linear | 0.670 | 0.799 | 0.662 | 0.643 | 0.820 | 0.623 | 0.662 | 0.413 |
| dummy | embed_linear | linear | 0.520 | 0.500 | 0.521 | 0.520 | 0.505 | 0.538 | 0.521 | 0.557 |
| lgbm | none | none | 0.515 | 0.500 | 0.500 | 0.340 | nan | 0.515 | 0.500 | 0.000 |
| rf | embed_linear | linear | 0.515 | 0.500 | 0.500 | 0.340 | nan | 0.515 | 0.500 | 0.000 |
| rf | none | none | 0.515 | 0.500 | 0.500 | 0.340 | nan | 0.515 | 0.500 | 0.000 |
| dummy | assoc | none | 0.515 | 0.500 | 0.500 | 0.340 | nan | 0.515 | 0.500 | 0.000 |
| dummy | pred | none | 0.500 | 0.500 | 0.500 | 0.500 | 0.484 | 0.516 | 0.500 | 0.500 |
| dummy | none | none | 0.495 | 0.530 | 0.495 | 0.491 | 0.478 | 0.512 | 0.495 | 0.500 |
| dummy | wrap | none | 0.487 | 0.556 | 0.487 | 0.486 | 0.476 | 0.498 | 0.487 | 0.469 |
| dummy | embed_lgbm | lgbm | 0.482 | 0.530 | 0.481 | 0.479 | 0.463 | 0.499 | 0.481 | 0.433 |

### 2.2 Derek’s feature selection method results

For Derek's feature selection task, the wrapper-based method has been used for the full dataset, male and female was the setup approach with liner as wrapper model and hs redundancy aware processing enabled.
Below is a summary of the findings for each of the 3 df-analyze runs followed by selected features.

**Full Dataset:**

- Wrapper method: StepUp
- Wrapper model: Linear
- Redundancy-aware: True
- Selected Features: ['Prescription_nan', 'HeartRate']

For the full dataset, the most important features selected by the model are related to the missing prescription data Prescription_nan) and the heart rate, which suggests that these may be the key predictors for the outcome in this specific dataset

**Female Dataset:**

- Wrapper method: StepUp
- Wrapper model: Linear
- Redundancy-aware: True
- Selected Features: ['Prescription_nan', 'Dominant_Hand_nan', 'BloodOxygenLevel']

For the Female dataset, the selected features are missing prescription information (Prescription_nan), dominant hand (Dominant_Hand_nan), and blood oxygen levels, which indicate that physical and prescription data can be significant for predicting dementia in females.

**Male Dataset:**

- Wrapper method: StepUp
- Wrapper model: Linear
- Redundancy-aware: True
- Selected Features: ['Dosage_in_mg_NAN', 'Family_History_nan', 'HeartRate']

For the male datasets, the most important key feature is the missing dosage information (Dosage_in_mg_NAN), family history (Family_History_nan), and heart rate. These might likely to contribute to the prediction accuracy for families

## 3. References


1. “Dementia Prediction Dataset”. Retrieved from https://www.kaggle.com/code/mdismielhossenabir/dementia-health-prediction/input
2. “Df-analyze python package”. Retrieved from GitHub Repository https://github.com/stfxecutables/df-analyze
3. Wikipedia contributors. (2024, October 10). *Automated machine learning*. Wikipedia. https://en.wikipedia.org/w/index.php?title=Automated_machine_learning&oldid=1193286380


# Appendix

### A.Mean and Standard deviation table for male and Female dataset

Table 1.2.1

**Table 1.2.1 Continuous values for male Dataset**

| Feature | Group of Interest (Dementia =1) | | Not Group of Intrest (Dementia =0) | |
|---|---|---|---|---|
| | Mean | Std Dev | Mean | Std Dev |
| Diabetic | 0.547 | 0.498 | 0.501 | 0.0500 |
| Alcohol Level | 0.099 | 0.057 | 0.101 | 0.057 |
| Heart Rate | 80.427 | 12.096 | 79.607 | 12.089 |
| BloodOxygen | 94.902 | 2.809 | 95.513 | 3.010 |
| BodyTemp | 36.825 | 0.426 | 36.719 | 0.430 |
| Weight | 72.947 | 15.386 | 75.170 | 13.882 |

| | | | | |
|---|---|---|---|---|
| Age | 9.060 | 6.599 | 76.070 | 8.567 |
| CognitiveTest | 74.526 | 9.443 | 9.027 | 0.815 |
| Dosage_in_mg | 3.543 | 2.301 | NA | NA |
| Dementia | 1 | 0 | 0 | 0 |

Table 1.2.2

**Table 1.2.2. Categorical Values for Male Dataset**

| Column | Value | Group of Interest Dementia=1 | Group of Not-Interest Dementia=0 |
|---|---|---|---|
| Prescription | Galantamine | 54 | Na |
| | Donepezil | 58 | Na |
| | Rivastigmine | 67 | Na |
| | Memantine | 62 | Na |
| Education level | Primary School | 92 | 105 |
| | Secondary School | 82 | 74 |
| | Diploma/Degree | 0 | 44 |
| | No School | 85 | 32 |
| Dominanat Hand | Left | 125 | 140 |
| | Right | 116 | 115 |
| Gender | Male | 241 | 0 |
| | Female | 0 | 0 |

| | | | |
|---|---|---|---|
| Family History | No | 122 | 107 |
| | Yes | 119 | 148 |
| Smoking Status | Current smoker | 0 | 49 |
| | Former Smoker | 120 | 101 |
| | Never Smoker | 121 | 105 |
| APOE_e4 | Positive | 215 | 134 |
| | Negative | 26 | 121 |
| Physical Activity | Sedentary | 79 | 90 |
| | Moderate Activity | 75 | 85 |
| | Mild Activity | 87 | 80 |
| Depression status | Yes | 116 | 0 |
| | No | 125 | 225 |
| Medical history | Yes | 122 | 125 |
| | No | 119 | 130 |
| Nutrition diet | Low Carb-78 | 81 | 78 |
| | Meditate | 87 | 92 |
| | Balanced | 73 | 85 |
| Sleep Quality | Poor | 125 | 135 |
| | Good | 116 | 120 |
| Chronic health condition | Diabetes | 132 | 128 |
| | Heart Disease | 32 | 37 |
| | Hypertension | 33 | 42 |
| | Nan | 44 | 48 |

Table 1.3.1

| The table 1.3.1 Continuous values for Female dataset samples. | | | | |
|---|---|---|---|---|
| **Feature** | **Group of Interest (Dementia=1)** | | **Group of Not-Interest (Dementia=0)** | |
| | Mean | Std Dev | Mean | Std Dev |
| Diabetic | 0.536 | 0.499 | 0.491 | 0.500 |
| Alcohol Level | 0.098 | 0.059 | 0.098 | 0.0583 |
| Heart Rate | 79.536 | 12.127 | 79.238 | 12.098 |
| Blood Oxygen | 95.010 | 2.866 | 95.42 | 2.957 |
| Body Temp | 36.776 | 0.423 | 36.747 | 0.438 |
| Weight | 73.584 | 15.027 | 75.016 | 13.890 |
| Dosage_in_mg | 9.213 | 6.493 | NA | NA |
| Age | 74.325 | 9.339 | 75.456 | 8.832 |
| Cognitive | 3.620 | 2.310 | 8.984 | 0.808 |
| Dementia | 1 | 0 | 0 | 0 |

Table: 1.3.2

| Table 1.3.2 Categorical values for Female dataset samples. | | | |
|---|---|---|---|
| Column | Value | Group of Interest Dementia=1 | Group of Not-Interest Dementia=0 |
| Prescription | Galantamine | 71 | Na |
| | Donepezil | 55 | Na |
| | Rivastigmine | 52 | Na |
| | Memantine | 66 | Na |
| Education level | Primary School | 89 | 103 |
| | Secondary School | 73 | 75 |

| | | | |
|---|---|---|---|
| | Diploma/Degree | 29 | 57 |
| | No School | 53 | 25 |
| Dominanat Hand | Left | 126 | 128 |
| | Right | 118 | 132 |
| Gender | Male | 0 | 0 |
| | Female | 244 | 260 |
| Family History | No | 133 | 118 |
| | Yes | 112 | 142 |
| Smoking Status | Current smoker | 0 | 41 |
| | Former Smoker | 132 | 105 |
| | Never Smoker | 112 | 114 |
| APOE_e4 | Positive | 220 | 125 |
| | Negative | 24 | 135 |
| Physical Activity | Sedentary | 79 | 83 |
| | Moderate Activity | 83 | 75 |
| | Mild Activity | 82 | 102 |
| Depression status | Yes | 129 | 0 |
| | No | 115 | 260 |
| Medical history | Yes | 129 | 138 |
| | No | 115 | 122 |
| Nutrition diet | Low Carb-78 | 81 | 90 |
| | Meditate | 82 | 77 |
| | Balanced | 81 | 93 |
| Sleep Quality | Poor | 139 | 135 |
| | Good | 105 | 125 |

| | | | |
|---|---|---|---|
| Chronic health condition | Diabetes | | 125 |
| | Heart Disease | 33 | 53 |
| | Hypertension | 39 | 39 |
| | Nan | 44 | 43 |

## B. Statistical analysis Table for overall accuracy of Male and Female

Table 3.1

**Table 3.1: Male Dataset Overall Accuracy**

| model | selection | embed_selector | acc |
|---|---|---|---|
| sgd | wrap | none | 1.000 |
| rf | none | none | 1.000 |
| lgbm | pred | none | 1.000 |
| lgbm | embed_linear | linear | 1.000 |
| lgbm | assoc | none | 1.000 |
| knn | wrap | none | 1.000 |
| rf | embed_linear | linear | 1.000 |
| sgd | pred | none | 0.995 |
| sgd | embed_lgbm | lgbm | 0.985 |
| sgd | embed_linear | linear | 0.985 |
| sgd | assoc | none | 0.985 |

| | | | |
|---|---|---|---|
| sgd | none | none | 0.970 |
| knn | pred | none | 0.960 |
| lgbm | embed_lgbm | lgbm | 0.887 |
| knn | embed_lgbm | lgbm | 0.854 |
| knn | assoc | none | 0.749 |
| knn | none | none | 0.744 |
| knn | embed_linear | linear | 0.613 |
| dummy | embed_lgbm | lgbm | 0.533 |
| rf | wrap | none | 0.513 |
| rf | pred | none | 0.513 |
| dummy | assoc | none | 0.513 |
| rf | embed_lgbm | lgbm | 0.513 |
| rf | assoc | none | 0.513 |
| lgbm | wrap | none | 0.513 |
| dummy | wrap | none | 0.513 |
| dummy | pred | none | 0.513 |
| dummy | embed_linear | linear | 0.513 |
| lgbm | none | none | 0.513 |

| | | | |
|---|---|---|---|
| dummy | none | none | 0.437 |

Table 3.2

**Table 3.2: Female Dataset Overall Accuracy**

| Model | Selection | Embed_Selector | Acc |
|---|---|---|---|
| SGD | Wrap | None | 1.000 |
| LGBM | Embed_Linear | Linear | 1.000 |
| SGD | Embed_LGBM | LGBM | 1.000 |
| RF | None | None | 1.000 |
| RF | Embed_Linear | Linear | 1.000 |
| LR | Embed_Linear | Linear | 1.000 |
| KNN | Wrap | None | 1.000 |
| LR | Embed_LGBM | LGBM | 1.000 |
| LGBM | Pred | None | 1.000 |
| SGD | Pred | None | 0.995 |
| SGD | None | None | 0.995 |
| SGD | Assoc | None | 0.995 |
| LR | Pred | None | 0.995 |

| | | | |
|---|---|---|---|
| LR | None | None | 0.995 |
| LR | Assoc | None | 0.990 |
| SGD | Embed_Linear | Linear | 0.985 |
| KNN | Embed_Linear | Linear | 0.876 |
| KNN | None | None | 0.837 |
| KNN | Embed_LGBM | LGBM | 0.805 |
| KNN | Assoc | None | 0.792 |
| KNN | Pred | None | 0.763 |
| LR | Wrap | None | 0.705 |
| Dummy | Embed_Linear | Linear | 0.534 |
| LGBM | None | None | 0.515 |
| RF | Pred | None | 0.515 |
| Dummy | None | None | 0.515 |
| Dummy | Wrap | None | 0.515 |
| RF | Wrap | None | 0.515 |
| RF | Embed_LGBM | LGBM | 0.515 |
| LGBM | Embed_LGBM | LGBM | 0.515 |

| | | | |
|---|---|---|---|
| RF | Assoc | None | 0.515 |
| LGBM | Assoc | None | 0.515 |
| LGBM | Wrap | None | 0.515 |
| Dummy | Assoc | None | 0.505 |
| Dummy | Embed_LGBM | LGBM | 0.485 |
| Dummy | Pred | None | 0.445 |

### C.5-fold performance on the holdout set result Table for Female and Male dataset

Table: 2.2

**Table: 2.2. 5-fold performance on holdout set table for Female Dataset**

| Model | Selection | Embed_Selector | Acc | AUROC | Bal-Acc | F1 | NPV | PPV | Sens | Spec |
|---|---|---|---|---|---|---|---|---|---|---|
| SGD | Wrap | None | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| LGBM | Embed_Linear | Linear | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| SGD | Embed_LGBM | LGBM | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| RF | None | None | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| RF | Embed_Linear | Linear | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| LR | Embed_Linear | Linear | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| KNN | Wrap | None | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| LR | Embed_LGBM | LGBM | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| LGBM | Pred | None | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| SGD | Pred | None | 0.995 | 1.000 | 0.995 | 0.995 | 1.000 | 0.991 | 0.995 | 0.990 |
| SGD | None | None | 0.995 | 1.000 | 0.995 | 0.995 | 1.000 | 0.991 | 0.995 | 0.990 |
| SGD | Assoc | None | 0.995 | 1.000 | 0.995 | 0.995 | 1.000 | 0.991 | 0.995 | 0.990 |
| LR | Pred | None | 0.995 | 1.000 | 0.995 | 0.995 | 1.000 | 0.991 | 0.995 | 0.990 |
| LR | None | None | 0.995 | 1.000 | 0.995 | 0.995 | 1.000 | 0.991 | 0.995 | 0.990 |
| LR | Assoc | None | 0.990 | 1.000 | 0.990 | 0.990 | 1.000 | 0.982 | 0.990 | 0.979 |
| SGD | Embed_Linear | Linear | 0.985 | 0.986 | 0.986 | 0.985 | 0.973 | 1.000 | 0.986 | 1.000 |
| KNN | Embed_Linear | Linear | 0.876 | 0.996 | 0.872 | 0.873 | 1.000 | 0.809 | 0.872 | 0.745 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| KNN | None | None | 0.837 | 0.832 | 0.832 | 0.832 | 0.960 | 0.772 | 0.832 | 0.694 |
| KNN | Embed_LGBM | LGBM | 0.805 | 1.000 | 0.800 | 0.736 | 1.000 | 0.805 | 0.800 | 0.600 |
| KNN | Assoc | None | 0.792 | 0.922 | 0.785 | 0.778 | 0.971 | 0.723 | 0.785 | 0.590 |
| KNN | Pred | None | 0.763 | 0.903 | 0.756 | 0.748 | 0.935 | 0.696 | 0.756 | 0.551 |
| LR | Wrap | None | 0.705 | 1.000 | 0.700 | 0.602 | 1.000 | 0.705 | 0.700 | 0.400 |
| Dummy | Embed_Linear | Linear | 0.534 | 0.559 | 0.534 | 0.526 | 0.538 | 0.535 | 0.534 | 0.491 |
| LGBM | None | None | 0.515 | 0.500 | 0.500 | 0.340 | NaN | 0.515 | 0.500 | 0.000 |
| RF | Pred | None | 0.515 | 0.500 | 0.500 | 0.340 | NaN | 0.515 | 0.500 | 0.000 |
| Dummy | None | None | 0.515 | 0.500 | 0.500 | 0.340 | NaN | 0.515 | 0.500 | 0.000 |
| Dummy | Wrap | None | 0.515 | 0.500 | 0.500 | 0.340 | NaN | 0.515 | 0.500 | 0.000 |
| RF | Wrap | None | 0.515 | 0.500 | 0.500 | 0.340 | NaN | 0.515 | 0.500 | 0.000 |
| RF | Embed_LGBM | LGBM | 0.515 | 0.500 | 0.500 | 0.340 | NaN | 0.515 | 0.500 | 0.000 |
| LGBM | Embed_LGBM | LGBM | 0.515 | 0.500 | 0.500 | 0.340 | NaN | 0.515 | 0.500 | 0.000 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| RF | Assoc | None | 0.515 | 0.500 | 0.500 | 0.340 | NaN | 0.515 | 0.500 | 0.000 |
| LGBM | Assoc | None | 0.515 | 0.500 | 0.500 | 0.340 | NaN | 0.515 | 0.500 | 0.000 |
| LGBM | Wrap | None | 0.515 | 0.500 | 0.500 | 0.340 | NaN | 0.515 | 0.500 | 0.000 |
| Dummy | Assoc | None | 0.505 | 0.500 | 0.508 | 0.498 | 0.493 | 0.524 | 0.508 | 0.583 |
| Dummy | Embed_LGBM | LGBM | 0.485 | 0.500 | 0.484 | 0.480 | 0.469 | 0.499 | 0.484 | 0.467 |
| Dummy | Pred | None | 0.445 | 0.550 | 0.446 | 0.443 | 0.435 | 0.455 | 0.446 | 0.448 |

Table 2.3

**Table: 2.3. 5-fold performance on holdout set table for Male Dataset**

| model | selection | embed_selector | acc | auroc | bal-acc | f1 | npv | ppv | sens | spec |
|---|---|---|---|---|---|---|---|---|---|---|
| sgd | wrap | none | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| rf | none | none | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| lgbm | pred | none | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| lgbm | embed_linear | linear | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| lgbm | assoc | none | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| knn | wrap | none | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| rf | embed_linear | linear | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| sgd | pred | none | 0.995 | 0.995 | 0.995 | 0.995 | 0.990 | 1.000 | 0.995 | 1.000 |
| sgd | embed_lgbm | lgbm | 0.985 | 0.984 | 0.984 | 0.985 | 1.000 | 0.973 | 0.984 | 0.969 |
| sgd | embed_linear | linear | 0.985 | 0.998 | 0.984 | 0.985 | 1.000 | 0.973 | 0.984 | 0.968 |
| sgd | assoc | none | 0.985 | 0.985 | 0.985 | 0.985 | 0.982 | 0.990 | 0.985 | 0.989 |
| sgd | none | none | 0.970 | 0.979 | 0.970 | 0.970 | 0.979 | 0.963 | 0.970 | 0.958 |
| knn | pred | none | 0.960 | 0.985 | 0.959 | 0.960 | 1.000 | 0.931 | 0.959 | 0.919 |
| lgbm | embed_lgbm | lgbm | 0.887 | 0.908 | 0.886 | 0.883 | 0.877 | 0.892 | 0.886 | 0.853 |
| knn | embed_lgbm | lgbm | 0.854 | 0.852 | 0.852 | 0.851 | 0.960 | 0.792 | 0.852 | 0.732 |
| knn | assoc | none | 0.749 | 0.822 | 0.744 | 0.737 | 0.888 | 0.689 | 0.744 | 0.556 |
| knn | none | none | 0.744 | 0.886 | 0.737 | 0.723 | 0.947 | 0.676 | 0.737 | 0.503 |
| knn | embed_linear | linear | 0.613 | 0.715 | 0.605 | 0.578 | 0.705 | 0.586 | 0.605 | 0.358 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| dummy | embed_lgbm | lgbm | 0.533 | 0.557 | 0.535 | 0.528 | 0.528 | 0.547 | 0.535 | 0.537 |
| rf | wrap | none | 0.513 | 0.500 | 0.500 | 0.339 | nan | 0.513 | 0.500 | 0.000 |
| rf | pred | none | 0.513 | 0.500 | 0.500 | 0.339 | nan | 0.513 | 0.500 | 0.000 |
| dummy | assoc | none | 0.513 | 0.500 | 0.500 | 0.339 | nan | 0.513 | 0.500 | 0.000 |
| rf | embed_lgbm | lgbm | 0.513 | 0.500 | 0.500 | 0.339 | nan | 0.513 | 0.500 | 0.000 |
| rf | assoc | none | 0.513 | 0.500 | 0.500 | 0.339 | nan | 0.513 | 0.500 | 0.000 |
| lgbm | wrap | none | 0.513 | 0.500 | 0.500 | 0.339 | nan | 0.513 | 0.500 | 0.000 |
| dummy | wrap | none | 0.513 | 0.500 | 0.500 | 0.339 | nan | 0.513 | 0.500 | 0.000 |
| dummy | pred | none | 0.513 | 0.500 | 0.500 | 0.339 | nan | 0.513 | 0.500 | 0.000 |
| dummy | embed_linear | linear | 0.513 | 0.500 | 0.500 | 0.339 | nan | 0.513 | 0.500 | 0.000 |
| lgbm | none | none | 0.513 | 0.500 | 0.500 | 0.339 | nan | 0.513 | 0.500 | 0.000 |
| dummy | none | none | 0.437 | 0.500 | 0.436 | 0.433 | 0.415 | 0.454 | 0.436 | 0.391 |

—-----—---------- END OF THE REPORT —----------

## Files Included

### 1. Results Files

- `Main_Dataset.zip`
  The generated result is shown in this folder for “dementia_patients_health.csv” (Full Dataset)
- `Male_Dataset.zip`
  The generated result is shown in this folder for the Male Dataset
- `Female_Dataset.zip`
  The generated result is shown in this folder for the Female Dataset

### 2. Analysis Scripts

- **`df-analyze-script.sh`**
  A shell script containing the commands to run `df_analyze` on each dataset.

### 3. Python Script

- **`mean_SD_CALC.py`**
  A Python script that calculates the **mean** and **median** values for any provided dataset (CSV format). It reads the CSV file and computes the statistics. Also, new Python files will be generated when we execute this file.