MLD Project Phase 4

Topic: Dementia Prediction using ML

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

2. Gagandeep Singh [Student ID: **202303876**]

Abstract

The study focuses on using machine learning to predict dementia, a complex and increasingly prevalent health condition.[3] The main dataset consists of data samples of **1000** patients with **23** columns[3]. Some of the features presented are Age, Heart_Rate, BloodOxygenLevels, Weight, with lifestyle factors like Smoking_Status, Physical_Activity, and Depression_Status and other category genetic information such as APOE_ε4 and Family_History of dementia[3]. For the final phase, we removed the **Prescription**, **Dosage in mg** and **Cognitive_Test _Scores** columns from all of the runs as per the feedback given in the phase 3 results. We integrated Gandalf in our runs to observe differences in results from the previous results and existing publications. Interestingly, **LGBM** and **SGD** have been seen to be the best performers for most of the runs and we have seen the drop in accuracy when columns have been removed. However, Gandalf showed **100%** accuracy for the run for a full dataset with all columns presented along with other models presented.

Introduction

Dementia is a progressive and debilitating condition affecting millions worldwide, characterized by cognitive decline that interferes with daily functioning[1]. Early detection is crucial to managing symptoms and improving patient outcomes[5]. Using a dataset[3] from Kaggle that includes both demographic factors and health-related metrics[3], we examined the effectiveness of these models in predicting dementia across multiple experimental conditions. These conditions included full dataset runs, gender-specific segmentations, and selective removal of key feature columns to observe the models' sensitivity to feature availability. By comparing the predictive accuracy of each model across these variations, we aim to focus on the role of specific features, also, we assess Gandalf’s potential for traditional ML models, particularly for applications that require deep learning in handling tabular complex feature selection. Our findings could contribute to insights into feature selection, model robustness and potential impact on early dementia detection with implications for developing more targeted predictive tools in healthcare.

1. Results

1.1 Comparison of Phase 2 3, and Current Results

Phase 2 Results: DF Analyze Run

For Phase 2 of the df-analyze process run, we have conducted a run on the full dataset and have come to the result that the best-performing model is **LGBM** with [none, pred, assoc], **SGD** with [embed_linear, assoc], and **Linear Regression (LR)** with [pred, embed_linear, assoc]. The following models have given impressive results with the most predictions with **100% accuracy**. The performance of these models has shown the effectiveness of the selected feature sets and high accuracy.

Phase 3: Dataset Segmentation and Accuracy Comparison

For phase 3, we have introduced the segregation of the full dataset which includes all samples into a **male** dataset and **female** dataset, which is based on the feedback from the previous phases. After that run, for the **full dataset**, we observed similar results of phase 2, with similar leading models having high accuracy. However, **KNN h**as been seen as a new strong performer which has achieved **100%** accuracy while using **embed_linear** selection.

Furthermore, when we divided the dataset into a **male** and **female** separate dataset and ran them separately in df-analzye, distinct patterns were noticed in the model performance. With respective of **Male Dataset**, we can see that the best-performing models are **SGD** with wrap, **RF** with none, **LGBM** with pred, embed_linear, assoc, and **KNN** with wrap. When observing the results of the **Female** dataset, the leading model is observed to be **SGD** with wrap, **RF** with embed_linear, **LGBM** with embed_linear, assoc, and pred, as well as **LR** with wrap and embed_lgbm. If we list down the features selected for the full dataset which is, for males ['Prescription_nan', 'Nutrition_Diet_Mediterranean Diet', 'Weight'], and was seen similar for Females, these described results have shown that different feature selections are well more suited for each gender-specific dataset which has led to variations in prediction accuracy.

Phase 4: Integration of Gandalf and Updated DF Analyze Results

Phase 4 has a significant update with the integration of **Gandalf** which is a deep-learning model designed especially for tabular data, which is integrated into **df-analzye.[4]** so as per the feedback now received from phase 4, we have removed columns **[Prescription], [Dosage in mg]** and **[Cognitive_Test_Scores]** from the **full dataset**, **male dataset**, and **female dataset**. Now, it is notable that for phase 4, after the feature columns are removed, we have seen drastic results differences in results compared to earlier phases; which is discussed further down.

Now, for the **Full Dataset**, the leading prediction model is **SGD** with none selection having **75.7%** accuracy. Noticeably, there is not much difference in accuracy as compared to the other results, but has dropped approx. 25% as compared to all columns presented.

Observing the **Male** dataset results, **LGBM** is the best model selected with none. LGBM showed **76.2% accuracy**. As mentioned before, some columns are more important and show better correlations, that is why by removing that column we can see the drop in overall accuracy.

Lastly, results for the **Female dataset** run, we can see the top-performing model as **LGBM** with **wrap** having **75.8%** accuracy. With all columns available, LGBM showed 100% accuracy.

Noteworthy Differences and Impact of Gandalf

The introduction of **Gandalf** and removing features across the dataset in phase 4 has shown up noteworthy difference compared to the results of the earlier phases.

The prediction accuracy for the full dataset with some columns removed underscores a potential that **Gandalf** in predictive modelling for tabular data, which also demonstrated its capability to match the accuracy of best-performing models and may provide valuable insights for future use of this model i.e df-analzye with Gandalf.

1.2 GANDALF Assessment

In this phase of analysis, we have integrated **Gandalf**, which is a deep learning model specially designed for tabular datasets integrated within df-analzye[4] and had successful runs on all 3 datasets i.e **full dataset**, the **male dataset**, the **female datase**t and has removed feature columns as per the feedback given in phase 3.

For the **Full dataset, Gandalf** has achieved a predictive accuracy of **predictive accuracy of 72.6%** and is in 19th position in the result list compared to other models. Although the accuracy is not as high as compared to leading models, it is still reasoning closely which means that Gandalf has performed well on this dataset variant. The **embed_linear** model selection has been chosen as bets performing for Gandalf which indicates the feature is partially relevant to the model's prediction capability.

For the **male dataset**, Gandalf's performance has been lower compared to the full dataset, with a **predictive accuracy of 66.6%**, placing it in the **23rd position** in the result section list. For this dataset, Gandalf selected **none** as the model selection. This also suggests, that Gandalf still performed well, but was not able to predict as effectively as other models on the male dataset, as we have the **LGBM** model with **76.2%** accuracies which is on top, where selected features may not have aligned with the dataset characteristics, especially with Gandalf.

Lastly, for the **female** dataset, Gandalf has shown a **predictive accuracy of 73.4%**, which is ranked **8th position** in the results list. In this instance, Gandalf has selected **pred** as the model selection method, which highlights that the selected features for the female dataset are more relevant to the model's predictions which leads to better performance compared to the male dataset.

In comparison to previous leading models, Gandalf's performance can be stated as mixed. While it has accuracy not being the highest, it still has demonstrated potential, especially for the f**ull dataset** and **female dataset.** Gandalf's selection methods here which include —**embed_linear**, **none**, and **pred**— have a very important role in determining accuracy and prediction. We can state that the model selections are tailored to specific datasets but also have not outperformed the previous leading models like **LGBM** and **SGD** in most run cases.

Now, when combined with **Derek's feature selection method,** we observe that Gandalf was not able to emerge overall leading method, nor did it surpass other learners using **Derek's feature selection.** However, we can still consider it as having the ability in certain datasets that would be a valuable addition to the already existing set of models of predictive tasks specifically which when combined with other models or optimized further.

2. Discussion

For this phase, of our analysis, the most interesting findings which came across were the performance model of **LGBM** and **SGD** models, which have been the leading machine learning techniques across all the 3 datasets. Both of these models were able to achieve an accuracy of approx **75%**. Specifically, **LGBM** has been the top performer across most of the dataset results compared to the previous results. This result is partially noteworthy as it aligns with the hypothesis that if certain feature columns have been removed, in our case [Prescrription, Dosage in mg, Cognitive_Test_Scores] it would impact the overall performance of the model as in previous phases, we see that the machine is biased towards specific column in the dataset. The accuracy dropped after these columns had been removed. This work highlights the value and importance of different machine learning models and their sensitivity to dataset feature column changes, which provide insights into the robustness limitations and biases by ML for predicting dementia.

Hypothesis 1: Full Dataset [Removed Columns]

We primarily hypothesized based on phase 3 feedback, that we could have significant changes in prediction accuracy if we remove certain columns from the full dataset as we see the machine is being biased towards certain columns. Now, Upon removing specific columns which are **Prescription**, **Dosage_in_mg**, and **Cognitive_Test_Scores**, we see a notable drop in accuracy with this run **full dataset** accuracy from near 100% to **75.6%.** These changes confirm that our hypothesis that feature columns have a role in predicting dementia and that the ML model is biased toward predicting dementia is true.

Hypothesis 2: Male Dataset

This dataset is segmented and focuses on male dataset entries from our main full dataset entries which includes all patients with male gender to explore gender-based differences while removing certain feature columns and adding **Gandalf** from the dataset. We saw that in this run, **LGBM** emerged as the leading model with an accuracy of **76.2%**. Interestingly, the accuracy performance for this dataset has shown more accuracy as compared to the full dataset run which could be due to many reasons one of which could be that it is focused on the one kind of attribute which could have strong correlations. Which predicts that the machine is ias towards some attributes.

Hypothesis 3: Female Dataset

Similarly, we have extracted female entries which correspond to the female gender to further hypothesize on gender-specific trends in prediction accuracy by removing certain feature columns. For the **female dataset**, **LGBM** has again been the leading model with an accuracy of **75.8%.** The performance is similar but has a slight reduction of accuracy compared to the male dataset. The results now suggest that gender does affect the performance of predictive models even if the difference in accuracy is slightly smaller, it might still be valuable information for long-term and more advanced ML model development for dementia prediction.

Hypothesis 4: Full Dataset [Without Removed Columns]

For the full dataset run, which includes all samples and no feature columns being removed, the hypothesis is to predict dementia and observe the leading models with Gandalf. In this latest full dataset run, the leading models **LGBM**, **SGD**, **LR**, **RF**, and **Gandalf**—achieved **100% accuracy**, with almost half of the models in these selections showing perfect performance. These results are noticeable as they indicate that the full dataset if properly trained and modelled, allows for accurate prediction of dementia which supports the hypothesis that a comprehensive approach can successfully predicate the target variable. The addition of **Gandalf** in this run along with other models with perfect accuracy is also an important observation showing the potential of deep learning in this dataset and domain.

2.1 Literature Comparison

The dataset[3] we have chosen from Kaggle does not have a research article or journal directly related to or published on it. However, in Phase 2, we referenced a research paper that is closely related and similar to our dataset. In this section, we compare the results of our Phase 4 analysis with the findings from this existing paper.

One of the closest studies based on our topic is the paper titled **"Early Prediction of Dementia Using Feature Extraction Battery (FEB) and Optimized Support Vector Machine (SVM) for Classification"** [2]. In this study, the authors have proposed a model using a Feature Extraction Battery (FEB) combined with an optimized Support Vector Machine (SVM) with a radial basis function for the classification of dementia.[2] In this study, they have an accuracy of **98.28%** on

their training dataset and **93.92% and 93.92%** on their testing set.[2] With this, the precision is reported as **91.80%** with a recall value of **86.59** an f1 score of **89.12% and** Matthew's Correlation Coefficient (MCC) of **0.4987.** [2]

While comparing the article results with the leading model of our dataset, we found that for the **Full Dataset** with all features columns available, around 50% of models showed **100% accuracy.** However, when we removed certain feature columns from the dataset, the accuracy dropped to **75.6%** with precision and the F1-scores significantly reduced to **50%** and **75.5% respectively**. This drop in performance showcases the impact of removing feature columns which is similar and aligns with findings from the study.

As we have hypothesized in phase 3, the precision has been decreased when specific columns have been removed. Df-analzye performed well with all columns but has shown decreased accuracy when certain columns have been executed. The authors of the referenced study have noticed similar challenges and have tried to solve them stating that since dementia is a rare disease, the classes in their dataset which are highly imbalanced can be biased machine learning model [2] the observation with this reference stands true for our dataset as well. This can conclude that column removal can result in reduced accuracy btu as also can increase the bias in the model for prediction and classification.

So to conclude with this literature comparison in this section, the comparison sheds light on challenges which are faced with imbalanced datasets regarding column removal and model being biased. These results show the importance of careful feature selection, and its influence on model performance while dealing with complex and incurable diseases like dementia Further research can also explore various methods to address these bas or imbalances and improve the overall accuracy and robustness of predicting models in this domain and also in some other domain facing same challenges.

2.2 Redundancy-Aware Feature Selection

Accuracy/Error of Wrap Models Compared to Best-Performing Learners:

The Wrap model accuracy on the full dataset is **100%,** whereas excluding specific columns reduces the performance to **75.2%** on the full dataset. When we analyze male and female datasets the accuracy are little different, with the male dataset achieving **76.2%** and the female dataset **75.8%.** These results suggest that the wrap model can achieve the best performance on the full dataset, but its accuracy reduces when specific features are excluded or when gender-specific datasets are used. This variation in performance shows an influence on specific features on predictive percentages on models.

Features Reliant on Performance:

On **full dataset**, the wrap models depend on variety set of features which include **[‘Dosage_in_mg_NAN’, ‘AlchoholLevel’, ‘Smoking_Status_Never Smoked’, ‘Depression_Status_Yes’, ‘Prescription_Mematime’, ‘Weight’, ‘Dominant_Hand_Right’, ‘Chronic_Health_Conditions_Hypertension’, ‘Nutrition_Diet_Meditarranean Diet’, ‘Chronic_Health_Conditions,_Heart Disease’, ‘Family_History_Yes’, ‘Prescription_Galantamine’, ‘Gender_Male’, ‘Physical_Activity_Moderate Activity’, ‘Nutritional_Diet_Low_Carb Diet’, ‘Education_Level_Secondary School’, ‘Dosage_in_mg’, ‘Physical_Activity_Mild Activity’, ‘Prescription_Denpezil’]**. The feature which is selected in this **full dataset run while removing certain columns** were now [**'Depression_Status_Yes', 'BodyTemperature**', **'Gender_Male', 'Weight', 'Age',** '**HeartRate' 'BloodOxygenLevel'**]. For the gender-specific datset of **Male** and **Female,** displayed different feature selection set; wherein the **Male dataset** had features like **[‘Depression_Status_Yes’, ’Education_LEvel_Primary School’, ‘Weight’, ‘BodyTempreture,’ ’Age’, ‘HeartRate’,’BloodOxygenLevel’]** while female dataset the features were **[‘Depression_Status_Yes’, ‘Diabetic_1’, ‘BodyTempreture Age’]**. Differences between these indicate feature selection depends on gender-specific characteristics of the dataset and has potential relevance for predicting the target variable.

Reason for Predictive Features:

The predictive features selected for the full dataset, are **'Gender_Male' and 'Depression_Status_Yes',** which is likely due to greater relevance based on these variables within the dataset. Gender plays a significant role with males, making it more frequent in the dataset which influences the selection of the '**Gender_Male'** as a very important feature. Furthermore, males tend to have a higher rate of depression and prescription medication usage which is further explained strong correlation between these features and also with **diabetes. So overall, we can see that these features may have an impotence in predicting dementia.**

Do the Results Make Sense? Could They Be Caused by a Bug or Dataset Peculiarity?

Yes, the results do make sense based on the dataset’s characteristics. Due to the high prevalence of features like ‘**Gender_Male**’ and ‘**Depression_Status_Yes**’, it is expected and relevant, also no bugs or dataset peculiarities were found. The change in accuracy when certain columns were removed from the dataset sounds logical as it shows the influence of these features on predictive performance. Thus, the results do align with the expected behaviour given the dataset’s composition and model’s design.

Causal Relationship Between Features and Target:

The relationship between the predictor features and the target variable seems to be influenced by underlying patterns in the data. For instance, the dominance of '**Gender_Male**' in the feature selection may reflect the imbalance in the dataset, where males have higher rates of certain health conditions. The presence of depression and chronic health conditions, often more prevalent in males in the dataset, might exacerbate the likelihood of developing diabetes, making these features predictive of the target.

Do Predictor Features Cause the Target or the Other Way Around?

In this case, it appears that the target variable depends on the predictor features. When specific columns were removed from the dataset, the model's performance decreased, which indicates that the features were tightly linked to the target variable. The changes in feature selection after removing columns suggest that the predictors are primarily driven by the target, and the relationship is not simply a reflection of causation in the reverse direction. This further supports the idea that the target variable influences the model's ability to predict accurately.

Speculation on Correlation and Causation:

In machine learning the correlations make more sense than the causation, i.e. the potential relations between the predictor features and target can help interpret the results. In this case, the correlations between '**Gender_Male**', '**Depression_Status**', and chronic health conditions, with **'Diabetic'** as the target, suggest a deeper relationship. It is plausible that underlying health conditions and lifestyle factors contribute to both the presence of diabetes and the predictor features, making the model's results more interpretable within the context of the health domain.

2.3 Discussion of GANDALF Results

In this phase, we have integrated and implemented **Gandalf** across various hypothetical runs which evaluates its performance on **full dataset**, **male dataset**, and **female dataset**. The results from this assessment show both the strengths and weekends of Gandalf.

For the full Dataset with all feature columns included Gandalf demonstrated **100% accuracy** which matches with performance of other models. But even though achieving the same accuracy, it is ranked below other models which show better results compared to Gandalf. This also suggests that even though Gnadalf is capable of showing high accuracy, it also did not have a distinct advantage over other models in this scenario. It shows equal performance which indicates that Gandalf being a potential deep learning model did not outperform our already leading methods like **LGBM** or **SGD**.

We observed a significant decline in Gandalf's accuracy when we examined datasets with certain feature columns being removed. For the **full Dataset,** the accuracy has dropped to **72.6%.** For the male dataset, the accuracy is dropped to **66.6%** and for female, it is even less to **62.3%**. This decline suggests that Galdaf is very sensitive to the presence of specific features in the dataset. When certain features are removed, the model's ability to make predictions has subsided highlighting a reliance on these key features for optimal performance and prediction. This also showcases that Gandalf was daily performed but was also not that robust compared to other prediction models which showed better consistency in prediction even with missing feature columns. The result from Gandalf suggests both its potential and limitations. One of the strengths is the ability to achieve **100%** accuracy when all features have been included, which demonstrates that it is been highly effective when it comes to a complete set of data with all columns included. However, the performance seems to decrease when some of the feature columns have been removed revealing its vulnerability to feature selection.

3. Conclusions.

In the project phase, our analysis of dementia prediction using various machine learning models, including the deep learner Gandalf, demonstrated the importance of feature selection and dataset configuration. LGBM and SGD contributed to being the most emerging model in our runs.

- For a **full dataset** with all columns included, the best prediction accuracy was **100%** with **LGBM**, and for with removed columns was **75.6%** with **SGD**
- For the **Male dataset**, with columns removed, the highest accuracy was **76.2%** with **LGBM**
- For the **Female dataset**, with columns removed, **LGBM** achieved **75.8%** accuracy.

When comparing these results with those from previous phases, **Gandalf** did not lead in performance, except when all features were included, where it showed **100% accuracy**, aligning with other models. A noteworthy observation from this phase was the significant drop in prediction accuracy when we removed specific columns—namely **'Prescription'**, **'Dosage in mg'**, and **'Cognitive_Test_Scores'**. This removal caused the accuracy to decrease drastically, with the performance dropping from **100%** to approximately **76%** across all models.

Overall, **DF-Analyze** was able to predict with fair accuracy in most of the runs, underscoring the importance of certain features in explaining the dataset. This suggests that identifying and prioritizing key features is crucial for improving prediction accuracy. These findings highlight that the machine and model were getting biased with certain columns and relying on prediction. When removed, the prediction accuracy was decreased which satisfies our hypothesis for this run. Looking ahead, further analysis of the identified important features could significantly enhance our ability to predict dementia accurately and provide opportunities for early treatment.

4. References

1. "World Health Organization. (2023). “Dementia”. Retrieved from https://www.who.int/news-room/fact-sheets/detail/dementia

2.Javeed, A., Dallora, A.L., Berglund, J.S., Idrisoglu, A., Ali, L., Rauf, H.T., & Anderberg, P. (2023). “Early prediction of dementia using feature extraction battery (FEB) and optimized support vector machine (SVM) for classification”. *Biomedicines, 11*(2), 439. https://doi.org/10.3390/biomedicines11020439.

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

4.“Df-analyze python package”. Retrieved from GitHub Repository https://github.com/stfxecutables/df-analyze

5."National Institute on Aging. (2021). “What Is Dementia? Symptoms, Types, and Diagnosis”. Retrieved from https://www.nia.nih.gov/health/alzheimers-and-dementia/what-dementia-symptoms-types-and-diagnosis

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

Files Included

1. Results Files

With all columns:

- `Full_Data_.zip`
- `Male_all.zip`
- `Female_all.zip`

With removed Columns:

- `Main_Dt_remove.zip`
- `Male_removed.zip`
- `Female_removed.zip`

2. Analysis Scripts

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

3. Python Script

- `MLD_W_Pres.py`

4. Dataset

- `Dementia_patients_health_data.csv`
- `Male.csv`
- `female.csv`
- `Dementia_wp_pres.csv`
- `Male_drop.csv`
- `Female_drop.csv`

5. Presentation Slides

- `MLD PHASE 4 ppt.pptx`
- `MLD PHASE 4 ppt.pdf`