ANALYSING AUSTRALIAN IMAGING BIOMARKER LIFESTYLE DATA USING NAÏVE BAYES

Abstract

The study used supervised and unsupervised machine learning models for predicting healthy control and non-healthy controls using AIBL lifestyle data. Supervised models of Naïve Bayes, XGBoost model, Decision Tree Model, and random tree forest were used for prediction and the analysis showed that random model summary was the best predictor showing 93% accuracy while Hierarchical clustering, K-means clustering, and Dimensionality reduction using PCA were the unsupervised learning models. The study showed that the unsupervised models are not suitable for this kind of data as the data is curated for supervised learning.

Keywords: Supervised learning, Unsupervised learning, AIBL Lifestyle Data

I. INTRODUCTION

The chance of developing dementia (typically caused by Alzheimer's Disease (AD)) is increased in people who have mild cognitive impairment (MCI), with an annual progression rate of up to 10-20 percent [1]. Although clinical criteria for mild cognitive impairment (MCI) and Alzheimer's disease (AD) have been developed to formalize the assessment of the gradual progression of the disease, it remains difficult to predict when and which individuals who meet the criteria for MCI will eventually progress to AD dementia at a given time at the outset. [1]

Early detection of Alzheimer's disease dementia has traditionally been modeled as a pattern classification challenge. By categorizing MCI individuals into two groups based on the length of their follow-up, for example, a binary classifier can be trained on baseline data to distinguish between progressing MCIs (pMCIs) and stable MCIs (sMCIs), and this may be used to discriminate between pMCIs and sMCIs. Machine learning techniques have been implemented to develop classifiers based on neuroimaging data to predict Alzheimer's disease dementia in its early stages.

For this study, When it comes to Alzheimer's disease (AD), which is the most frequent type of dementia, the data is heterogeneous. It has been observed that COVID-19 caused a significant worsening of the neuropsychiatric symptoms associated with dementia. Making use of the capabilities of artificial intelligence and machine learning to help healthcare across the board is important, and dementia is no exception, particularly in these extreme conditions. The data (http://adni.loni.usc.edu/data-samples/access-data/), which has been approved for use in teaching by the AIBL Management Committee, can be found in the BBL Assessment area. The Data Dictionary search box may be located at http://adni.loni.usc.edu/data-dictionary-search/, where you can look for the description of the data. Using AIBL data, three clinical diagnostic outcomes were identified: Healthy Control (HC), Mild Cognitive Impairment (MCI), and Alzheimer's Disease (AD) (AD). This coursework will look at two different types, namely HC and NonHC (combining MCI and AD).

This study looked at the dataset and attempted to control the data inconsistencies before attempting to combine the data to give sense to the dataset. The missing variables were subsequently filled in, which helped with feature engineering and modeling to properly forecast whether a person has HC or not, respectively.

A machine learning algorithm is an artificial intelligence technique for selecting the best model from a collection of possibilities to fit a set of data. It is used to find the best model for a piece of data. Machine learning algorithms provide many advantages, including nonlinearity, fault tolerance, and real-time operation. As a result, they are well suited for demanding applications. A growing number of precision psychiatric procedures are based on machine learning techniques. However, the majority of current research has depended on brain imaging equipment that is not practicable in real-world situations. This validates the method taken in this study, which is to use machine learning approaches for categorization rather than human judgment.

Machine learning and supervised algorithms such as Naïve Bayes, XGBoost< Neural networks can be used for the prediction of healthy and nonhealthy controls. By training the algorithms of data with known or unknown results, HC or non-HC can be determined with a high degree of certainty as against traditional and clinical models.

Using data from the Australian Imaging, Biomarker Lifestyle Flagship Study of Aging, our model is trained and validated (AIBL).

II. DATA PREPROCESSING

a. *Dataset Description*

The AIBL data has three clinical diagnostic results (DXCURREN). These are the Healthy Control (HC), Mild Cognitive Impairment (MCI), and Alzheimer's disease (AD). The data consists of nine different datasets. These are:

- Apoeres_data containing information on genotype
- Cdr_data containing CDR score on cognitive assessment
- Lab_data contains data on the Blood test
- Med_hist contains information about Medical history

- Mmse_data containing information on MMSE score on cognitive assessment
- Neurobat with information on Logical memory recall score Cognitive assessment
- Pdxconv containing information on Clinical diagnosis
- Ptdemog containing information on Gender and DoB.

b. Importing and Exploring the Dataset

The datasets were imported via the pd.read_csv () command and this was done for all the eight datasets. Some of the datasets were explored to review their characteristics. The next step down was to filter the "bl" before combining the datasets into a single data frame for ease of analysis. After combining the datasets in a single data frame, data with columns having the same information were dropped via df.drop([], axis=1) function. This was done as part of the data cleaning process, after combining the dataset, data was checked if there are duplicated data and the result returned as false, hence, this gives a clear match for data exploration.

III. EXPLORATORY DATA ANALYSIS

Any research analysis must include exploratory data analysis (EDA). The main goal of the exploratory analysis is to look for distribution, outliers, and anomalies in the data to lead to particular testing of your hypothesis. It also includes tools for generating hypotheses by viewing and comprehending data, which is commonly done through graphical representation. EDA is designed to aid the analyst in recognizing natural patterns. [2]

The characteristics of the new data such as df.info(), data shape, and columns were explored. The data were also checked for missing data. After this was done was the correlational analysis of the presented data. Correlation analysis is a widely used technique for identifying intriguing relationships in data that is becoming increasingly popular. These interactions assist us in determining the importance of traits to the target class to be forecasted. [3]. The correlation graph of all the 32 variables were shown in figure 1 below.

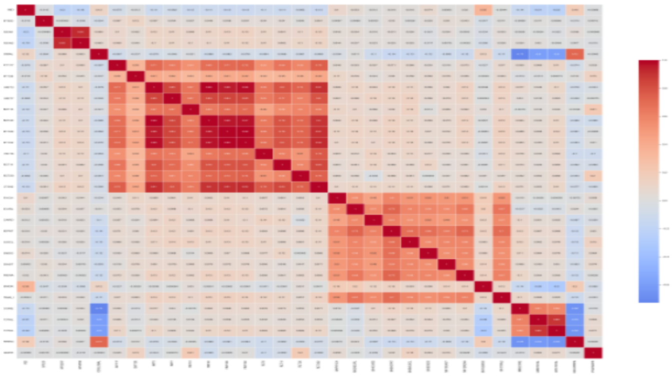
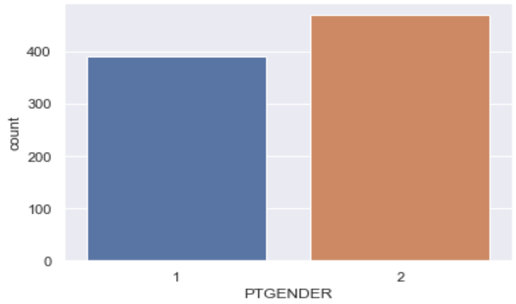


Figure 1: correlation Matrix of the variables

The red plots shows the variables that were strongly correlated while the blue shows variables with weak correlation. A copy reference was made for the variable to be able to reference and easily access the original data frame with the command df1=df.copy ()

IV. DATA CLEANING

In cleaning the data, columns with the same information and high cardinality were dropped while anomalies were discovered in one of the columns of PTDOB. The new data were then checked for missing values and there were no missing values in the data. It was observed that there are 862 entries and 31 variables in the dataset. Further exploration of the dataset reveals the properties of the dataset and is shown in Figure 2&3.



ɘ Male come for diagnosis

1 stands for **Female**
2 stands for **Male**

Figure 2: Gender of Respondents

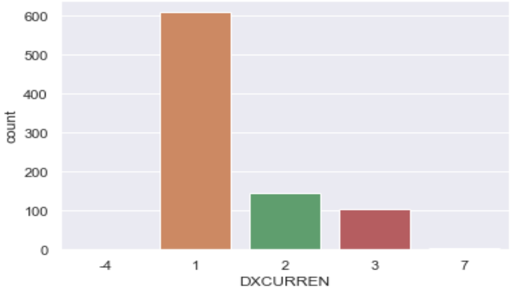


Figure 3: DXCURREN Data

-4 AND 7 are very rare events, hence they will be replaced with the most frequent"1'. I stand for healthy control (HC) which is the most frequently diagnosed result. 2 stands for mild cognitive impairment (MCI) of which the frequency is very low compared to HC while 3 stands for Alzheimer's disease (AD) which is least frequent diagnosis results. The least important values with the most frequent value, combine 2 and 3 to form Non-HC. This gives an updated plot after resolving the outliers.

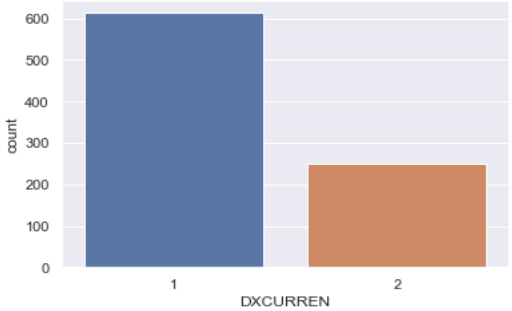


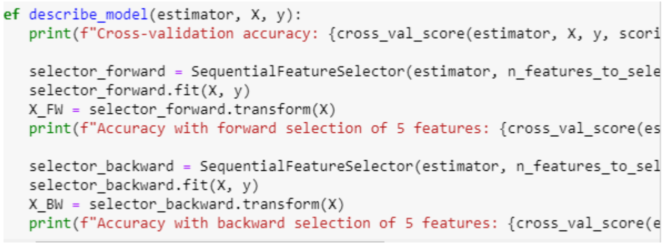
Figure 4: NEW DXCURREN

V. MODEL BUILDING

The data was split in train and test data set. The data was then balanced by over sampling the training data with the command:

Smote=SMOTE(random_state=42)
X_train, y_train= smote.fit_resample(X_train, y_train)
mms=MinMaxScaler()
X_train=mms.fit_transform(X_train)
X_test=mms.transform(X_test)

Writing a function



```
ef describe_model(estimator, X, y):
   print(f"Cross-validation accuracy: {cross_val_score(estimator, X, y, scori

   selector_forward = SequentialFeatureSelector(estimator, n_features_to_sele
   selector_forward.fit(X, y)
   X_FW = selector_forward.transform(X)
   print(f"Accuracy with forward selection of 5 features: {cross_val_score(es

   selector_backward = SequentialFeatureSelector(estimator, n_features_to_sel
   selector_backward.fit(X, y)
   X_BW = selector_backward.transform(X)
   print(f"Accuracy with backward selection of 5 features: {cross_val_score(e
```

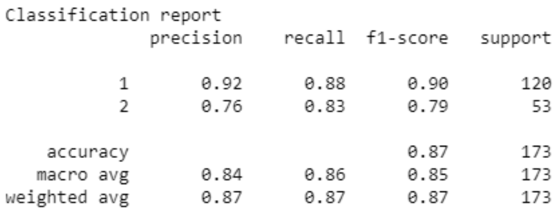
Figure 5: Writing Function Code

Modeling approach

1. Instantiate the classifier
2. Fit the model with the train set
3. Predict the model's performance
4. Plot the confusion matrix for the model

Modeling of data was done by Gaussian Naive Bayes. The ideas of Naive Bayes (NB) are based on Bayes' Theorem, which is a mathematical formula. The classification is based on the assumption that the prediction is independent of the predictors. Using the Naive Bayes as the baseline model, cross-validation accuracy is 0.899 with accuracy with the forward selection of 5 features as 0.946 while accuracy with the backward selection of 5 features as 0.948. using the fitting model. It shows the accuracy of training accuracy is 0.89 while the testing accuracy is 0.867. the modeling accuracy score is 0.867.

The classification report was also presented in Table 6.



```
Classification report
              precision    recall  f1-score   support

           1       0.92      0.88      0.90       120
           2       0.76      0.83      0.79        53

    accuracy                           0.87       173
   macro avg       0.84      0.86      0.85       173
weighted avg       0.87      0.87      0.87       173
```

Figure 6: Classification Report

Evaluating the model performance is predicted as 87.0%. it shows that the model performed well and hence hyper-parameter tuning is not necessary.

AIBL with Xgboost

The Xgboost was also used to predict the possible result of some clinical diagnoses. One of the most essential factors in the success of XGBoost has been its capacity to scale in all situations. The variable 'df2' is the data we will be using for further procedures. '-4' and '7' are very rare events, hence they will be replaced with the most frequent '1' 1 stands for healthy control (HC) which is the most frequent diagnosis result 2 stands for mild cognitive impairment (MCI) of which the frequency is very low compared to HC 3 stands for Alzheimer’s disease (AD) which is least frequent diagnosis result. 2 and 3 would be combined for the Non-HC class, this was the same that was done with Naïve Bayes.

Model Training with Xgboost algorithm

Extreme Gradient Boosting is a distributed gradient-boosted decision tree machine learning toolkit that is scalable. xgboost has both Regression and Classification algorithms but in this case, we are using the boosting algorithm which will predict three(3) classes. It does not also require feature scaling because it's comprised of Decision Trees, hence the data wasn't scaled. The accuracy of the Booster is 92%. The model performed well and hence hyper-parameter tuning is not necessarily. The plot is presented below.

AIBL with Decision Tree algorithm

The decision tree was also used to predict the likelihood of the diseases/illnesses one is likely to be affected with.
Decision Trees are a sort of Supervised Machine Learning in which the data is continually split according to a parameter (you explain what the input is and what the related output is in the training data). Two entities, namely decision nodes, and

leaves can be used to explain the tree. The decisions or consequences are represented by the leaves. Decision Tree has both Regression and Classification algorithm but in this case, we are using a classifier that will predict three(3) classes It does not also require feature scaling, hence the data wasn't scaled. The accuracy of the decision tree classifier was seen as 88%.

Hyparameter Tuning

This was done by fitting 5 folds for each of 10 candidates totaling 50 fits. The model was rebuilt with the best estimator with an accuracy of 88%.

AIBL with Random Forest

Data were split, feature importance in Random Forest Model was determined, and the important variables were also analyzed before building the model for the random forest with a training accuracy of 1.0 and a test accuracy of 0.93.
The plot Model for the supervised learning was presented in Appendix IV

Hierarchical, Clustering, and PCA

Data was replaced and cleaned for exploration, the data were also checked for correlation, and standardization of data was done to ensure all features have equal length before the clustering was done. The Plot is presented in Appendix I.

K-Means Clustering

We perform K-Means with nine clusters and iterations at a random state of 42wcss=[]. We then plot within the cluster, the sum of squares to determine the number of the cluster for the analysis. (See Appendix II). We created k-means with 2 clusters and created a copy of the original dataset for clustering and determined the mean value for clusters. We then determined the number of observations and proportions of observations for each cluster. We then determined age with cognitive assessment using the K-Means, age with CDR Scores, and Logical memory recall scale(SEE Appendix III).

Dimensionality Reduction Using PCA

PCA for dimensionality reduction and to create components subset, data standardization with PCA was done and this displays how each variable explains the variance. we determined the number of components and showed the PCA results using K-means clustering with PCA results. The Model summary and supervised learning were presented as an interrelated cluster and were created to identify relationships between feature data.

VI. DISCUSSION AND CONCLUSION

It was observed that each of the models tested, performed differently since each model was trained with the same dataset. Here is the summary of this project

- The dataset had 862 rows and 36 columns
- The correlation report showed that rid, siteid, viscode, examdate, aptestdt have similar information, so they will be dropped in the first copy of the dataset used for the naive Bayes modeling and subsequently in other supervised learning models
- The missing values report showed that there were no missing values in the dataset
- After dropping the columns with similar information, the shape of the updated dataset is now having 862 rows and 31 columns
- The dataset showed that more males went for diagnosis than female
- The dataset showed that the records had a large proportion of the people falling under the healthy control group (HC) some outliers [-4, 7] were replaced with 1.

VII. CONCLUSION

From the confusion matrix, naive Bayes predicted the true positive value well (HC), but predicted the non-HC value very poorly amongst the other models, random forest performed better than other models in predicting the true negatives and was 1 correct value away from tieing with naive Bayes., although, with more data and further finetuning, naive Bayes, decision tree and xgboost could improve on their performance.

VIII. SUGGESTIONS

1. For better analysis, more data is needed for both supervised and unsupervised learning. So going forward, there is a need to collate more dataset
2. The nature of data preprocessing and EDA can be improved upon when the first suggestion is taken care of
3. It was observed that unsupervised learning is not a suitable method for this kind of dataset, the nature of the dataset looks curated for supervised learning.

IX. REFERENCES

[1] Li H, Habes M, Wolk DA, Fan Y; A" Alzheimer's Disease Neuroimaging Initiative and the Australian Imaging Biomarkers and Lifestyle Study of Aging. A deep learning model for early prediction of Alzheimer's disease dementia based on hippocampal magnetic resonance imaging data. Alzheimer's Dement." 2019 Aug;15(8):1059-1070. doi: 10.1016/j.jalz.2019.02.007. Epub 2019 Jun 11. PMID: 31201098; PMCID: PMC6719787

[2] Komorowski, M., Marshall, D. C., Salciccioli, J. D.& Crutain, Y. " Exploratory Data Analysis", MIT Critical Data, Secondary Analysis of Electronic

Health Records, 2016, 185-205, DOI 10.1007/978-3-319-43742-2_15

[3] Kumar, S. &Ching, I., “Correlation Analysis to Identify the Effective Data in Machine Learning: Prediction of Depressive Disorder and Emotion States”, 2018, International Journal of Environmetal Research and Public Health.