Predictive Modeling for Alzheimer's Disease Risk Assessment

Aleksandr Dulepov
Aleksandr.dulepov@student.oulu.fi

Claudia Trujillo
claudia.trujilloolvera@student.oulu.fi

Fahim Himel
fahim.himel@student.oulu.fi

Zareen Tasnim
Zareen.Tasnim@student.oulu.fi

Abstract

Alzheimer's disease (AD) is an extremely challenging disease worldwide due to high prevalence and profound societal impact. This study explores the ability of predictive analytics using machine learning algorithms to predict the onset of the disease based on Magnetic Resonance Imaging (MRI) data, demographic characteristics and clinical measurements. Four distinct classifiers-random forest, logistic regression, k-nearest neighbors, and decision tree - were investigated in the study. Analyses established the decision tree classifier as the top performer, as it showcased superior metrics including an accuracy of 88.6%, a recall of 73.3%, and a ROC AUC of 86.7%. Logistic regression emerged as the second-best classifier (accuracy = 77.1%, recall = 86.7%, ROC AUC = 85.0%). On the other hand, random forest (accuracy = 54.3%, recall = 73.3%, ROC AUC = 76.3%) had a better ROC AUC value than k-nearest neighbors (accuracy = 71.4%, recall = 80%, ROC AUC = 75.7%), however, it did poorly in accuracy and recall. Overall, depending on the results, our study strongly suggests that decision tree classifier is the most reliable option for predictive modeling to assess the onset of Alzheimer's disease and could be used in the field of medicine for effective and early diagnosis of the diseases for patients.

1. Introduction

Alzheimer’s disease (AD) is the most common form of dementia and the most common neurodegenerative disease. Clinical manifestation includes a decline in short-term memory, personality change, cognition impairment that affects daily activities and finally having patient’s death as outcome. The greatest risk factor is age, the likelihood of having AD increases after 65 years from about 3% to 30% by age 85 [1]. Some other risk factors include genetics, inflammation, cerebral and cardiovascular diseases, diet and hypertension [1,2]. After symptom’s onset, life expectancy often goes from 3–10 years with complications such as immobility, pneumonia or blood clots [1].

The high detrimental impact in life-quality, increase in its prevalence, and absence of effective treatment to reverse or stop AD progression, highlights the need to develop models that can help to predict its manifestation and identify key factors that can work as protective or preventive factors and help medical decision-making.

The generation of huge volumes of medical and health data has opened the possibility to perform data-driven diagnostic or prognostic models for diseases, which include from classical statistical methods to machine learning models. Nowadays, machine learning methods are of interest as they have shown a superior performance compared to statistical models, when it comes to complex data [2].

In this project, we used four different machine learning (ML) based algorithms: “random forest, logistic regression, k-nearest neighbors and decision trees” to identify dementia based on medical data obtained from Magnetic Resonance Imaging (MRI) images, so as known risk factors and features, obtained from “MRI and Alzheimer's Magnetic: Resonance Imaging Comparisons of Demented and Nondemented Adults” data set [3].

2. Related Work

Previous studies have developed machine learning models for predicting Alzheimer’s disease (AD) and identify key factors for targeted prevention. One of them was described by Maoni et. al. [2], they included 3 different populations aged +60 years and used six machine learning models (logistic regression with penalty (LR), support vector machine (SVM), decision trees, random forest (RF), extreme gradient boosting (XGB), and artificial neural network (ANN)) in 3 different scenarios to test their ability in predicting AD the above three populations. Results showed that logistic regression performed best, with an AUC value of 0.818. Syaifullah et. al [4] developed a software named BAAD, which uses ML algorithms for the diagnosis of Alzheimer's disease (AD) and prediction of mild cognitive impairment (MCI) progression, by combining a support vector machine (SVM) to classify and a voxel-based morphometry (VBM) to reduce concerned variables. They compared the accuracy of the diagnosis of their software against two radiologists. BAAD's SVMs outperformed radiologists for AD diagnosis in an MRI review. The accuracy of the two radiologists was 57.5 and 70.0%, respectively, whereas that of the SVMst was 90.5%. Li et. al. [5] built three machine learning-based MRI data classifiers to predict AD and infer the brain regions that contribute to disease development and progression and compared them (constructed based on Support Vector Machine (SVM), 3D Very Deep Convolutional Network (VGGNet) and 3D Deep Residual Network-ResNet). The classification accuracy for AD subjects from elderly control subjects was 90%, 95%, and 95% for the SVM, VGGNet and ResNet classifiers, respectively. The resulted maps consistently highlighted several disease-associated brain regions. A diversity of further studies like the ones cited show the high potential of machine learning algorithms as support in clinical practice in AD diagnosis and prediction.

3. Proposed Method

In our dataset we have labeled data; as we know the cases that present the symptoms (dementia) therefore, we should use supervised learning methods. According to the objective of the project and the data’s nature (binary outcome), we decided to use a classification method.

A brief description of models used is presented below:

3.1 Random Forest

Random forest is a supervised learning algorithm used for both classification and regression tasks. It works by creating a number of decision trees during the training phase. Each tree is constructed using a random subset of the dataset to measure a random subset of features in each partition. In prediction, the algorithm aggregates the results of all trees, either by voting (for classification tasks) or by averaging (for regression tasks). The advantages of this model include reduced overfitting compared to a single decision tree; handles missing values and maintains accuracy and it is effective for large datasets with high dimensionality. However, disadvantages include less interpretable compared to a single decision tree; requires more computational resources and time for training and might not perform well on very imbalanced datasets [6].

3.2 Logistic regression

Logistic regression is a supervised learning algorithm used for binary classification problems. It models the probability that an element belongs to a given class or not. Logistic regression is a statistical algorithm that analyzes the relationship between two data factors. It uses a logistic function (also known as a sigmoid function) to map predicted values to probabilities, which lie between 0 and 1. The advantages of this model are: efficient for small datasets with limited features; provides probabilities for outcomes; easy to implement and interpret and performs well when the relationship between features and target variable is linear. On the other hand, disadvantages include it assumes a linear relationship between the features and the target; it is limited to linear decision boundaries, and it is not suitable for complex relationships between features [7].

3.3 k-nearest neighbors

The k-nearest neighbors (KNN) algorithm is a non-parametric, supervised learning classifier, which uses proximity to make classifications or predictions about the grouping of an individual data point. It works by finding the K nearest neighbors to a given data point based on a distance metric, such as Euclidean distance. The class or value of the data point is then determined by the majority vote or average of the K neighbors. The advantages of using this method include easy implementation, easy adaptability; few hyperparameters; its non-parametric nature makes it effective for complex decision boundaries and it is robust to noisy training data. On the other hand, disadvantages include its computationally expensive during prediction, especially with large datasets; it does not perform well with high-dimensional data inputs; prone to overfitting: sensible to irrelevant features and the scale of the data and need to choose an appropriate value for k, which can affect the model's performance [8].

3.4 Decision trees

A decision tree is a supervised learning technique that can be used for both classification and regression problems. It builds a flowchart-like tree structure where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label. It is constructed by recursively splitting the training data into subsets based on the values of the attributes until a stopping criterion is met. Advantages of this method include easy to understand and interpret; it can handle both numerical and categorical data; it is robust to outliers [9].

4. Experiments

The supervised classification approach has been applied because of several reasons:

1) The presence of prelabeled data.
2) The outcome is binary type of data (demented or nondemented).
3) In clinical settings, the price of false negative error is very high, and we need very high transparency and interpretability of the results for clinical doctors.

The whole dataset has been divided into 3 sets:

1) Training set 50%.
2) Validation set 25%.
3) Test set 25%.

These proportions were chosen to provide enough data for validation and test tasks, which increases robustness of the tested models.

All the features have been scaled to [0-1] range by applying the MinMaxScaler function from the sklearn library. To prevent the train-test contamination (data leakage) feature scaling for teaching and validation data has been applied separately from the test data.

The classifiers have been used and compared: random forest, decision tree, logistic regression, and k-nearest neighbor classifier.

The performance measure for the validation was accuracy. The hyperparameters that have been validated were:

1) The number of neighbors (k-value) for the k-nearest neighbor classifier.
2) The C-value for logistic regression. It regulates the intensity of the regularization.
3) The maximum depth of the decision tree classifier.
4) The number of trees in the random forest classifier.

The dataset is balanced, with a roughly equal quantity of samples labeled as '1' and '0' (183 demented and 190 nondemented). In medical diagnostics, our priority is achieving a high true positive rate to identify all patients with Alzheimer's disease at the earliest opportunity. Simultaneously, we aim to minimize the false positive rate to avoid misdiagnosing healthy adults as having dementia and initiating unnecessary medical interventions. Therefore, the ROC AUC (Receiver Operating Characteristic Area Under the Curve) serves as a key performance metric for evaluating model performance in this context.

As additional performance metrics confusion matrix, accuracy and recall/sensitivity have been chosen. They are also commonly used with balanced data and provide robust estimation of the models' performance.

4.1. Dataset and pre-processing

The dataset has been collected from the Open Access Series of Imaging Studies (OASIS) project. The original dataset consists of 15 features and 373 observations that include 1 to 4 visits. The outcome is a categorical variable of 3 classes: Demented, Nondemented, and Converted (initially were classified as nondemented but at the second or third visit were classified as Demented). To simplify the analysis the class "Converted" was changed to "Demented"; thus, we are dealing with a binary outcome.

Data was further preprocessed by eliminating unnecessary features, such as more than 2 visits, MRI ID (not relevant for the analysis), and hand (all participants are right-handed).

Likewise, sex was modified to binary notation (0,1), so as demented and non-demented and missing values were identified and non-considered (8). After these modifications, the final dataset included 142 observations.

The model was constructed using 8 features: sex, age, educational level, socioeconomic status, Mini Mental State Examination, Estimated Total Intracranial Volume, Normalize Whole Brain Volume and Atlas Scaling Factor ('M/F', 'Age', 'EDUC', 'SES', 'MMSE', 'eTIV', 'nWBV', 'ASF').

4.2. Software

The visualization and the preprocessing for the visuals have been done in R 4.3.2. The libraries used for this in R are as follows: readxl, dplyr, gplots, ggplot2.

The preprocessing, teaching models and assessing performance have been done in Python 3.11.5 version. The libraries that have been used are pandas, matplotlib, numpy, sklearn, scipy, skimage, imblearn.

4.3. Exploratory Data Analysis and Visualization

In the exploratory data analysis, we attempted to understand the relationships and patterns between the selected features and the outcome. The socioeconomic status structure was almost similar between the “demented” and the “nondemented” groups, with an exception in the 2nd level. In case of sex, there was noteworthy difference between males and females for the “nondemented” group, however, in the “demented” group there was no notable difference (Figure 1).

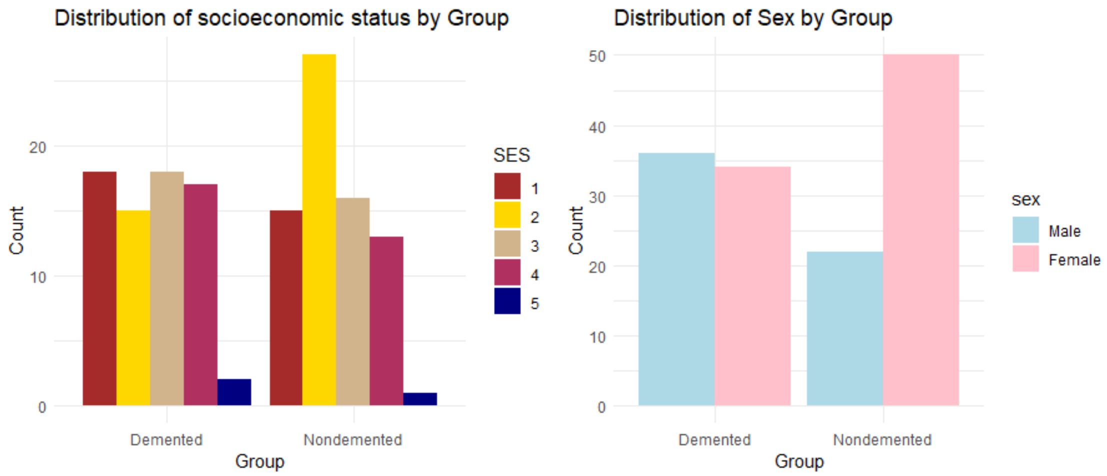

Figure 1: Distribution of socioeconomic status and sex by group

The boxplots suggest that the median value of ASF is negligibly higher in the “demented” group compared to the “nondemented” group. Whereas eTIV does not show any visible difference between the two groups (Figure 2).

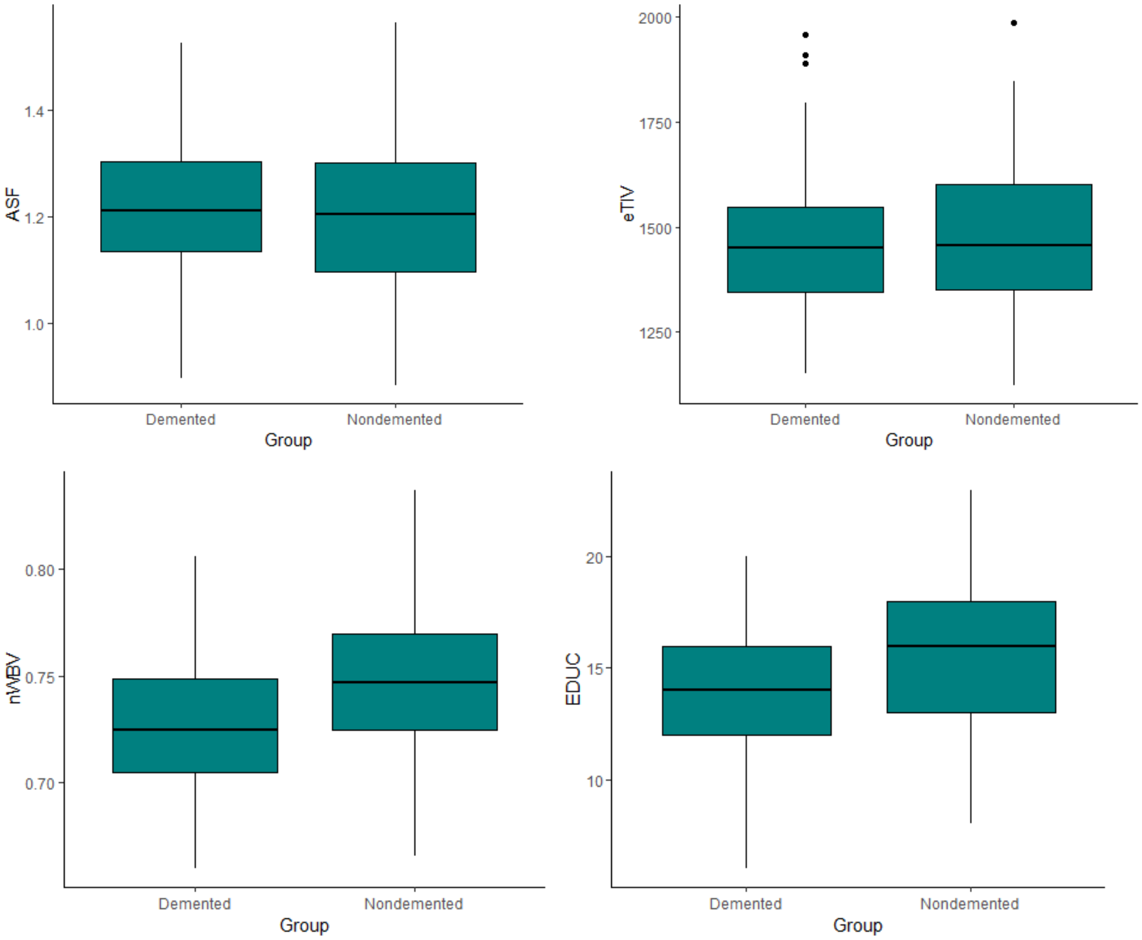

Figure 2: boxplots for ASF, eTIV, nWBV, and education level by group

On the other hand, the nWBV is significantly lower among the demented participants (Figure 2). Similar to nWBV, the "demented" group exhibits a lower level of education compared to nondemented participants.

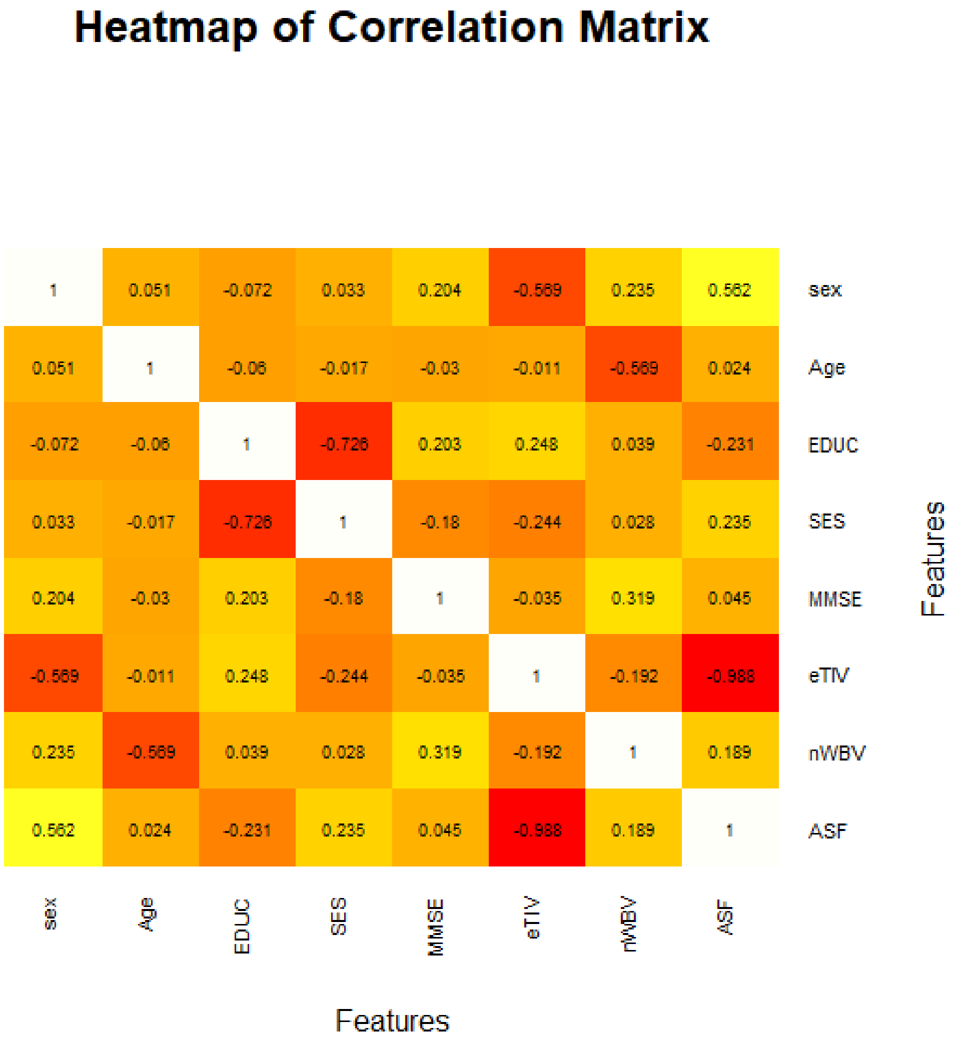

Figure 3: Heatmap of correlation matrix of the selected features

To visualize the patterns of the pairwise correlation among the selected features, we generated a heatmap. The heatmap suggests a high correlation between two background characteristics of the participants: socioeconomic status and years of education. This could be due to the influence of educational attainment on income, which increases the socioeconomic status. Alongside, the strong negative correlation with a correlation coefficient of -0.988 between Estimated Total Intracranial Volume (eTIV) and Atlas Scaling Factor (ASF) implies that larger intracranial volumes correspond to lower scaling factors, which indicates to an adjustment for variations in head size when normalizing brain volume measurements (Figure 3).

5. Results and Discussion

After splitting the dataset into training, validation and test sets the distribution of the samples has been assessed. Teaching data "class 0 (Nondemented)" has 33 samples and in "class 1 (Demented)" there are 38 samples Test data "class 0" has 20 samples and in "class 1" there are 15 samples. Thus, we can conclude that the data has been distributed roughly equally between the data sets. The graph for sample's distribution is presented below (Figure 4).

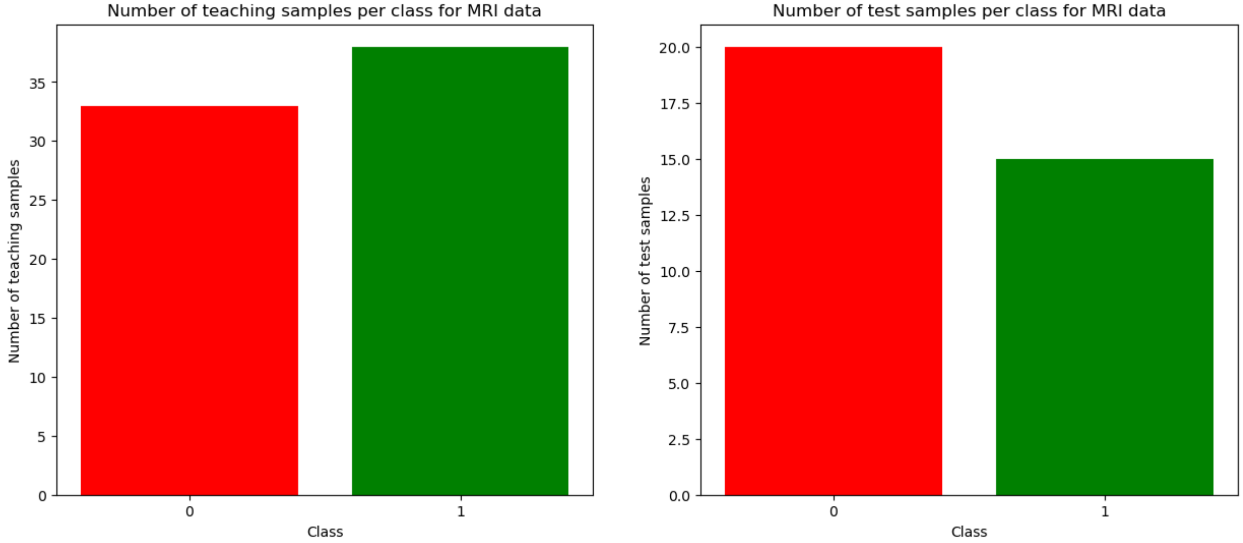

Figure 4. Number of samples in training and test set.

The validation has been applied for all tested models on the validation set. It will be described in the following sections.

Validation of the k-nearest neighbors classifier

Validation has been performed by running k-values through the loop from 1 to 33 with a step of 1. Then the effect of the k-value on the classification accuracy was plotted on the graph, and the best k-value was selected from it. The best k-value with the highest accuracy is 9. The graph for performance of different k-values is presented below (Figure 5).

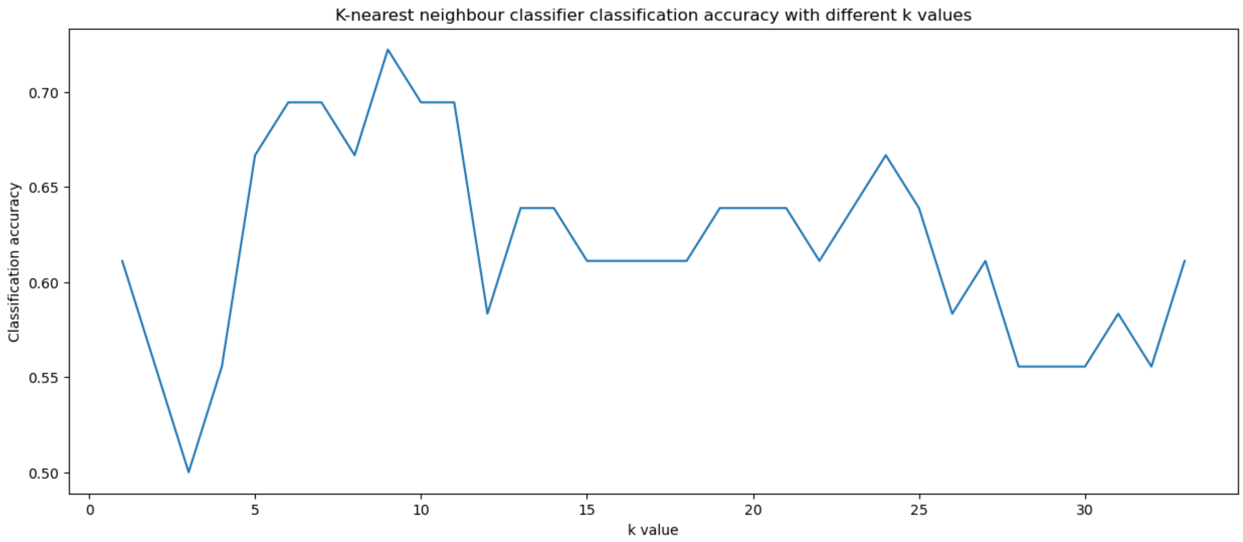

Figure 5. Performance of different k-values.

Validation of the logistic regression classifier

Validation has been performed by running C-values as exponent of 10^n, where n is –5, -4, ..., 14, 15 with a step of 1. Then the effect of the exponent value of 10 (C-value) on the classification accuracy was plotted on the graph, and the best C-value was selected from it. The best k-value with the highest accuracy is 2 (10^2). The graph for performance of different C-values is presented below (Figure 6).

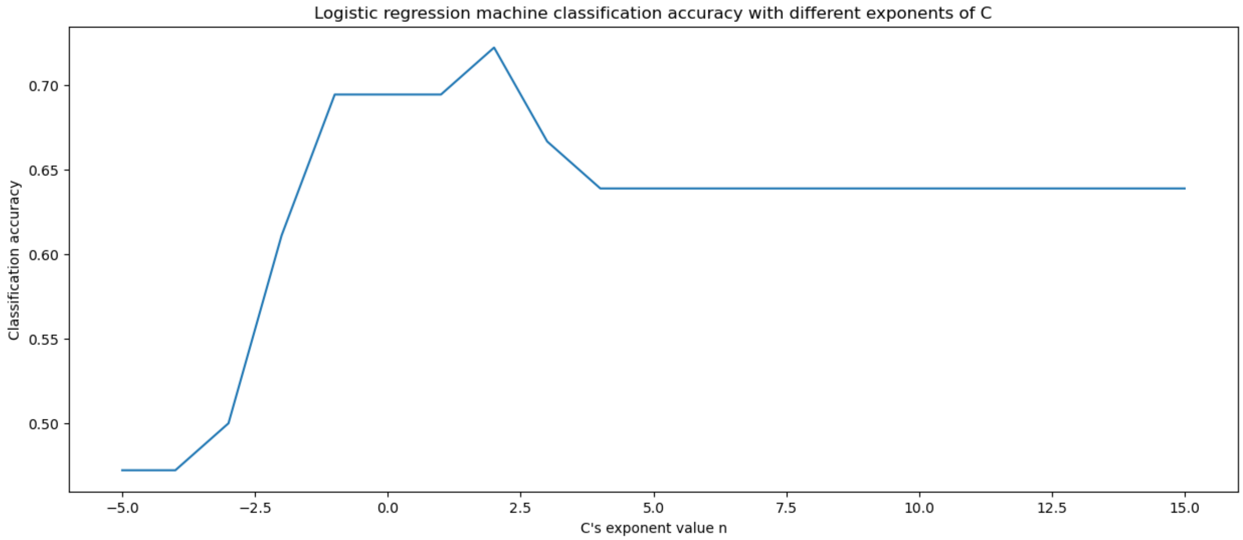

Figure 6. Performance of different C-values.

Validation of the decision tree classifier

Validation has been performed by running maximum depth value through the loop from 1 to 9 with a step of 1. Then the effect of the max depth value on the classification accuracy was plotted on the graph, and the best max depth value was selected from it. The max depth value with the highest accuracy is 1. The graph for performance of different max depth values is presented below (Figure 7).

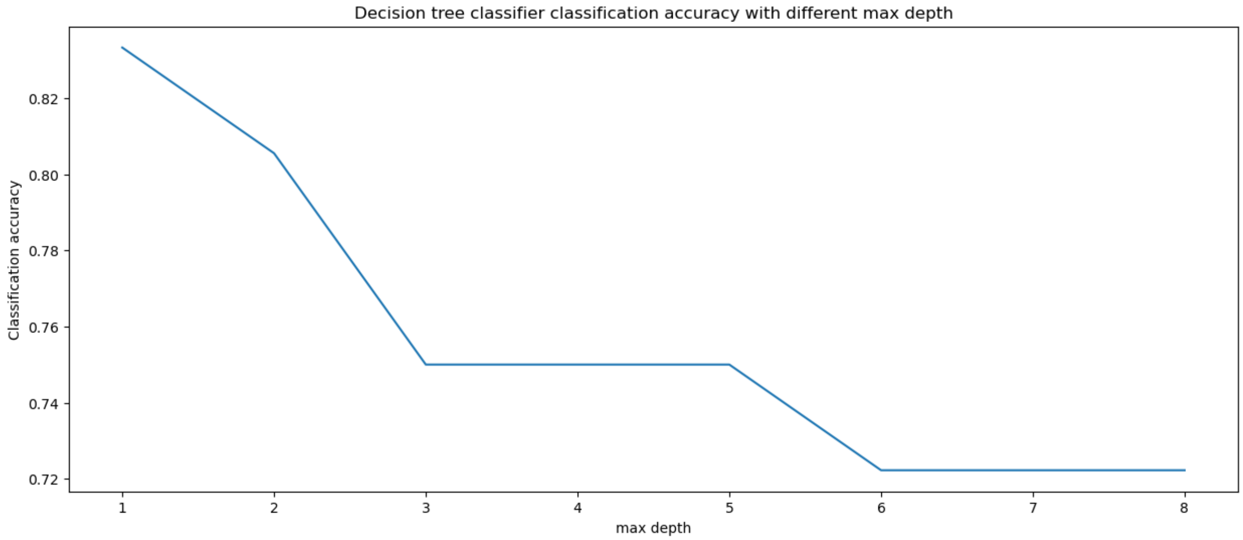

Figure 7. Performance of different max depth values.

Validation of the random forest classifier

Validation has been performed by running the number of trees (number of estimators) through the loop from 1 to 400 with a step of 10. Then the effect of the number of trees on the classification accuracy was plotted on the graph, and the best number of trees value was selected from it. The best number of trees with the highest accuracy is 91. The graph for performance of different numbers of trees is presented below (Figure 8).

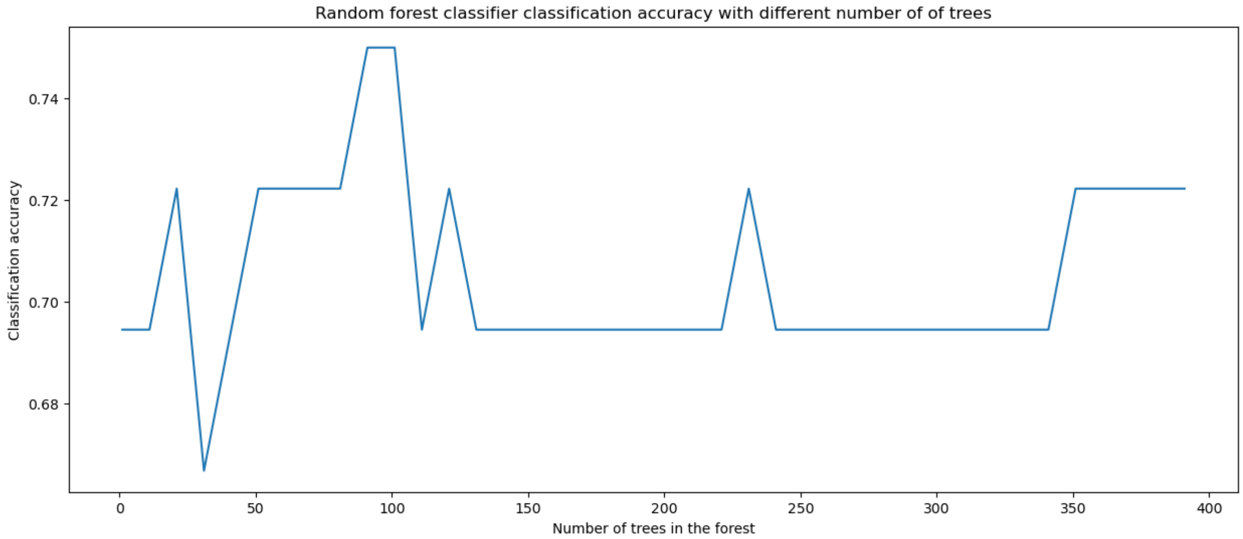

Figure 8. Performance of different number of trees in the forest

The table with optimal hyperparameters is presented below (Table 1). The models were retaught on the training set with the evaluated hyperparameters.

Classifier	**Best hyperparameter value**

K-nearest neighbors classifier	9
logistic regression	2
Decision tree	1
Random forest	91

Table 1. The best hyperparameter values.

After the validation part, the performance part was performed. The confusion matrixes for each model are shown below (Figure 9). According to these results, the decision tree classifier distinguished groups the most effectively because it has the most correct number of predictions (only 4 samples were false negatives and 0 false positives).

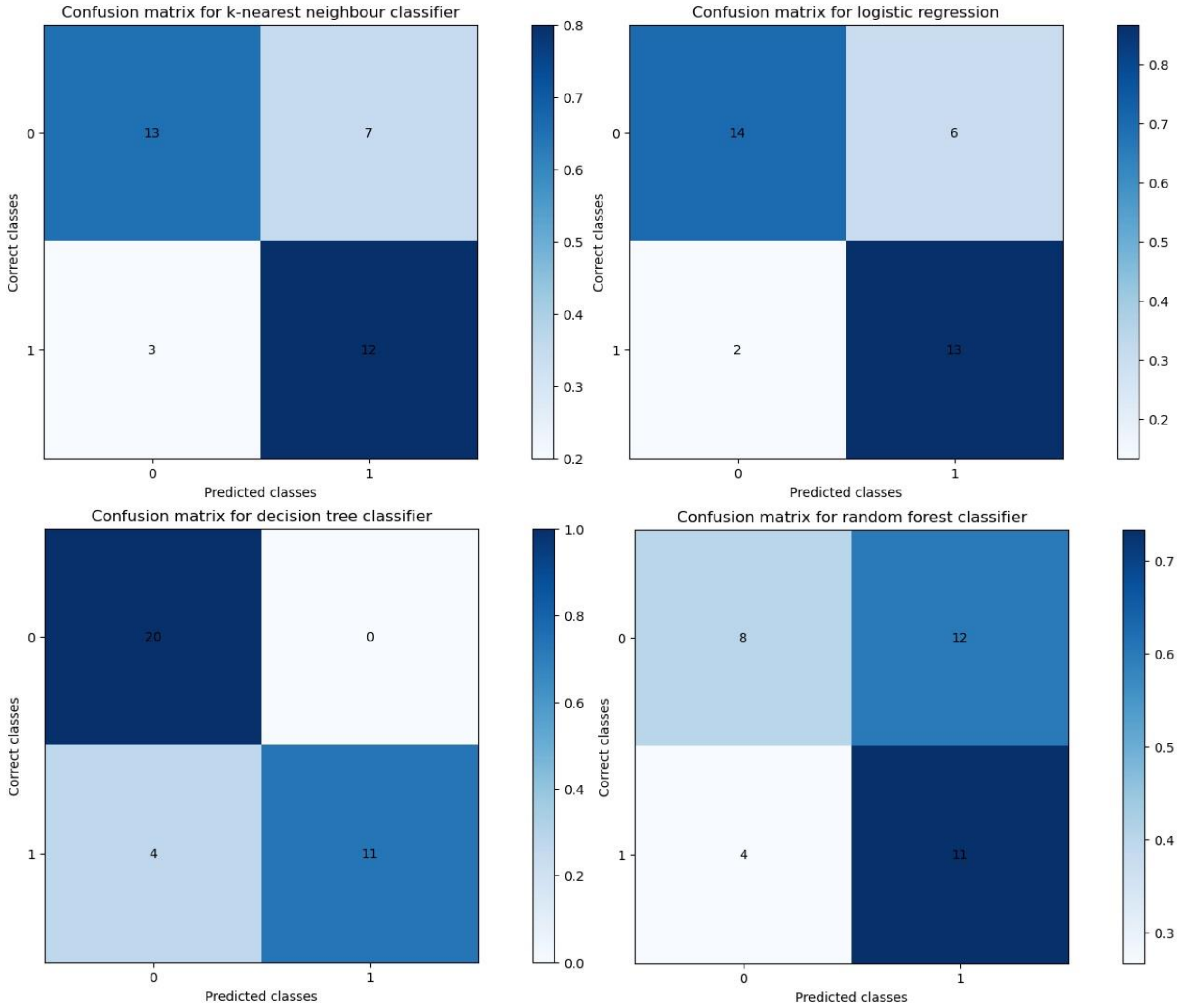

Figure 9. Confusion matrixes.

The ROC curves have also been calculated for all models (Figure 10). Based on these results the decision tree classifier performed the best (ROC AUC value is 0.867). The random forest, logistic regression, k-nearest neighbor had ROC AUC 0.763, 0.85, 0.757 respectively.

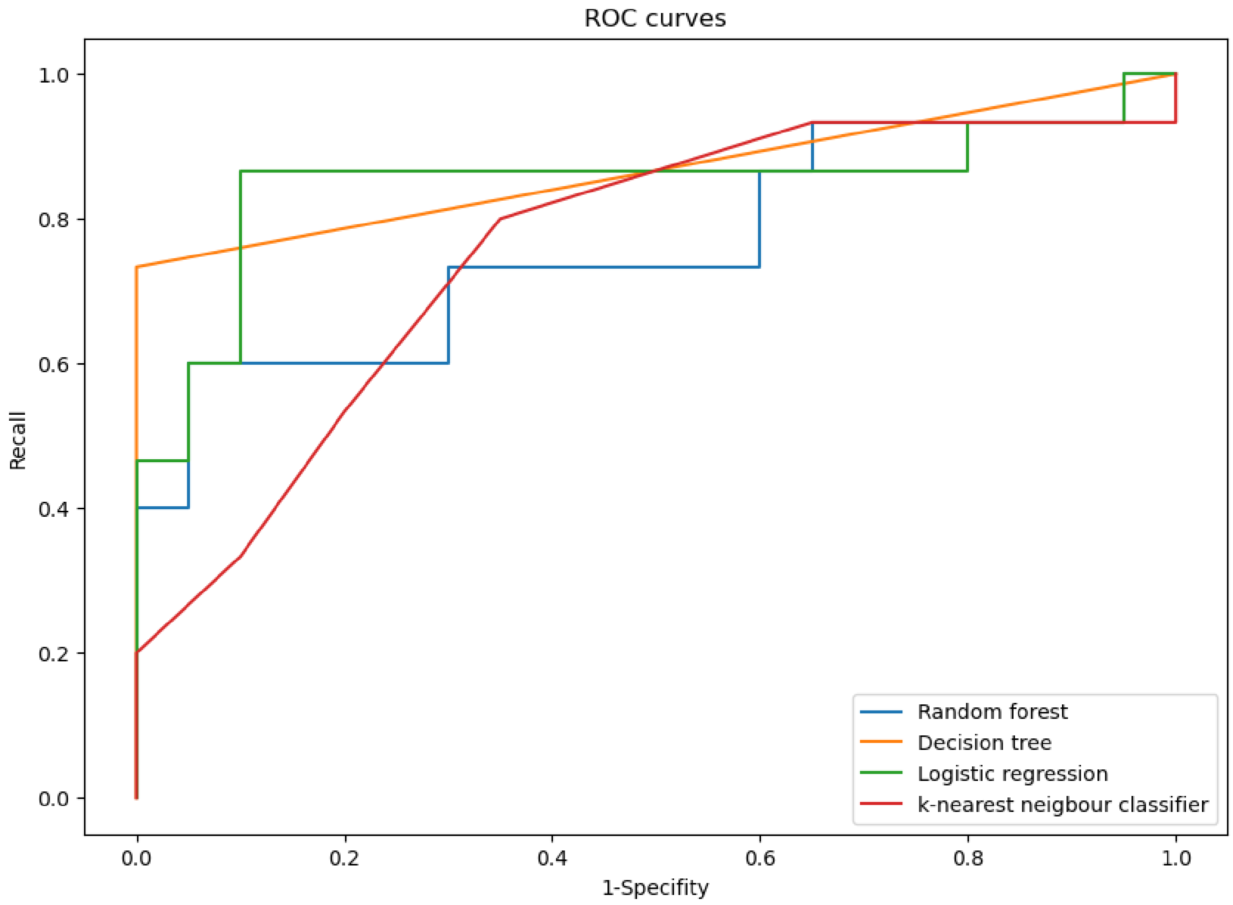

Figure 10. ROC curves.

Accuracy and Recall/Sensitivity were calculated and presented in the final table of results (Table 2).

Model	Accuracy	Recall/ Sensitivity	ROC AUC
Random forest	0.543	0.733	0.763
Decision tree	0.886	0.733	0.867
Logistic regression	0.771	0.867	0.85
k-nearest Neighbors classifier	0.714	0.8	0.757

Table 2. Performance metrics for different models.

According to the final results, the best performed classifier was the decision tree (accuracy = 88.6%, recall = 73.3%, ROC AUC = 86.7%). The second-best model was the logistic regression (accuracy = 77.1%, recall = 86.7%, ROC AUC = 85.0%). The third-best performed model based on ROC AUC as the primary metric was the random forest classifier, however it performed worse than k-nearest neighbor classifier in recall and accuracy measures.

6. Conclusions

To conclude, this study aimed to predict the Alzheimer disease risk by using the machine learning algorithms trained on MRI data and known risk factors. Four classifiers such as random forest, logistic regression, k-nearest neighbors and decision trees were compared based on their metrics. The decision tree classifier and logistic regression demonstrated the strongest performance. The robust performance of decision trees, especially in minimizing false negatives, highlights its potential for clinical use in early AD diagnosis. However, although the random forest classifier achieved a better ROC area under curve than k-nearest neighbors; however, lack of accuracy made it a less effective method. Decision tree and logistic regression models offer promising avenue for early diagnosis of AD and risk management. Further research on larger data sets is needed to enhance the generalizability and reliability of these predictive models in clinical practice.

7. Acknowledgements

Some of the code lines for data preprocessing in Python were borrowed from Hyunseok Choi at kaggle website (https://www.kaggle.com/code/hyunseokc/detecting-early-alzheimer-s)

The MRI longitudinal data for this project was introduced by Open Access Series of Imaging Studies (OASIS) - a project aimed at making MRI data sets of the brain freely available to the scientific community.

8. Contributions

Aleksandr Dulepov – responsible for choosing validation methods, performance metrics, and machine learning models; developing code lines in Python for preprocessing, training, validating, and testing machine learning models; methods (first half), results, and discussion sections of the report.

Claudia Trujillo – introduction, related work, methods description, dataset and pre-processing section.

Zareen Tasnim – responsible for developing codes in R version 4.3.2 for preprocessing the data and producing visualizations, writing the abstract, exploratory data analysis and visualization section of the report.

Fahim Himel – contributed to the code development of data visualization and writing the conclusion section of the report, assisted in the developing code for teaching machine learning models.

References

[1] - Sheppard Olivia; Coleman Michael. (2020). Alzheimer’s Disease: Etiology, Neuropathology and Pathogenesis. In X. Huang (Ed.), Alzheimer’s disease: Drug discovery. (pp. 1–22). Exon Publications.

[2] - Maoni Jia, Yafei Wu, Chaoyi Xiang, Ya Fang; Predicting Alzheimer’s Disease with Interpretable Machine Learning. Dement Geriatr Cogn Disord 2 October 2023; 52 (4): 249–257. https://doi.org/10.1159/000531819

[3] - Magnetic Resonance Imaging Comparisons of Demented and Nondemented Adult. https://www.kaggle.com/datasets/jboysen/mri-and-alzheimers/data .

[4] - Syaifullah, A. H., Shiino, A., Kitahara, H., Ito, R., Ishida, M., & Tanigaki, K. (2021). Machine Learning for Diagnosis of AD and Prediction of MCI Progression From Brain MRI Using Brain Anatomical Analysis Using Diffeomorphic Deformation. Frontiers in neurology, 11, 576029. https://doi.org/10.3389/fneur.2020.576029

[5] - Li, Q., & Yang, M. Q. (2021). Comparison of machine learning approaches for enhancing Alzheimer's disease classification. PeerJ, 9, e10549. https://doi.org/10.7717/peerj.10549

[6] - IBM. (2024) What is a Random Forest? https://www.ibm.com/topics/random-forest#:~:text=Random%20forest%20is%20a%20commonly,both%20classification%20and%20regression%20problems.

[7] - IBM. (2024) What is Logistic regression? https://www.ibm.com/topics/logistic-regression

[8] - IBM. (2024) What is the KNN algorithm? https://www.ibm.com/topics/knn#:~:text=The%20k%2Dnearest%20neighbors%20(KNN,of%20an%20individual%20data%20point.

[9] - IBM. (2024) What is a Decision Tree? https://www.ibm.com/topics/decision-trees#:~:text=A%20decision%20tree%20is%20a,internal%20nodes%20and%20leaf%20nodes.

Appendix

The source code and the dataset of the project:

https://github.com/Arnaden/Ml_in_medicine