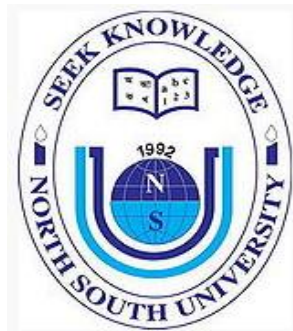


Project Report
CSE445 – Machine Learning

**Dementia Prediction Using Different Machine Learning
Model**



Submitted By

ID: 1621148642 – Name: Md. Tahrir Faroque Tushar

ID: 1721395042 – Name: S. M. Al Faruqui

ID: 1620345042 – Name: Morshedul Bari Antor

ID: 1620346042 – Name: Shafayet Jamil Zim

Submitted To

Syed Athar Bin Amir – SAA3

Lecturer

ELECTRICAL AND COMPUTER ENGINEERING
NORTH SOUTH UNIVERSITY

Fall 2020

Abstract

This paper represents the result and analysis regarding detecting dementia. This project's main goal was to recognize dementia among various patients. For this, we have used a data set from OASIS. Though the data set is small, it has some significant values. We analyzed the data set and applied several models: Support Vector Machine, Logistic Regression, Random Forest, and Decision tree. First, we run them without fine-tuning, then with the fine-tuning. We compared all the results and found that SVM provides better results than other models. It has the best accuracy in detecting dementia among numerous patients.

Table of Contents

Abstract.....	1
Introduction.....	3
Background.....	3
Design Methodology.....	4
Data Preprocessing & Feature Engineering	5
Evaluation Metrics	5
Training and Fine-Tuning	6
Support Vector Machine	6
Random Forest	6
Logistic Regression.....	6
Decision Tree	6
Results & Discussion	7
Conclusion	9
Authorship	10
Source Code.....	10
Reference	10

Introduction

Alzheimer's syndrome is an inherited, irreversible brain condition that steadily affects the ability to perform the necessary things, memory and reasoning skills, and ultimately. A massive proportion of neurons stop working in Alzheimer's disease, losing synaptic connections. Between a person's 30s and mid-60s, early-onset Alzheimer's happens and is very rare. Symptoms can include a shift in sleep habits, depression, anxiety, difficulties doing basic tasks such as reading or writing, aggressive actions, and poor decision making also happened in Alzheimer's disease. Alzheimer's disease and initial changes in the brain begin 10-20 years before the onset of symptoms. It progressively leads to memory damages and decreases thinking abilities. The leading cause of this disease is dementia.

Dementia is the failure of brain function, understanding, recognizing, thinking, and behavioral skills to such a level that an individual's everyday life and behaviors are interfered with. Few people with dementia are unable to deal with their emotions, and their personality can be changed. From the mildest stage, dementia varies in severity. It mainly affects older people. No cure is available rather than treatment.

Studies show that if we can detect early Alzheimer's disease, it may help improving therapy. For this, they have to predict the progress of the disease accurately from mild condition to dementia. Machine learning technology can help to predict accurately early Alzheimer's disease. Therefore, we developed a model that can indicate early Alzheimer's disease, using machine learning to support medical technicians. It will verify and show the result if anyone has Alzheimer's disease or not.

Background

Machine Learning is defined as the study of computer programs that leverage algorithms and statistical models to learn through inference and patterns without being explicitly programmed.

We have used the SVM model. A support vector machine (SVM) is a supervised machine learning model that uses classification algorithms for two-group classification problems. Support Vector Machines is a fast and dependable classification algorithm that performs very well with a limited amount of data to analyze.

We also used the Logistic Regression model. The logistic regression model is the appropriate regression analysis. Logistic regression is predictive regression analysis. To classify data and illustrate the relationship between one dependent binary variable and one or more independent nominal, ordinal, interval, or ratio-level variables, logistic regression is used.

Another model we used is the Decision tree. A decision tree is a machine learning algorithm that partitions the data into subsets. A decision tree's purpose is to sum up, the training data in the smallest tree possible.

The last model we applied is random Forest. Random forest is a supervised learning algorithm. Random forest is a versatile, easy-to-use machine learning algorithm that provides, most of the time, a fantastic result even without hyper-parameter tuning. Due to its simple design and variety, it is also one of the most used algorithms.

Design Methodology

This project's primary goal is to predict dementia in different patients based on various attributes such as Mini-Mental State Examination (MMSE), Socioeconomic Status (SSE), Estimated Total Intracranial Volume (eTIV), Normalize Whole Brain Volume (nWBV), etc. For this project, we have used longitudinal MRI data from OASIS. We have used four machine learning models, Support Vector Machine (SVM), Logistic Regression, Random Forest, and Decision Tree for the classification.

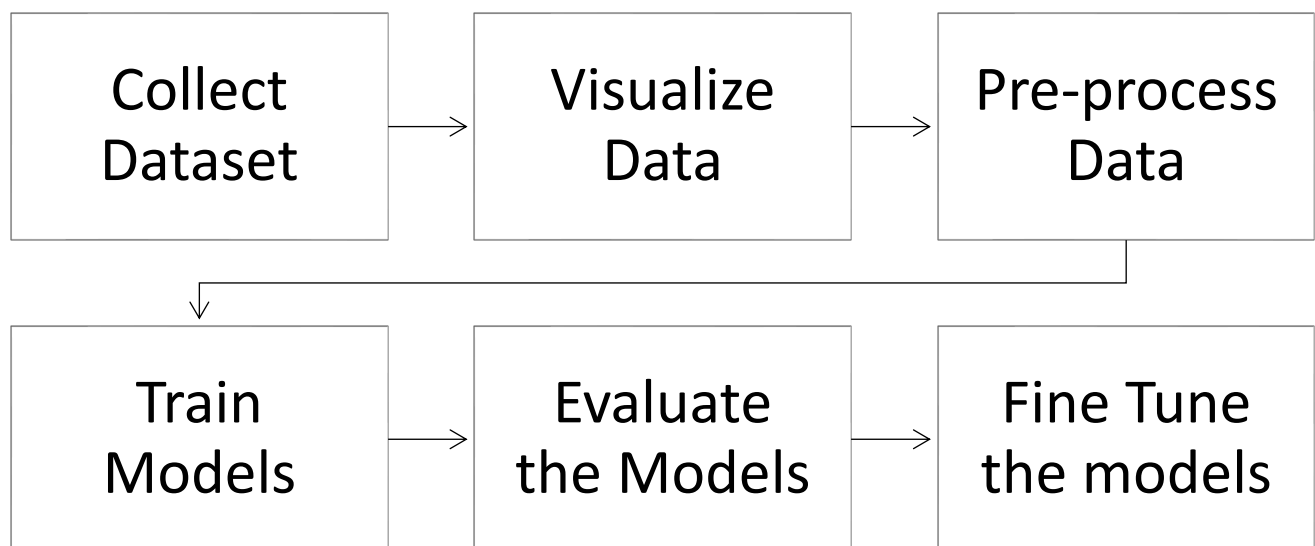


Figure: Development Stages

Data Preprocessing & Feature Engineering

At first, we analyzed the dataset for any categorical values, and there are several of them. Among them, Gender and Group attribute columns are converted into numeric values 0 and 1. Then, we checked the correlation between attributes using the 'correlation_matrix' function based on Group attributes and plotted them to understand better. We found that Gender, SES, and ASF showed a closer correlation with the Group attribute. After that, the dataset is checked for any null or missing values. SES and MMSE columns have 19 and 2 missing values, respectively. As mentioned earlier, the SES feature has a close correlation with the target attribute. For that, we did not delete those rows with missing values. Instead, the median value is used to fill those missing values for both of the features.

Next, we assigned the features to make the prediction and the target value that the model will predict. Then, we split the dataset for training-validation and testing. We considered using random sampling for the split, but this creates an imbalance between train and test split. So, we applied stratified sampling with a train-validation size of 80% and a test size of 20%. After that, standardization has been applied to do the scaling of the features. Furthermore, we have done some histograms and scatterplot visualization on the training split to understand the scenario better. Then we began our training for this project. We have implemented all of the models using the scikit-learn library.

Evaluation Metrics

We used 4 evaluation metrics Accuracy, Recall, Area Under the Curve(AUC), and Confusion Matrix, to evaluate our model. Here, accuracy is not a good measurement for a classification task. So, we mainly focused on Recall, AUC, and the Confusion Matrix.

Training and Fine-Tuning

Support Vector Machine

At first, we implemented the Support Vector Machine without any fine-tuning. Without fine-tuning, SVM takes regularization parameter C as 1, and for the kernel, it uses the Radial basis function (RBF). We also calculated the confusion matrix based on this version. After that, we applied the grid search to fine-tune the model. As for the parameter combinations, we took different regularization parameter values C , gamma values, and four types of kernels, such as the rbf, linear, poly, sigmoid kernel. Also, 5-fold cross-validation was applied to evaluate all possible combinations. Then, we trained the model again, and there was a significant improvement.

Random Forest

After that, we applied the Random Forest model without fine-tuning. And then, Just like the SVM, grid search was used here with 5-fold cross-validation and different parameter combinations such as the #trees in the random forest ($n_estimators$), what function to use for the #features to consider at every split, levels in the tree, and method of selecting samples for training each tree. To measure the quality of the tree, we used the Gini criterion. We tried with the Entropy criterion, but Gini provides better accuracy.

Logistic Regression

Also, we applied the same approach with the Logistic Regression model, just like the SVM and Random Forest. Here, the only difference is that we used the l_2 penalty and different regularization parameter values C for the fine-tuning.

Decision Tree

In the Decision Tree model, the same approach is followed. We train the model without fine-tuning and then used the grid search to find the best parameter values to fine-tune the model. Here, we consider the Gini criterion as a fixed value to evaluate the tree's quality and choose a range of 1 to 10 for evaluating the depth of the tree.

Finally, to show the comparison between these four models, we plotted the ROC curve based on the False Positive Rate vs. True Positive Rate.

Results & Discussion

Since, this is a classification problem the evaluation metrics used in our project were: Accuracy, Recall, Area Under the Curve (AUC) and confusion matrix. After compiling the results from all the models, it is evident that Support Vector Machine (SVM), more specifically Support Vector Classifier gave the best overall result in all metrics. Though Decision Tree classifier gave the best True Positive Result. We have used Grid Search for fine tuning our models. So, the results obtained are the best possible result for our particular dataset. We noticed some overfitting with Decision Tree and Random Forest models.

Our dataset has a dimension of 373 rows x 15 columns, which is relatively small in the field of Machine Learning. Theoretically, SVM works best with smaller datasets.

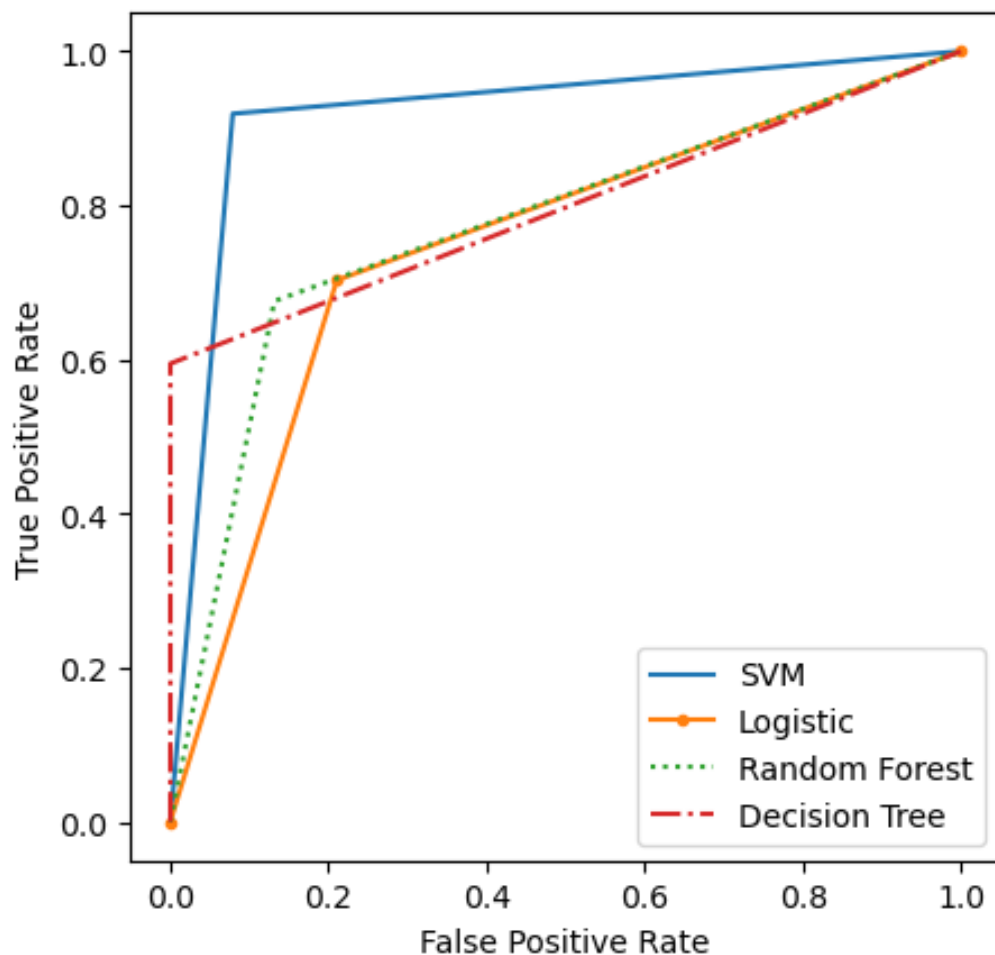


Figure: ROC curve for finding AUC of 4 different models

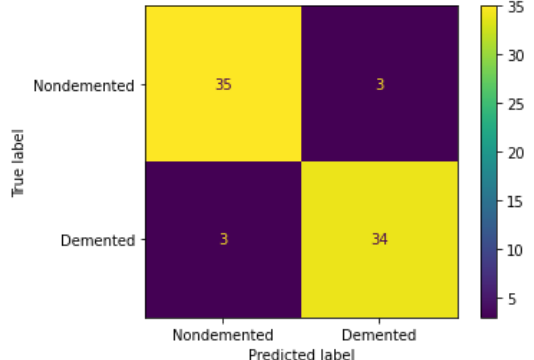
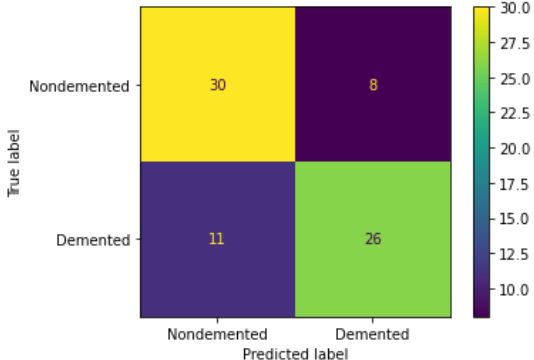
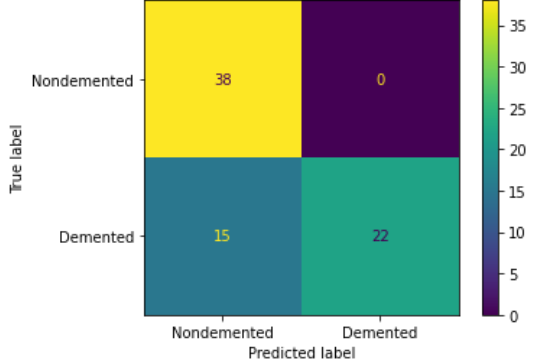
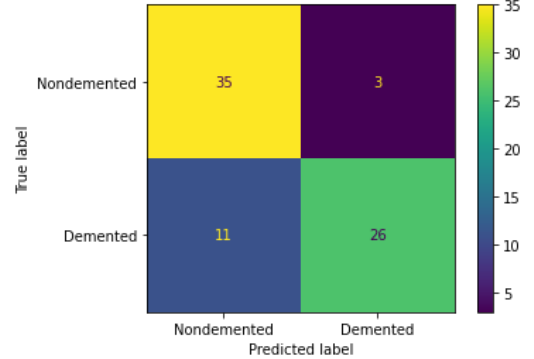
Model	Accuracy	Recall	AUC	Confusion Matrix									
Support Vector Machine	0.92	0.919	0.919	 <table border="1"><thead><tr><th>True label \ Predicted label</th><th>Nondemented</th><th>Demented</th></tr></thead><tbody><tr><th>Nondemented</th><td>35</td><td>3</td></tr><tr><th>Demented</th><td>3</td><td>34</td></tr></tbody></table>	True label \ Predicted label	Nondemented	Demented	Nondemented	35	3	Demented	3	34
True label \ Predicted label	Nondemented	Demented											
Nondemented	35	3											
Demented	3	34											
Logistic Regression	0.747	0.703	0.746	 <table border="1"><thead><tr><th>True label \ Predicted label</th><th>Nondemented</th><th>Demented</th></tr></thead><tbody><tr><th>Nondemented</th><td>30</td><td>8</td></tr><tr><th>Demented</th><td>11</td><td>26</td></tr></tbody></table>	True label \ Predicted label	Nondemented	Demented	Nondemented	30	8	Demented	11	26
True label \ Predicted label	Nondemented	Demented											
Nondemented	30	8											
Demented	11	26											
Decision Tree	0.800	0.594	0.797	 <table border="1"><thead><tr><th>True label \ Predicted label</th><th>Nondemented</th><th>Demented</th></tr></thead><tbody><tr><th>Nondemented</th><td>38</td><td>0</td></tr><tr><th>Demented</th><td>15</td><td>22</td></tr></tbody></table>	True label \ Predicted label	Nondemented	Demented	Nondemented	38	0	Demented	15	22
True label \ Predicted label	Nondemented	Demented											
Nondemented	38	0											
Demented	15	22											
Random Forest	0.813	0.703	0.812	 <table border="1"><thead><tr><th>True label \ Predicted label</th><th>Nondemented</th><th>Demented</th></tr></thead><tbody><tr><th>Nondemented</th><td>35</td><td>3</td></tr><tr><th>Demented</th><td>11</td><td>26</td></tr></tbody></table>	True label \ Predicted label	Nondemented	Demented	Nondemented	35	3	Demented	11	26
True label \ Predicted label	Nondemented	Demented											
Nondemented	35	3											
Demented	11	26											

Table: Compilation of the **test** result from different ML models

The best way to solve a Machine Learning problem is through trial and error. It is essential to have a good and comprehensive dataset. We also need to apply as many algorithms as possible to get the best possible result. But due to time constraints, we limited our dataset to only one source. And we were able to explore only four classification algorithms. In order to improve the results, we should incorporate more data from other sources. We can also try out other Machine Learning models like AdaBoost, K-Nearest Neighbor, Bagging, etc.

The most notable problem we faced was figuring out the pattern in our dataset. Even after visualizing our dataset with histograms and scatter plots, it wasn't quite apparent which attributes contributed most towards dementia. Since we are all computer science students, and we aren't very familiar with the medical terms and the value of certain attributes. That's why we didn't include any custom transformations using the attributes of our dataset. It might be possible to get slightly better results with our existing dataset and models with a little more feature engineering.

Conclusion

For predicting Alzheimer's disease or dementia in adult patients, we used the "MRI and Alzheimers" dataset provided by the Open Access Series of Imaging Studies (OASIS) project. We visualized the dataset and filled in the missing values. We pre-processed the data by removing some unnecessary features. We standardized the values to make sure they fit our ML models. Then we used our dataset to train SVM, Logistic Regression, Decision Tree, and Random Forest, models. We used accuracy, recall, AUC, and confusion matrix as evaluation metrics. To improve our result, we fine-tuned all our models using the Grid Search method. For our particular dataset, we got the best result using SVM. More complex model like Random Forest classifier suffered from overfitting issue.

For deployment, we can use our SVM model. In the future, we can improve our model by using a larger dataset.

Our project can help the general public get an idea about the possibility of dementia in adult patients by simply inputting MRI data. Hopefully, it will prompt the patient(s) to get early treatment for dementia and improve their life.

Authorship

Name	ID	Work
Md. Tahrim Faroque Tushar	1621148642	Data visualization & Support Vector Machine
S. M. Al Faruqui	1721395042	Data Pre-processing & Logistic regression
Morshedul Bari Antor	1620345042	Decision Tree
Shafayet Jamil Zim	1620346042	Random Forest

Source Code

Trained on Kaggle notebook. Link: <https://www.kaggle.com/tusharfaroque/cse445-project>

Reference

- [1] Radiological Society of North America (RSNA) and American College of Radiology (ACR), "Alzheimer's Disease," Diagnosis, Evaluation, and Treatment. [Online]. Available: <https://www.radiologyinfo.org/en/info.cfm?pg=alzheimers>
- [2] "What Is Dementia? Symptoms, Types, and Diagnosis," National Institute on Aging. [Online]. Available: <https://www.nia.nih.gov/health/what-dementia-symptoms-types-and-diagnosis>
- [3] B. Stecanella, "An Introduction to Support Vector Machines (SVM)," Monkey Learn Blog, 22-Jun-2017. [Online]. Available: <https://monkeylearn.com/blog/introduction-to-support-vector-machines-svm/>.
- [4] S. Swaminathan, "Logistic Regression - Detailed Overview," Medium, 18-Jan-2019. [Online]. Available: <https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc>.
- [5] P. Gupta, "Decision Trees in Machine Learning," Medium, 12-Nov-2017. [Online]. Available: <https://towardsdatascience.com/decision-trees-in-machine-learning-641b9c4e8052>.
- [6] T. Yiu, "Understanding Random Forest," Medium, 14-Aug-2019. [Online]. Available: <https://towardsdatascience.com/understanding-random-forest-58381e0602d2>.