IOSR Journal of Engineering (IOSRJEN) **www.iosrjen.org**
ISSN (e): 2250-3021, ISSN (p): 2278-8719
Vol. 09, Issue 4 (April. 2019), ||S (II) || PP 10-16

A Detector and Predictor System for Alzheimer's Disease using Naives Bayes and Decision Tree Algorithm

Linda Mary John[1], Ashima Sharma[2], Siddhant Gujarathi[3]
[1](Department of Information Technology, St. John College of Engineering and Management, India)
[2](Department of Information Technology, St. John College of Engineering and Management, India)
[3](Department of Information Technology, St. John College of Engineering and Management, India)

Abstract: All over the world, people are suffering from various types of diseases. Among them, the diseases related to the brain have a huge impact on the lives and health of the patients. Dementia is one such disease. Dementia is a term that describes a group of symptoms that is associated with rapidly declining mental and memory skills. Alzheimer's Disease is a form of dementia that slowly destroys memory and thinking skills and makes it impossible to perform simple tasks. Alzheimer's disease is one of the types of the dementia which contribute to 60-80% of mental disorders. Alzheimer's is an irreparable brain disease, impairs thinking and memory while the aggregate mind size shrinks which at last prompts demise. Diagnosis of this disease at an early stage will help the patients to lead a quality life for the remaining of their life. Early diagnosis of AD is important for the progress of more powerful treatments.

Keywords- Alzheimer's Disease, Dementia, Data Mining, Machine Learning

---

Date of Submission: 27-03-2019 Date of acceptance: 12-04-2019

I. INTRODUCTION

Alzheimer's disease (AD)[1], a type of dementia, is characterized by progressive problems with thinking and behaviour that starts in the middle or old age. The pathologic characteristics are the presence of neurotic plaques in the brain and degeneration of explicit brain cells. The symptoms usually develop slowly and get serious enough to interfere in daily life. Although the paramount risk factor is oldness but AD is not just an old age disease. In its early stages, the memory loss is mild while in the later stages, the patient's conversation and their ability to respond degrades dramatically. The current treatments cannot stop

Alzheimer's disease (AD) from developing but early diagnosis can aid in precluding the severity of the disease and help the patients to improve the quality life. It has been reported that the number of individuals effected with AD will double in next 20 years. The stages of Alzheimer's Disease each has different symptoms and signs associated with it.[2]

A. Mild Alzheimer's disease (early stage)

In the early stage of Alzheimer's, a person may function on his/her own. He or she may still be able to drive car, work and be part of social activities. Despite this, the person may feel as if he or she is having memory disorder or some kind of uneasiness, such as not able to remember familiar words or the location of everyday objects and work places.

Friends, family or others close to the individual begin to notice difficulties while remembering their names. During a detailed medical interview, doctors may be able to detect problems in memory or concentration of the person towards their day to day tasks.Common difficulties include:

- Problems coming up with remembering the right word or name.
- Problem remembering names when introduced to new people
- Challenges performing tasks in social or day-to-day work settings.
- Forgetting material that one has just read from a book or some reading material.
- Not able to find or misplacing a valuable object
- Increasing trouble with planning or organizing the tasks or activities.

B. Moderate Alzheimer's disease (middle stage)

Moderate Alzheimer's is typically the longer stage and can make the person to suffer for many years. As the disease progresses, the person with Alzheimer's will require a greater immense of care.

During the moderate stage of Alzheimer's, the dementia symptoms are more severe and visible. A person may face greater problems while performing tasks, such as paying bills, but they may still remember some significant details about their life.

At this point, symptoms will be visible/noticeable to others and may include:

- Fugue of events or about one's own personal history
- Feeling blackout or withdrawn, especially in socially or mentally challenging situations
- Being unable to remember their own address or telephone number or the high school or college from which they graduated
- Not able to recognise about where they are or what day it is
- They are needed for help choosing proper clothing for the season or the occasion
- Problems in controlling bladder and bowels in some individuals
- Changes in sleeping patterns/habits, such as sleeping during the day and becoming disturbed at night
- An increased exposure/hazard of wandering and becoming lost
- Personality and behavioural changes, including suspiciousness and delusions or compulsive, repetitive behaviour like hand-wringing or tissue shredding.

C. Severe Alzheimer's disease (late stage)

In the final stage of this disease, dementia symptoms are relentless. Individuals to lose the ability to correspond to their environment, to convey on a conversation and, eventually, to control their movement. They may still say words or phrases, but communicating pain becomes severe. As memory and cognitive skills continue to decline, significant personality changes may take place and individuals need tremendous help with daily activities.

At this stage, individuals may:

- Need 24/7 assistance with daily activities and personal care.
- Lose consciousness of recent experiences as well as of their surroundings
- Experience changes in physical abilities, including the ability to walk, sit and, eventually swallow eating materials
- Have increasing difficulty communicating with other people
- Become vulnerable to infections, especially pneumonia.

Machine learning is used to interpret and analyse data. Furthermore it can classify patterns and model data. It permits decisions to be made that couldn't be made generally utilizing routine systems while sparing time and endeavours.

II. DATA MINING

Data Mining is the process of sorting through large sets of data to identify patterns and establish relationships to solve problems and provide efficient results through data analysis. Data Mining tools allow enterprises to forecast future trends. Many other terms carry a similar or little different meaning to data mining, such as knowledge mining from data, knowledge extraction, data/pattern analysis, data archaeology, and data dredging. Many people think data mining as an equivalent for another popularly used term, Knowledge Discovery from Data, or KDD.Alternatively, others prospect data mining as simply an essential step in the process of knowledge discovery.

Data preparation and cleaning is an often ignored but extremely important step in the data mining process. The old saying "garbage-in-garbage-out" is particularly applicable to the typical data mining projects where large data sets collected through some automatic methods (e.g., via the Web) serve as the input to the analysis. Also, the method by which the data where gathered was not tightly controlled, and so the data may contain out-of-range values (e.g., Income: -100), impossible data combinations (e.g., Gender: Male, Pregnant: Yes), and the like. Analysing data that has not been carefully screened for such problems can produce highly improper results, in particular in predictive data mining.

Data mining is known as the process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems. Data mining is an interdisciplinary subfield of computer science and statistics with an overall goal to extract information (with intelligent methods) from a data set and transforming the information into a comprehensible structure for further use. Data mining is the analytic step of the "knowledge discovery in databases" process, or KDD. Aside from the aforementioned raw analysis step, it also involves database and data management aspects.

Data preprocessing model and inference considers, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating.The difference between data analysis and data mining is that data analysis is used to summarize the history such as analyzing the effectiveness of a marketing campaign, in contrast, data mining mainly focuses on using specific machine learning and statistical models to predict the future and discover the patterns among data.

Machine learning (ML), in computing is the study of algorithms and statistical models that computer systems use to progressively improve their performance on a specific task. Machine learning algorithms build a mathematical model of sample data, which is known as "training data", in order to make predictions or decisions

without being explicitly programmed to perform the task.Machine learning is closely related to computational statistics, that focuses on making predictions using computers. The study of mathematical optimization delivers to the field of machine learning- methods, theory and application domains. Data mining is a field of study within machine learning, and it focuses on exploratory data analysis through unsupervised learning. It is used across business problems, machine learning is also referred to as predictive analytics.[3]

A. Importance of Data Mining

It can be seen that Data Mining has became beneficial to a lot of party and multiple range of level in the organization as the model or framework that is apply and can result in reducing the time and cost. Then, the results allow the responsible knowledge worker to transform the analytical results into the strategic value of information effectively by critically evaluate the result.The process should be carried out carefully to avoid the useful variables or algorithm being removed or not being included in the extraction of determined data. Data mining techniques will help in select a portion of data using appropriate tools to filter-out outliers and anomalies within the given set of data. In order to find the pattern of data, a few methodologies are use in clarifying the ambiguity as well as to identifying the relation among one variables and other variables within the databases whereas the result will guide in making decision or for planning the impact when the action were taken into consideration.

B. Data Mining in Medical field

Data mining technology is emerging as a promising field and is used in broad application areas like ecommerce, microarray- gene expression data, scientific experiments etc. This technology intermixture various analytic methods with advanced and sophisticated algorithms which helps in exploring large volumes of data. It also plays a coral role in the early detection of diseases. There exist number of application areas of data mining in medical industry. It is crucial that data mining techniques like classification, clustering etc. should be applied to hospital databases so that the right treatments can be provided to patients at the right time which in turn will lower loss of life rate. Classification approach works by first developing a model from training data and then it is applied on testing data for the prediction of unknown data. In healthcare sector, classification and prediction is used extensively in disease foreseeing.

A unique approach was developed in using ANN algorithm for the prediction of heart disease. The researchers developed an interactive prediction system using the classification through artificial neural network algorithm with the deliberation of 13 most important clinical factors. The proposed approach was very efficient and user friendly for heart disease prediction with 80% accuracy and can be of great used for healthcare specialist. An productive prediction system was designed in to predict the risk level of heart patients. The system could discover rules efficiently from the dataset using decision tree approach according to the given guideline related to patient`s health. Authors concluded that the system can detect the risk level of heart disease risk level to a great extent.

C. Naives Bayes

The Naïve Bayes classification algorithm is a probabilistic classifier. It is based on probabilistic model that incorporate strong independent assumptions. The independent assumptions often do not have an impact on reality. Therefore they are considered as naïve. Naivesbayes classifiers are scalable and require a number of parameters which is linear in the number of variables (features/predictors) in a learning problem. Naive Bayes is a very simple technique for construction of classifiers i.e., models that assign class labels to problem instances, which are represented as vectors of feature values, where the class labels are drawn from some finite set. It is not a single algorithm for training classifiers, but rather a family of algorithms based on a common principle: all naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable.[4]

D. Decision Tree

A decision tree is a decision support tool that uses a tree-like model of decisions and their possible consequences, which includes chance event outcomes, resource costs, and utility. Decision tree is one of the ways to display an algorithm that only contains conditional control statements. Decision trees are commonly used in operations research, especially in decision analysis, to help identify a strategy which most likely reaches a goal, but they are also a popular tool in machine learning.[5] The core algorithm for building decision trees called ID3 by J. R. Quinlan, employs a top-down, greedy search approach through the space of possible branches with no backtracking. ID3 uses Entropy and Information Gain data to construct a decision tree. In ZeroR model there is no predictor, in OneR model the algorithm tries to find the single best predictor. Naive Bayesian includes all predictors using Bayes' rule and the independence

assumptions between predictors but decision tree includes all predictors with the dependence assumptions between predictors.[6]

III. PROPOSED SYSTEM

For the purpose of detection and prediction of Alzheimer's Disease, we propose a system that will not only easily detect the disease, but also provide the patients and medical practitioners with accurate results.

The working of the system is divided into five major phases. These are:

PHASE 1- Firstly a questionnaire is prepared by a doctor which takes into account all the questions that can be used to diagnose a person suffering from Alzheimer's Disease. These questions should range from person details of the patient's life to verbal and mathematical questions. All these questions can help the doctor understand which stage of the disease the person is currently in. The questions can be divided stage-wise to ensure easy detection of the stage. Stage one questions can include all the personal stories and information about the patients like asking about their anniversary or birthdays. The second stage questions can include the state or country the person belongs in. Patients generally lose sense of time and questions like today's date or current year can help detect this stage. Questions in the final stage can include both verbal and mathematical riddles along with a few commands. This stage leads to the person being unable to follow commands and asking them to follow simple commands like 'Close your eyes' can help detect this stage. The questionnaire is answered both by the patient as well as the relative. Although, the answers provided by the relatives are only useful for stage one questions, since this stage consists of personal questions. The answers provided by both the patient and the next-of-kin are matched to provide the results of stage one answers. Other answers are also entered into the database and a resultant dataset is created for each user.

PHASE 2- This stage is divided into further substages.

PHASE 2.1 – This substage uses Decision Tree algorithm on the resultant dataset that helps us to classify whether the patient is suffering from Alzheimer's Disease or not. Decision Tree algorithms are a part of the supervised learning algorithms. Decision Tree algorithms are used for regression and classification problems. It is generally used to create a training model that can predict class or the value of target variables from the training data. The best attribute of the dataset is selected as the root of the tree. The training set data is split into subsets. While creating subsets, each subset must contain data with the same value for an attribute. These steps are repeated until all the leaf nodes are found.In decision trees, we start from the **root** of the tree for predicting a class label for a record. The values of the root attribute are compared with record's attribute. On the basis of comparison, the branch corresponding to that value is followed and we jump to the next node.

PHASE 2.2 – After it is detected that the person is in fact suffering from Alzheimer's Disease, Naives Bayes algorithm is used on the resultant dataset to detect the stage of the disease the person is currently in. This can help the doctors diagnose the severity of the disease and provide the appropriate prescription to the patient. Naives Bayes algorithm is a classification algorithm. It works on a large dataset with multiple attributes. Naives Bayes algorithm makes use of the Bayes Theorem. Bayes Theorem works on conditional probability, meaning the probability of an event happening, given that another event has already occurred. The conditional probability is calculated as:

$$P(H\ |E) = \frac{P(E\ |H)\ *\ P(H)}{P(E)}$$

P(H) is the likelihood of hypothesis H being true. P(H) is known as the prior probability.

P(E) is the probability of the event (irrespective of the hypothesis).

P(E|H) is the probability of the event given that hypothesis is true.

P(H|E) is the probability of the hypothesis given that the event is there.

In Naive Bayes classifier, it is assumed that all the features are unrelated to each other. Presence or absence of a feature does not affect the presence or absence of any other feature. A frequency table for each attribute is created and the likelihood of each feature is calculated. Based on the likelihood, the conditional probabilities for each classes is determined, and the class with the maximum conditional probability is considered.

PHASE 3- Based on the results, the patient is provided with an output, which helps the patient know which stage of Alzheimer's Disease he/she currently is in. Detecting the stage is a very important part because the stages are detected based on the answers provided by the patients. Also, detecting the stage helps the doctor understand the intensity and effect of the disease. Analysis of the disease is done using a visualization tool called Orange. Orange is an open source toolkit for data visualization, data mining, data analysis and machine learning. It uses common libraries of Python for computation. Recommending the person with the do's and don'ts helps prevent the impact of this disease. These recommendations will also include measures to reduce the further effects of the disease on the patient.

PHASE 4 – An electronic generated diagnosis is presented to the patient which includes the stage of Alzheimer's Disease the patient is currently in as well as the recommendations that are provided to the patients to curb or reduce the impact of the disease on the patient's brain and health.

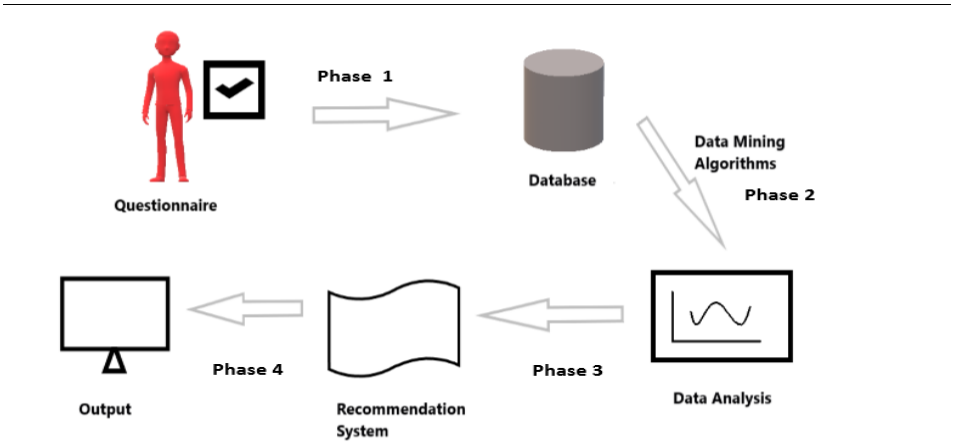


Fig 1: Proposed Architecture

IV. RESULTS AND DISCUSSIONS

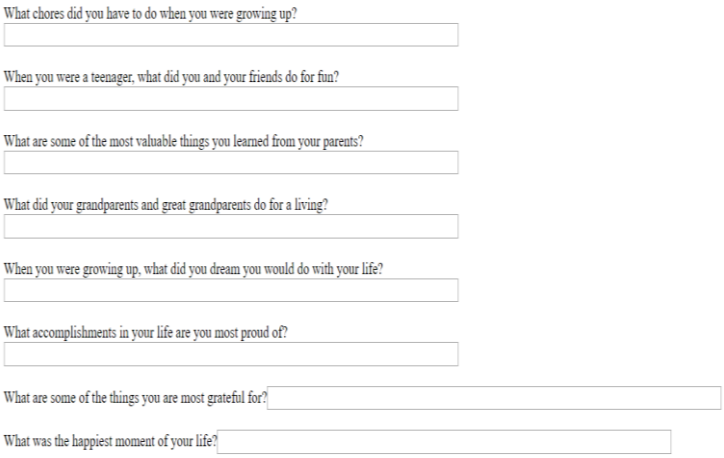


Fig 2: Stage 1 questions

This stage consists of the questions that will provide the doctors with an insight into the patient's personal life. These questions can help us determine stage 1 of AD.

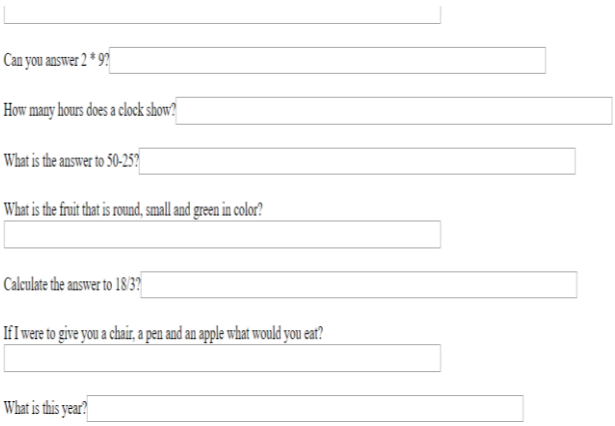


Fig 3: Stage 2 questions

This stage consists of the mathematical and verbal questions that will help the doctors determine stage 2 of AD

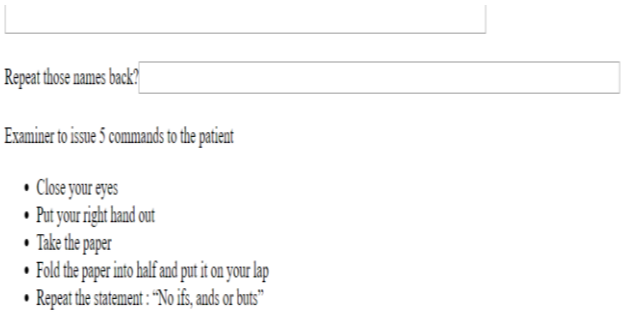


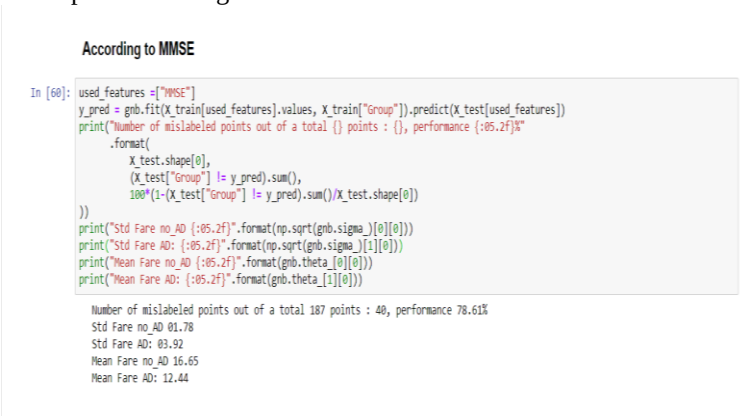
Fig 4: Stage 3 questions

Since stage 3 of AD can affect patients to strugle with commands and actions, these commands can help detect this stage successfully.

```
In [8]: df["SES"].fillna(df["SES"].median(), inplace=True)
        df["MMSE"].fillna(df["MMSE"].mean(), inplace=True)

In [9]: # Encode columns into numeric
        from sklearn.preprocessing import LabelEncoder
        for column in df.columns:
            le = LabelEncoder()
            df[column] = le.fit_transform(df[column])

In [10]: from sklearn.model_selection import train_test_split

         feature_col_names = ["M/F", "Age", "EDUC", "SES", "MMSE", "eTIV", "nWBV", "ASF"]
         predicted_class_names = ['Group']

         X = df[feature_col_names].values
         y = df[predicted_class_names].values

         X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.30, random_state=42)

In [53]: from sklearn.tree import DecisionTreeClassifier
         from sklearn import tree

In [59]: model = tree.DecisionTreeClassifier()
         model

Out[59]: DecisionTreeClassifier(class_weight=None, criterion='gini', max_depth=None,
                     max_features=None, max_leaf_nodes=None,
                     min_impurity_decrease=0.0, min_impurity_split=None,
                     min_samples_leaf=1, min_samples_split=2,
                     min_weight_fraction_leaf=0.0, presort=False, random_state=None,
                     splitter='best')
```

Fig 5: Decision Tree results

Decision tree algorithm helps us to successfully and efficiently detect the disease in the patient. The results from this algorithm can either be positive or negative.



According to MMSE

```
In [60]: used_features =["MMSE"]
         y_pred = gnb.fit(X_train[used_features].values, X_train["Group"]).predict(X_test[used_features])
         print("Number of mislabeled points out of a total {} points : {}, performance {:05.2f}%"
               .format(
                   X_test.shape[0],
                   (X_test["Group"] != y_pred).sum(),
                   100*(1-(X_test["Group"] != y_pred).sum()/X_test.shape[0])
         ))
         print("Std Fare no_AD {:05.2f}".format(np.sqrt(gnb.sigma_)[0][0]))
         print("Std Fare AD: {:05.2f}".format(np.sqrt(gnb.sigma_)[1][0]))
         print("Mean Fare no_AD {:05.2f}".format(gnb.theta_[0][0]))
         print("Mean Fare AD: {:05.2f}".format(gnb.theta_[1][0]))

         Number of mislabeled points out of a total 187 points : 40, performance 78.61%
         Std Fare no_AD 01.78
         Std Fare AD: 03.92
         Mean Fare no_AD 16.65
         Mean Fare AD: 12.44
```

Fig 6: Naives Bayes results

Naives Bayes algorithm helps us to successfully and efficiently detect the stage of the disease the patient currently suffers from. The results from this algorithm will determine the stage.

V. CONCLUSION

According to the 2018 reports by Alzheimer's Association, sharp increase has been witnessed in Alzheimer's prevalence, deaths and the costs of care.[7]In America alone, around 5.7 million people are suffering from this disease which has led to researchers and doctors trying to find ways to reduce the diseases' presence. Machine Learning is one of the ways in which the data can be successfully analyzed, patterns can be detected and appropriate actions can be taken. In this paper, the two main Machine Learning algorithms selected are Naives Bayes and Decision Trees.These algorithms were selected due to their accuracy and efficiency. These algorithms were appropriate to use with the dataset available with us. The sensitivity provided by these algorithms helped in the successful classification of the data. In review, other algorithms which provide high level of efficiency as well as accuracy can also be used

ACKNOWLEDGEMENTS

We would like to forward our sincere gratitude to the principal of our college for providing us with the opportunity to work on this paper. We would also like to thank him for providing us with all the resources which were required. We would also like to extend our gratitude to the teaching and non-teaching staff as well as our colleagues for making this paper a huge success and for investing their time and efforts in this paper.

REFERENCES

[1] https://www.alz.org/alzheimers-dementia/what-is-alzheimers, accessed on 15/10/2018
[2] https://www.alz.org/alzheimers-dementia/stages, accessed on 07/01/2019
[3] https://en.wikipedia.org/wiki/Machine_learning, accessed on 12/09/2018
[4] https://en.wikipedia.org/wiki/Naive_Bayes_classifier, accessed on 7/01/2019
[5] https://en.wikipedia.org/wiki/Decision_tree, accessed on 7/01/2019
[6] https://www.saedsayad.com/decision_tree.htm, accessed on 7/01/2019
[7] https://www.medgadget.com/2017/12/future-scope-of-alzheimers-disease-diagnostic-market-which-is-expected-to-grow-at-a-cagr-of-10-top-key-players-profile-forecast-to-2022.html, accessed on 09/01/2019
[8] https://alzheimerscareresourcecenter.com/2018-alzheimers-disease-facts-figures-report, accessed on 09/01/2019.