# A STUDY OF PERFORMANCE COMPARISON OF DIFFERENT MACHINE LEARNING ALGORITHMS

Arunkumar Sudhakaran Nair
MSc Data Analytics,
*School of Computing,*
*National College of Ireland,*
Dublin, Ireland
x19153317@student.ncirl.ie

***Abstract*—Machine learning analysis can help to explain how critical tasks can be accomplished by training models to make predictions or classifications based on observed data. The main aim of this research paper is to compare various machine learning methods. The machine learning methods used in this research paper are KNN, Decision Tree, Linear Regression, Random Forest, and Logistic regression being applied to 3 large datasets. The probe can be branched into three subsections, the first one being the analysis of factors affecting road accident severity. The second part is the prediction of suicide rates. The third part aims at predicting medical appointment no shows. The accuracy of each model is evaluated at the end of the analysis for comparison.**

***Keywords—KNN, Decision Tree, RandomForest, Linear Regression, KDD***

## I. INTRODUCTION

Road accidents have become a worldwide concern and the ninth leading cause of death worldwide in recent years. This is the main reason which made me take up this topic for analysis. The dataset chosen here is the data about accidents in the USA. It is a national data collection of traffic accidents affecting 49 US states. The dataset consists of 1.4 million rows and 49 columns which makes it a large dataset. Road accidents are the leading cause of death of US citizens flying abroad every year. Around 1.35 million fatalities occur in road accidents annually with an average of 3700 fatalities on road every day. In addition to that 20-50 million non-fatal accidents often lead to long term disabilities[1]. The research question being addressed here will be the factors affecting the severity of the accidents. The main aim of analyzing the severity of an accident is to address those factors and stamp them out. The machine learning methods used to analyze the factors affecting severity are K-Nearest Neighbour and Decision Trees. The prediction accuracy of the KNN algorithm has been analyzed and the performance improvement method was implemented for the model. The performance improvement method implemented was the Elbow method which assists in the selection of the optimal value of k for performing KNN.

The second topic of discussion is the overview of suicide rates from 1985 to 2016. The consideration of this topic is after estimating the fact that about 8 million people all over the world die due to suicide every year. This makes a count of one person every 40 seconds. This number is also possibly underestimated due to the stigma associated with suicide and the fact that it is illegal in some countries, with some suicides being listed as unintentional wounds. Suicides are also one of the leading causes of death in the world. The rate of deaths due to suicides is twice as that of murders. Suicides are common in most parts of the world it is almost up to ten to twenty times higher than murders[2]. A large number of people with psychological problems, depression, and drug abuse have attempted suicide. Unemployment and social alienation have been related to high suicide rates among social factors. The research question being addressed here is the prediction of suicide rates. The dataset chosen for the analysis is the 'suicide rates overview 1985 to 2016' dataset in Kaggle. The dataset contains around 28000 rows and 12 columns. This project uses historical data to make predictions of future suicide rates through developing predictive machine learning models of suicide rates using linear regression. Multiple factors can lead a person to commit suicide. These factors include mental illness, drug abuse, social circumstances, relationships, and life events. Here, I aim to predict the chance of suicide based on the age, sex, and generation of a person. The machine learning method used for prediction is linear regression. The accuracy of the Linear regression model is evaluated using RMSE value.

Community health centers provide primary care for poor and uninsured people. High rates of missing the appointment timing have been reported as one of the primary barriers in providing the required care for the people in need of real care. The third part of the research aims at analyzing the medical appointment no shows associated with healthcare. The research question being addressed at this part will be to predict the medical appointment No shows using Logistic Regression. No shows or patients missing their scheduled appointments are popular in healthcare and costly. A US study found that up to 30% of patients missed their appointments and 150 billion dollars are missed every year. Analyzing the reason for potential no-shows will help medical facilities introduce tailored interventions to reduce no-shows and losses[3]. The dataset was chosen from Kaggle. The dataset has around one hundred thousand rows and 14 columns. After cleaning and preprocessing, Random Forest and Logistic Regression methods were performed. The accuracy of Random Forest is

evaluated using a confusion matrix. The Logistic Model is evaluated using the ROC Curve.

Data mining refers to the process of analyzing a huge amount of data and extracting meaningful insights from it. The Knowledge Discovery in Databases is the data mining process model used in this project. The primary goal of the KDD process is the extraction of information from data in broad databases. The steps followed for interpreting patterns from data involved the application of KDD methods and procedures.

## II. Related Work

Over the years several studies were conducted to explore and analyze the factors contributing to the severity of motor vehicle crashes.

Machine learning methods have been used extensively during traffic prediction in recent years as they have been able to process multidimensional knowledge, flexibility in deployment, simplicity, and strong predictive abilities. Concerning the prediction of accident severity in the traffic, Kunt et al[4] used a pattern search and a multilayer (MLP) structural modeling in a neural artificial net (ANN) to predict the severity of road accidents by using twelve accidental parameters in GA pattern search and multilayer perceptron genetic algorithm(MLP). The models were developed based on 1000 accidents on the Tehran- Ghom Freeway in 2007. The best model was chosen based on R-value, RMSE, MAE, and SE. The highest value of R was obtained for ANN which indicated that ANN was the best predictor.

The solution to the potential study of traffic accidents using data mining techniques was suggested by S.krishnaveni and Dr.M Hemalatha [5]. Different classification methods were used to predict severity due to accidents. The injury due to accidents was classified using different classifiers like Naïve Bayes, Ada Boost, J48 Decision Tree, and Random Forest. In the end, Random Forest outperformed the other 4 algorithms.

TibebeBeshah, Shawndra[6] Hill used RTA data from Ethiopia to research the role of road-related factors in the severity of accidents. The main aim was to explore the underlying road-related variables which have a negative impact on the severity of car accidents and predict the severity of accidents using different data mining techniques. At the end of the analysis, Decision Tree and KNN received about 80 percent accuracy and Naïve Bayes got about 79 percent accuracy.

In another study, T.tesema, A. Abraham, and C.Grosnan [7] proposed a method to analyze accident crash severity. They used an algorithm to divide the dataset into subsets and they trained the classifiers for every subset of data. Upon completing the preprocessing, there were 4658 records of 16 attributes (13 bases and 3 derived) in the final dataset for modeling. The Levenberg-Marquardt algorithm has been used for MLP training and achieved a classification accuracy of 65.6 and 60.4 percent respectively during training and testing. They compared the results of the Multi-Layer Perception (MLP) and the Fuzzy ARTMAP and found that the MLP classifier has higher accuracy than that of Fuzzy ARTMAP.

Montella et al[8] used two machine learning algorithms namely rules discovery and classification trees for analyzing motorcycle dependencies and differences among characteristics which lead to a motorcycle crash. The result of this analysis provided insights into developing methods for safety improvement. The conclusion was that both rules discovery and classification trees provided meaningful insights about crash characteristics.

From the literature, it has been noticed that multiple factors were identified to contribute to crash severity including weather, behavioral patterns, and type of vehicle[24]. Also, multiple studies were carried out using different machine learning methods which yielded distinctive results

Several studies were performed to predict the factors affecting suicide rates. Some of them include a study conducted by Pasos et al.[9]. He distinguished suicide attempters from nonsuicide attempters among patients with mood disorders using machine learning algorithms with an accuracy of 65-72%. Delgado-Gomez, et al[10], collected 849 cases from a hospital in Spain and scored them manually. These scores were then used for the evaluation of classifiers including Boosting, Linear Discriminate Analysis, Fischer Linear Discriminate Analysis, and Support Vector Machine. They found that these manual scores could be used to find people with suicidal tendencies. Huang, et al.,[11] used a technique of keyword matching on myspace.com to find out if the users have an intent to commit suicide. The method recognized 14% suicidal users.There were not many research papers available for review since the related works for factors affecting suicide rates is limited.

Few previous studies included patient characteristics in their scheduling strategy. One research proposes to reserve patients until the number of appointment slots available for the day exceeds the planned arrivals. The estimated number of arrivals is calculated based on certain conditions as if age is greater than or less than 14 and the number of previous appointments[12].

Some studies have also analyzed patients and healthcare providers and compared their points of view on the reason for not attending. The factors contributing to non-attendance relates to inaccessibility factors such as physical location, opening hours, language stigma, and cultural differences. Some studies within psychiatry found that alcohol and drug users will have a profound impact on attendance rates. These studies have produced conflicting results[14].

The study conducted by Williamson et al. and Ellis et al.[13] emphasized on patient's demographic and realistic variables that predict missed serial appointments. Logistic regression is mostly used for predicting the probability of no show in appointments. It is done by fitting a numerical or categorical data into a logit function[15].

Most research in this field is related to factors contributing to nonattendance in specialty and all appointments from GP. The data found from hospitals administrative database stated that a variety of factors were found effective on a patient’s attendance in pediatric urology unit[16], pulmonary rehabilitation[17][18], psychiatric[19]-[20] and HIV[21], primary care [22], inpatient and outpatient in the hospital[23] by analyzing multiple correlations.At present, there is no proper understanding as to what works to reduce missing appointments. While considering a large number of variables, we used a deep learning approach and extract key features and

complexities from large datasets and more specifically at the individual level.

## III. METHODOLOGY

The data mining method followed here is Knowledge discovery in databases. It follows a series of steps to extract meaningful information from data. The steps followed during the analysis are explained as follows.

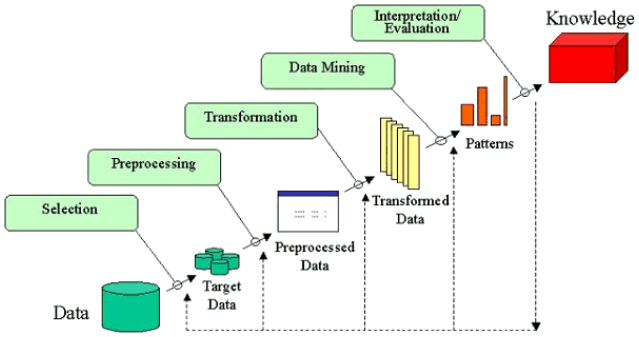



*Figure 1: KDD Process Flow*

### A. Data Source Selection

Three datasets were chosen for analysis. The datasets include the US Accidents dataset which consists of 1048576 rows and 49 columns. The second dataset is the suicide rate overview 185 to 2016 data which consists of 27821 rows and 12 columns. The third dataset is the Medical Appointment No-Shows data which consists of 110528 rows and 14 columns.

### B. Data preprocessing

The accident data consisted of more than 1 million rows, so the data size was reduced to 10000 rows for analysis using the 'nrows' function and the values were stored in a data frame. Also, there were 49 columns and only useful columns were chosen for analysis. My primary aim was to check for variables that can be used to help to predict the severity of accidents. So humidity and visibility were considered.

```
'data.frame':   10000 obs. of  3 variables:
 $ Humidity...   : num  91 100 100 96 89 97 100 100 99 100 ...
 $ Visibility.mi.: num  10 10 10 9 6 7 7 7 5 3 ...
 $ Severity      : int  3 2 2 3 2 3 2 3 2 3 ...
```

*Figure 2: data frame*

After selecting the required variables, the next step was to check for NA values. This was done using 'is.na' function. Both the variables were having NA values and they were removed using 'na.omit'. A similar operation was performed for all datasets.

```
> sum(is.na(accident_subset$Humidity...))
[1] 124
> sum(is.na(accident_subset$Visibility.mi.))
[1] 96
> colSums(is.na(accident_subset))#humidity and visibility has
   Humidity... Visibility.mi.       Severity
           124             96              0
> cleaned_data<-na.omit(accident_subset)
> colSums(is.na(cleaned_data))# checking if there is any na v
   Humidity... Visibility.mi.       Severity
             0               0              0
```

*Figure 3: omitting NA values*

### C. Transformation

The cleaned data was then transformed into a format that is suitable for the function to be performed. The categorical variables are changed into factors with their respective number of categories.

```
$ Scholarship   : int  0 0 0 0 0 0 0 0 0 0 ...
$ Hipertension  : int  1 0 0 0 1 1 0 0 0 0 ...
$ Diabetes      : int  0 0 0 0 1 0 0 0 0 0 ...
$ Alcoholism    : int  0 0 0 0 0 0 0 0 0 0 ...
$ Handcap       : int  0 0 0 0 0 0 0 0 0 0 ...
```

*Figure 4: integer values*

The values of scholarship, hypertension, diabetes, alcoholism, and handicap were transformed into factors with the number of levels equal to their respective number of categories.

```
$ Scholarship   : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
$ Hipertension  : Factor w/ 2 levels "0","1": 2 1 1 1 2 2 1 1 1 1 ...
$ Diabetes      : Factor w/ 2 levels "0","1": 1 1 1 1 2 1 1 1 1 1 ...
$ Alcoholism    : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
$ Handcap       : Factor w/ 5 levels "0","1","2","3",..: 1 1 1 1 1 1 1 1 1 1 ...
```

*Figure 5: Transformed into factors*

Once the data has been transformed into the datatype suitable for feeding into the model, the data is then partitioned into test and train datasets. This can be easily done using the 'catools' library. Figure 6 shows the data after partition into the test and train datasets.

| Data | |
|---|---|
| data | 110527 obs. of 14 variables |
| new | 110527 obs. of 1 variable |
| new_data | 110527 obs. of 11 variables |
| rf | List of 19 |
| test | 21915 obs. of 11 variables |
| train | 88612 obs. of 11 variables |

*Figure 6: Test and Train*

The next step is to feed the train data into the model to train the model. After the model is trained, the test data is fed into the model for predictions, and accuracy is checked.

### D. Data Mining

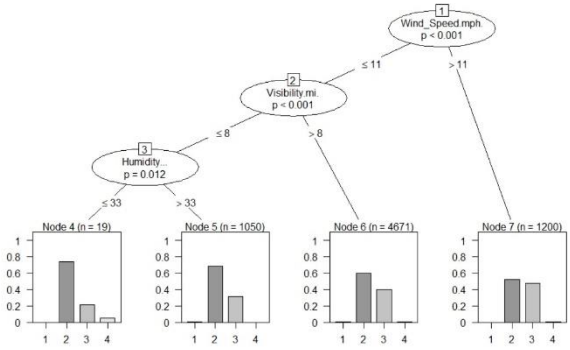
In this step, we have to decide the type of machine learning models whether we have to use classification, regression, or clustering. In this research, I have used five machine learning models. The models used are explained below.

➢ KNN

The K-Nearest Neighbour Algorithm is applied to the US Accidents dataset. KNN is a form of supervised machine learning algorithm which can be used for both classification and regression problems. KNN is used to analyze the factors affecting the severity of accidents. The actual k value if found using square root of 'Nrow'. The k value was found to be 83 and nearset neighbour as 84. Elbow method was used for performance improvement and the highest accuracy was found at k=1.

➢ DECISION TREE

A decision tree is a form of supervised machine learning. The goal of the decision tree is to construct a model that predicts the value of the target variable by learning from the data rules. Here, the decision tree model is performed using a function named 'party'. The data is trained based on factors such as visibility, humidity, and wind speed.prediction is done for the test data and misclassification error is noted.




➢ RANDOM FOREST

Random Forest is a type of machine learning algorithm which is a combination of different algorithms. It works itself by creating a poll for the final class and since it is a combination of multiple algorithms, the one with the highest vote gets the final poll. Random forest is used to predict the No Shows in medical appointments and the accuracy of the model is evaluated using a confusion matrix.

➢ LOGISTIC REGRESSION

Logistic Regression is a machine learning algorithm used for solving classification problems where the output tends to be binary. Here Logistic Regression is used to predict the No-Shows in a medical appointment. Here, the expected output of the problem is either a 'yes' or 'no'. The model predicts whether a patient will show up for the medical appointment or not based on the data.

➢ LINEAR REGRESSION

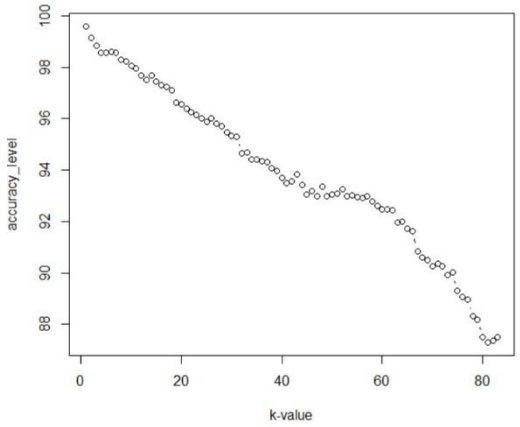
Linear Regression is used to find out relationships between variables and forecasting. In our research paper, I have used linear regression to predict the rate of suicide rates. The predicted values are compared with actual values to gain insights on the performance of the model.

## IV. EVALUATION

Model evaluation metrics are required to evaluate a model's efficiency. The choice of evaluation method depends on the machine learning task. The main evaluation methods used to evaluate the accuracy of the model include the confusion matrix and AUC. The evaluation of different models used in different datasets are illustrated below.

### *A. US Accidents Data*

The KNN algorithm provided an accuracy of 87.466 which is a good value . However, using elbow method the accuracy can further be improved to a value of almost 99%. It is shown in the figure below.




Also, the evaluation method used for decision tree is misclassification error which is 0.397 for train data and 0.389 for test data.

### *B. Suicide dataset*

The machine learning method used for this analysis is linear regression. The evaluation method used for linear regression is Root Mean Square Error or RMSE. The RMSE value for this data is 27.65.

```
> rmse
[1] 27.65515
```

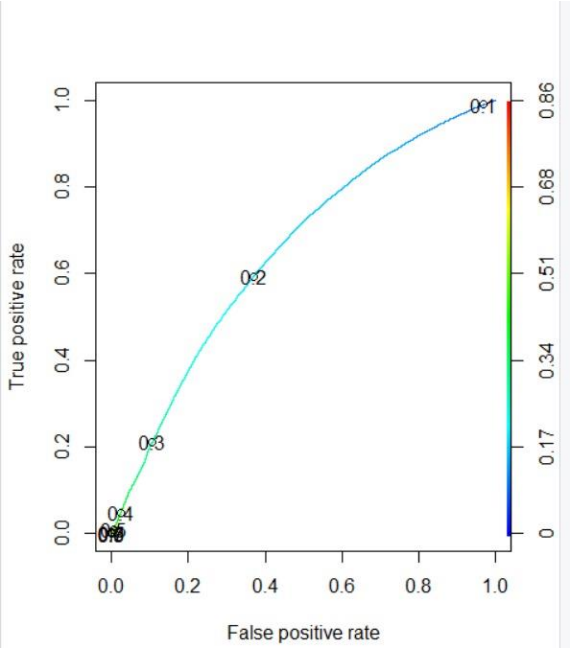
C. Medical No shows data

The machine learning method used for this analysis is random forest and since the classification was performed, a confusion matrix is used for evaluation.

```
> confusionMatrix(p1,train$No.show)
Confusion Matrix and Statistics

          Reference
Prediction    No   Yes
       No  70622 16707
       Yes     1  1159

               Accuracy : 0.8112
                 95% CI : (0.8086, 0.8138)
    No Information Rate : 0.7981
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.0997

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.99999
            Specificity : 0.06487
         Pos Pred Value : 0.80869
         Neg Pred Value : 0.99914
             Prevalence : 0.79810
         Detection Rate : 0.79809
   Detection Prevalence : 0.98689
      Balanced Accuracy : 0.53243

       'Positive' Class : No
```

From the values, it is evident that the accuracy is 0.8112.

Also, on performing logistic regression, the threshold value was finalized to 0.3 which gave an accuracy of 75.71. The roc curve for this analysis is shown below.




## V. Conclusion and Future Work

On comparing five machine learning algorithms, we can conclude that KNN had the highest accuracy of 87.66. Then the logistic Regression comes second with accuracy or 75.71. Also, the Random forest had an accuracy of 0.812. the RMSE value of linear regression was around 27.65. The misclassification error of the decision tree was around 0.389. From the analysis, it is evident that KNN had the highest accuracy and by using the Elbow method, we can predict which value of k has the highest accuracy and implement it in our model so that the model performance can be improved. The future scope of this analysis includes applications to predict the crash severity using KNN with improved accuracy. Also in the field of suicide prediction, Linear regression can be used to predict the suicide rates with improved accuracy. The prediction of No-shows in medical appointments is a primary concern in the field of healthcare. Implementing a Logistic Regression model with 75.71% accuracy can improve the performance of the model and predict the occurrence of no shows with improved accuracy. Due to the limitation of time, the application of other machine learning models like Neural Networks with a huge scope of application into the areas of research was omitted. Also, the application of performance improvement methods like the Elbow method could have used to improve performance accuracy.

## References


[1]"Road Safety Facts — Association for Safe International Road Travel", Association for Safe International Road Travel, 2020. [Online]. Available: https://www.asirt.org/safe-travel/road-safety-facts/. [Accessed: 04- May- 2020].

[2]H. Ritchie, M. Roser and E. Ortiz-Ospina, "Suicide", Our World in Data, 2020. [Online]. Available: https://ourworldindata.org/suicide. [Accessed: 04- May- 2020]

[3]"Using RandomForest to Predict Medical Appointment No-shows", Medium,2020.[Online].Available:https://towardsdatascience.com/using-randomforest-to-predict-medical-appointment-no-shows-b33575e3ff42. [Accessed: 04- May- 2020]

[4]2020.[Online].Available:https://www.researchgate.net/publication/254310836_Prediction_for_traffic_accident_severity_Comparing_the_artificial_neural_network_genetic_algorithm_combined_genetic_algorithm_and_pattern_search_methods. [Accessed: 04- May- 2020]

[5Ijcaonline.org,2020.[Online].Available:https://www.ijcaonline.org/volume23/number7/pxc3873788.pdf. [Accessed: 04- May- 2020]

[6]"Mining Road Traffic Accident Data to Improve Safety - Artificial ... - MAFIADOC.COM", mafiadoc.com, 2020. [Online]. Available: https://mafiadoc.com/mining-road-traffic-accident-data-to-improve-safety-artificial-_59e225551723dd1c74b4a1f3.html. [Accessed: 04- May- 2020]

[7]T. Tesema, A. Abraham, and C. Grosan, "Rule mining and classification of road traffic accidents using adaptive regression trees. I,"Journal of Simulation, vol. 6, no. 10, pp. 80–94, 2005

[8]Montella A, Aria M, D'Ambrosio A, Mauriello F. Analysis of powered two-wheeler crashes in Italy by classification trees and rules discovery. Accid Anal Prev. 2012;49: 58–72. pmid:23036383

[9] Passos IC, Mwangi B, Cao B, Hamilton JE, Wu MJ, Zhang XY, et al.

[10] D. Delgado-Gomez, H. Blasco-Fontecilla, A. A. Alegria, T. Legido-Gil, A. Artes-Rodriguez, and E. Baca-Garcia, "Improving the accuracy of suicide attempter classification," Artificial intelligence in medicine, vol. 52, no. 3, pp. 165–168, 2011

[11]Y.-P. Huang, T. Goh, and C. L. Liew, "Hunting suicide notes in web 2.0-preliminary findings," in Multimedia Workshops, 2007. ISMW'07. Ninth IEEE International Symposium on. IEEE, 2007, pp. 517–521

[12]Shonick W and Klein BW. An approach to reducing the adverse effects of broken appointments in primarycare systems: development of a decision rule based on estimated conditional probabilities. Med Care977; 15: 419–429

[13] A. E. Williamson, D. A. Ellis, P. Wilson, R. McQueenie, and A. McConnachie, "Understanding repeated nonattendance in health services: a pilot analysis ofadministrative data and full study protocol for a nationalretrospective cohort," BMJ Open, vol. 7, no. 2, Feb. 2017

[14]K. E. Lasser, I. L. Mintzer, A. Lambert, H. Cabral, andD. H. Bor, "Missed appointment rates in primary care: theimportance of site of care," J. Health Care PoorUnderserved, vol. 16, no. 3, pp. 475–486, 2005

[15] A. Alaeddini, K. Yang, C. Reddy, and S. Yu, "A probabilistic model for predicting the probability of no-showin hospital appointments," Health Care Manag. Sci., vol. 14,no. 2, pp. 146–157, 2011.

[11]A. Wong and A. Wong, "M2's Original Shoot 'Em Up Ikusaba Set to Launch in Arcades in Fall 2020", *Siliconera*, 2020. [Online]. Available: https://www.siliconera.com/m2s-original-shoot-em-up-ikusaba-is-set-to-launch-in-arcades-in-fall-2020/. [Accessed: 04- May- 2020]

[17] R. Sabit et al., "Predictors of poor attendance at an outpatient pulmonary rehabilitation programme," Respir.Med., vol. 102, no. 6, pp. 819–824, 2008.

[18] C. Hayton et al., "Barriers to pulmonary rehabilitation: characteristics that predict patient attendance and adherence,"Respir. Med., vol. 107, no. 3, pp. 401–407, 2013.

[19] A. J. Mitchell and T. Selmes, "A comparative survey of missed initial and follow-up appointments to psychiatric specialties in the United Kingdom," Psychiatr. Serv., vol. 58,no. 6, pp. 868–871, 2007.

[20] A. J. Mitchell and T. Selmes, "Why don't patients attend their appointments? Maintaining engagement with psychiatric services," Adv. Psychiatr. Treat., vol. 13, no. 6,pp. 423–434, 2007.

[21]A. González-Rodríguez, O. Molina-Andreu, M. Imaz Gurrutxaga, R. Catalán Campos and M. Bernardo Arroyo, "A descriptive retrospective study of the treatment and outpatient service use in a clinical group of delusional disorder patients", 2020.

[22]H. Killaspy, S. Banerjee, M. King and M. Lloyd, "Prospective controlled study of psychiatric out-patient non-attendance", 2020. .

[23] D. Giunta, A. Briatore, A. Baum, D. Luna, G. Waisman,and F. G. B. de Quiros, "Factors associated with nonattendance at clinical medicine scheduled outpatient appointments in a university general hospital," Patient Prefer. Adherence, vol. 7, pp. 1163–1170, 2013.

[24]Naun.org,2020.[Online].Available:http://www.naun.org/main/NAUN/neural/2018/a022016-069.pdf. [Accessed: 04- May- 2020]