PREDICTING POOR COMPLIANCE TO PSYCHOTROPICS USING MACHINE LEARNING APPROACH

A Report Submitted
in Partial Fulfilment for the Degree
of

BACHELOR OF TECHNOLOGY

in

Department of Computer Science and Engineering

by

Dhrubang Utpal Talukdar

(CSB20049)

Under the Supervision of

Dr. Rosy Sarmah

Associate Professor

Department of Computer Science and Engineering, Tezpur University

&

Dr. Siddeswara BL

Assistant Professor

Department of Child & Adolescent Psychiatry, LGBRIMH

School of Engineering

Department of Computer Science and Engineering

Tezpur University

Napaam – 784028, Assam, India

June, 2024

Department of Computer Science and Engineering

Tezpur University

Napaam – 784028, Assam, India

DECLARATION

I hereby declare that the work presented in this report entitled "Predicting Poor Compliance to Psychotropics using Machine Learning Approach", is carried out by me. I have not submitted the matter embodied in this report for the award of any other degree or diploma of any other University or Institute. I have given due credit to the original authors/sources for all the words, ideas, diagrams, graphics, computer programs, experiments, results, that are not my original contribution. I have used quotation marks to identify verbatim sentences and given credit to the original authors/sources. I affirm that no portion of my work is plagiarized, and the experiments and results reported in the report are not manipulated. In the event of a complaint of plagiarism and the manipulation of the experiments and results, I shall be fully responsible and answerable.

Place: Name:

Date: Enrolment no:

Signature:

Department of Computer Science and Engineering

Tezpur University

Napaam – 784028, Assam, India

CERTIFICATE

This is to certify that the dissertation entitled "Predicting Poor Compliance to Psychotropics using Machine Learning Approach" submitted by Dhrubang Utpal Talukdar bearing enrolment number CSB20049 has carried out his work under my supervision and guidance for partial fulfilment of the requirements and the regulations for the award of the degree of Bachelor of Technology in Computer Science & Engineering during the session 2020 – 2024 at Tezpur University. To the best of my knowledge, the matter embodied in the dissertation has not been submitted to any other university/institute for the award of any Degree or Diploma.

Dr. Rosy Sarmah

Associate Professor

Place: Department of Computer Science & Engineering

Date: Tezpur University

Department of Child & Adolescent Psychiatry
Lokopriya Gopinath Bordoloi Regional Institute of Mental Health
Tezpur– 784001, Assam, India

CERTIFICATE

This is to certify that the dissertation entitled "Predicting Poor Compliance to Psychotropics using Machine Learning Approach" submitted by Dhrubang Utpal Talukdar bearing enrolment number CSB20049 has carried out his work under my supervision and guidance for partial fulfilment of the requirements and the regulations for the award of the degree of Bachelor of Technology in Computer Science & Engineering during the session 2020 – 2024 at Tezpur University. To the best of my knowledge, the matter embodied in the dissertation has not been submitted to any other university/institute for the award of any Degree or Diploma.

Dr. Siddeswara BL

Assistant Professor

Place: Department of Child & Adolescent Psychiatry

Date**:** Lokopriya Gopinath Bordoloi Regional Institute of Mental Health

Department of Computer Science and Engineering

Tezpur University

Napaam – 784028, Assam, India

CERTIFICATE

This is to certify that the dissertation entitled "Predicting Poor Compliance to Psychotropics using Machine Learning Approach" is submitted by Dhrubang Utpal Talukdar bearing enrolment number CSB20049. He has completed his project work successfully as needed for partial fulfilment of the requirements and the regulations for the award of the degree of Bachelor of Technology in Computer Science and Engineering during the session 2020-2024 at Tezpur University. To the best of my knowledge, the matter embodied in the dissertation has not been submitted to any other university/institute for the award of any Degree or Diploma.

Head of the Department

Place: Department of Computer Science & Engineering

Date**:** Tezpur University

Department of Computer Science and Engineering

Tezpur University

Napaam – 784028, Assam, India

Certificate by the Examiner

This is to certify that the report entitled "Predicting Poor Compliance to Psychotropics using Machine Learning Approach" submitted by Dhrubang Utpal Talukdar bearing enrolment number CSB20049 in partial fulfilment of the requirements for the degree of Bachelor of Technology in Computer Science & Engineering during the session 2020-2024 at Tezpur University has been examined by me and is found satisfactory for the award of the degree.

This approval does not necessarily endorse or accept every statement made, opinion expressed, or conclusion drawn as recorded in the report. It only signifies the acceptance of this report for the purpose for which it is submitted.

Place:

Date:

ACKNOWLEDGEMENT

I take this opportunity to extend my heartfelt gratitude to Dr. Rosy Sarmah, Associate Professor, Dept. of Computer Science and Engineering, Tezpur University, Dr.Siddeswara BL, Assistant Professor, Department of Child & Adolescent Psychiatry, Lokopriya Gopinath Bordoloi Regional Institute of Mental Health for giving me the opportunity to carry out my work under them and providing me ample guidance and support through the course of the project.

I would also like to thank the Head of the Department, Dept. of Computer Science and Engineering, and all the faculty members and Staff of the dept. of CSE, Tezpur University for the valuable guidance and cooperation throughout the project.

My sincere thanks and appreciation go to my friends and family members who have directly or indirectly helped me out with their abilities.

ABSTRACT

Predicting poor medicine compliance in paediatric population is crucial in managing mental health disorders. Towards this end, we focus on using machine learning (ML) approaches and have proposed an ensemble-based feature selection to reduce the number of features for classification. We perform an empirical study of 5 classification models viz. KNN, SVM, Random Forest (RF), AdaBoost and XGBoost. Our analysis using 4 evaluation metrices show that the Random Forest classifier gives the best result. We further developed 2 modified versions of the RF model (Meta RF and RF using weighted entropy) to study the effect of different approaches of RF. The first method creates an ensemble of random forests, where predictions are made based on majority votes from all the individual RFs. The second method utilizes weighted entropy as the splitting criterion within the RF algorithm. Our exhaustive experimentation shows that even though initially we observe a slight increase in evaluation scores of the modified versions of the RF model, with an increase in biasness in data the original RF performs better. We also develop an ML tool to assist clinicians in predicting compliance incorporating: (i) our own ensemble-based feature selection to reduce the number of features and (ii) RF classifier to predict medicine compliance. Our ML tool is tested on a real dataset of 470 children, <18 years of age, presenting at a tertiary care hospital in Tezpur with various mental health issues. The dataset comprises 247 features such as physiological, sociodemographic, economic, and psychotropic treatment characteristics. All the ML models have been experimented on this data (70% training set and 30% testing set) w.r.t. accuracy, weighted precision, weighted sensitivity and weighted F1 Score. We conclude from our performance evaluation that RF model with our own feature selection method gives better prediction which has also been validated by clinicians.

LIST OF TABLES

Page no.

LIST OF FIGURES

Page no.

Page no.

LIST OF ABBREVIATIONS

Random Forest – RF

Support Vector Machine – SVM

K Nearest Neighbor – KNN

Extreme Gradient Boosting - XGBoost

Adaptive Boosting – AdaBoost

Machine Learning – ML

Pearson correlation – PC

Mutual Information – MI

Explainable AI - XAI

Local Interpretable Model-agnostic Explanations – LIME

Shapley Additive Explanations - SHAP

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

1.1 Non-Compliance to psychotropic treatments in children

The World Health Organization defines medication adherence as how well a person follows agreed upon health recommendations, including taking medication and making lifestyle changes. Compliance refers to how accurately a patient follows the prescribed medication schedule and dosage [11,39]. When a patient does not adhere to their treatment, it means they are not properly managing their healthcare conditions, often ignoring prescribed medications and schedules. This issue is common today, especially among children and adolescents, with non-adherence rates between 50% and 88% [30]. Such non-compliance can lead to increased morbidity and more frequent hospitalizations.

1.2 Factors affecting compliance

The high rates of non-adherence in children can largely be attributed to their immature cognitive and emotional development. Additionally, several factors, including family background, influence compliance levels; for instance, lower-income families are generally less likely to follow prescription regimens closely. It was found that households with incomes below Rs. 10,000 and those dealing with illnesses lasting one to two years show particularly low compliance [23]. Compliance generally decreases with increasing illness severity [33]. Certain mental illnesses have lesser overall compliance, like Schizophrenia was the most common mental health condition linked to noncompliance, often due to fear of side effects [32, 34]. Lack of support, desperation, and inadequate assistance also contribute. Females typically show lower compliance than males [36], and illiterate households have lower adherence [26]. Key reasons for noncompliance include lack of knowledge, adverse drug reactions, diminishing hope for a cure, inadequate care or support [23].

1.3 Effects of low psychotropic compliance in children

Early-onset of psychiatric disorders in children often lead to severe illness progression, substance abuse, suicidal thoughts, and complications like poor performance in school or work, interpersonal conflicts, and legal issues [4,8,28]. Despite significant efforts, treatment challenges for psychiatric diseases in children remain common, and the prevalence of illness-related effects, such as mortality, is notably high [7].

1.4 Improving compliance in children

Improving compliance requires regular communication between the child and paediatrician. Paediatricians should provide simple plans, including medication schedules and side effect information [13]. It is critical that the child and the paediatrician communicate effectively. Building a trustworthy connection is essential for treatment plan adherence, and it can be accomplished by establishing a pattern of honest, open, and compassionate communication. Language barriers can hinder compliance. It was found that translating patient sheets to make them fully understandable to the patient, can enhance patients' compliance [20]. Additionally, some creative solutions like health-focused video games were found to improve children's self-care [21]. Machine learning (ML) [18] could automate the health care systems and cut costs for medical treatments. Recently published reviews discussed how effective ML solutions are to improve and predict adherence. ML can track down some strong predictors, which largely influence the predictions [19].

1.5 Machine Learning

As a branch of artificial intelligence, machine learning looks for patterns in big datasets to draw insightful conclusions, usually for forecasting [16,17]. Reinforcement learning, supervised learning, and unsupervised learning are its three primary subtypes [15]. Despite the fact that the idea is more than a century old [24], it has advanced more quickly in recent years due to technical developments and the accessibility of large-scale data collecting and storage [35]. Major organisations have revolutionised the industry with the development of prominent frameworks like PyTorch, Keras, and TensorFlow. Access to AI algorithms has been made easier because to cloud computing, which has made it possible to solve issues that were previously unsolvable [18]. These days, machine learning is essential for jobs like fraud detection, medical diagnosis, and picture and speech recognition.

1.6 Supervised Learning

An algorithm is trained on a dataset containing feature-target pairings in supervised learning, enabling it to forecast results based on patterns found. For instance, to estimate home pricing, a real estate company may consider features like square footage and number of stories. To confirm correctness, the model is tested on test data after learning from training data. Regression, which forecasts numerical values, and classification, which puts data points into categories, are frequent tasks. [16,17,25].

1.7 Unsupervised Learning

Unsupervised learning, which is frequently used for anomaly detection, association, and grouping, finds patterns in data without the need for predetermined labels. For example, clustering may show that a dataset's homes are grouped according to their architectural style. Finding innate structures in the data is the aim to divide it into meaningful groupings, which is the main objective of unsupervised learning [16,17,25].

1.8 Machine Learning in psychotropics

Medical treatment costs could be reduced, and health care systems automated with machine learning [18]. The effectiveness of machine learning solutions in enhancing adherence has been reviewed in recently published reviews [2]. Some strong predictors that have a significant impact on the predictions can be found using machine learning [19]. Over time, Machine Learning techniques will become more and more important in medical systems, including those that prescribe psychotropics. Understanding the importance of predicting adherence in children undergoing psychiatric medicines, we have carried out an in-depth study on compliance to psychotropics in paediatric population. To this end we delved into the two broad categories of ML: Unsupervised and Supervised.

Our work considers the significance of predicting adherence in children to their undergoing psychotropic treatments. We propose to use supervised learning to form a hypothesis to roughly predict compliance in children using K Nearest Neighbor (KNN), Support Vector Machine (SVM), Adaptive boosting (AdaBoost), Extreme Gradient Boosting (XGBoost) and RF. We evaluated the performance of various algorithms in terms of accuracy, weighted precision, weighted sensitivity, and weighted F1 score. Additionally, we adjusted several parameters for each algorithm to assess their impact on compliance prediction. We further built two modified versions of the RF model (Meta RF and RF utilizing weighted entropy). Using the majority votes from each individual RF, an ensemble of random forests is created in the first approach, and predictions are formed from there. In the second approach, the RF algorithm's splitting criteria is weighted entropy. We analysed the effect of parameter variations on accuracy for these methods. The algorithm that delivered the best results was selected for developing our app tool and web tool, aimed at predicting compliance of children in advance.

CHAPTER 2: LITERATURE REVIEW

The need for using strong predictors has been highlighted using machine learning algorithms to predict adherence, which will have a significant impact on our findings [6]. Additionally, suggestions have been made for applying machine learning to anticipate low moods or feeling like depressions in patients [29]. Patients with depression were seen to be taking their medications as prescribed, both now and prospectively, using interactive voice response tests [29]. 208 patient ratings of their interactive voice responses provided the study's data. Projections were then made using these evaluations. Age, baseline physical functioning, and prior medication adherence were among the characteristics that improved the efficacy of the mode. Logistic regression was the sole technique employed to forecast future medication adherence. 53 schizophrenia patients' medication adherence was tracked by the authors using facial recognition and computer vision software [9]. The research team received the drug intake data that the study participants had taken with the cameras on their smartphones. Subsequently, algorithms for facial recognition and computer vision were employed to warn patients who exhibited questionable behaviour, suggesting a poor likelihood of long-term compliance. Moreover, this approach served as a reminder to patients to take their drugs at a certain time every day. The primary source of bias in the study was the patients' ability to select between in-person observation and app-based monitoring. Four unique models were created to predict a target result at various stages of a clinical trial utilizing the adherence classification [38]. The daily values of adherence, adjusted adherence, number of interventions, dosage delay, and dose duration were employed as dynamic characteristics in the model at different intervals. In all model types, the remaining features—condition, trial length, and micro reimbursements—performed as static predictors. Based on the designated intervals, sub-datasets containing daily and common properties were generated. Based on dynamic and static features from patient videos that were shot on cell phones, the XGBoost classifier was used to predict medication compliance [38]. There is a dearth of research on the variables influencing children's treatment compliance. The gap in drug treatment compliance that occurs for children alone has not been examined in any studies. To close this gap, our research will discover the critical variables affecting paediatric populations' drug adherence, which will enhance children's treatment regimens by predicting the compliance in advance.

CHAPTER 3: BACKGROUND

3.1 Feature Selection

The rapid advancement of modern technology has resulted in the production of unprecedented volumes of data from a wide range of sources, including text, speech, video, photos, social media, the Internet of Things, and cloud computing. With this high-dimensional data, analysis and decision-making are challenging. Feature selection, which has been demonstrated to enhance learning efficiency, reduce data processing volume, simplify learning objectives, speed up learning, and boost accuracy, extracts a relevant subset of characteristics from the original dataset [5, 31]. There are three methods for selecting features: unsupervised feature learning, supervised feature learning, and semi-supervised feature learning. In our work, we derive the feature ranking by averaging the ranks obtained by RF, Pearson correlation (PC), and Mutual Information (MI).

3.1.1 Supervised Feature Learning

Particularly for classification tasks, supervised feature learning concentrates on the significance or correlation between features and class labels. By applying significant metrics, it seeks to identify a subset of features that optimises model accuracy [1]. RF and SVM are two popular techniques.

Random Forest (RF): Renowned for its efficient feature selection, RF is a potent supervised learning method [15]. By examining each feature's effect on the predicted accuracy of several decision trees, it assesses each feature's significance and assigns importance scores correspondingly. By consistently identifying important predictors across tree ensembles, this strategy improves dimensionality reduction and is useful for enhancing both the interpretability and performance of the model.

3.1.2 Unsupervised Feature Learning

To improve clustering accuracy, unsupervised feature learning chooses feature subsets according to assessment criteria or clustering. It uses unsupervised filter and wrapper techniques to cut down on the complexity and time required for clustering [1]. Methods such as PC and PCA are commonly employed.

Pearson Correlation (PC): With coefficients ranging from -1 to 1, PC quantifies the linear link between two variables and indicates both its direction and intensity [15]. This statistical method is crucial for feature selection in data analysis because it works well for determining linear correlations and assessing feature redundancy. The following formula is used to get the PC coefficient:

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} \quad (3.1)$$

where :

$\rho_{X,Y}$ is the Pearson correlation coefficient between X and Y

$cov(X,Y)$ is the covariance of X and Y

σ_X is the standard deviation of X

σ_Y is the standard deviation of Y

3.1.3 Semi-Supervised Feature Learning

To enhance learning performance, semi-supervised feature learning makes use of a dataset that includes both labelled and unlabelled samples. To choose pertinent characteristics, it makes use of filter models and score functions like the Laplacian and Fisher scores. Label propagation and MI are examples of methods.

Mutual Information: MI measures how much information is shared and how dependent two random variables are on one another. Since strong reliance is indicated by high MI values, MI is especially useful for capturing non-linear interactions. For informative feature selection in a variety of machine learning applications, this makes it a useful tool. The following formula is used to calculate MI:

$$I(X;Y) = H(X) + H(Y) - H(X,Y) \quad (3.2)$$

where:

$I(X;Y)$ Is the mutual information between X and Y

$H(X)$ Is the entropy of X

$H(Y)$ Is the entropy of Y

$H(X,Y)$ Is the joint entropy between X and Y

3.2 Classification

Sorting data into predefined classes or categories is the fundamental step in the machine learning process known as classification. It is crucial in a wide range of industries, including image recognition and email spam detection. Classification algorithms foresee cases that were previously unknown and classify them into new groups based on patterns found in labelled data. These algorithms encompass a broad spectrum of methods, such

as decision trees, SVM, and neural networks, each with unique benefits and applications. Successful classification models accurately assign class labels to incoming data, allowing for automation and informed decision-making in a range of real-world scenarios. In our project we have incorporated classification algorithms like RF [15], AdaBoost [37], XGBoost [10], SVM [12] and KNN [14].

Random Forest: RF is an ensemble learning technique that generates the mean prediction (regression) or mode of the classes (classification) for each individual decision tree by using many decision trees that are generated during training. It has a well-deserved reputation for being resilient and capable of handling large, highly dimensional datasets while averaging to minimise overfitting.

AdaBoost: Adaptive Boosting, or AdaBoost, is an ensemble approach that builds a powerful classifier by combining weak classifiers. It concentrates more on difficult-to-classify cases in later iterations by successively modifying the weights of misclassified instances, which enhances the accuracy and performance of the model overall.

XGBoost: A gradient boosting system that is tuned for speed and performance is called XGBoost (Extreme Gradient Boosting). It is a well-liked option for structured or tabular data in machine learning contests and applications because it effectively handles sparse data and employs sophisticated regularization algorithms to minimize overfitting.

SVM: Encouragement For tasks involving regression and classification, supervised learning models called vector machines are employed. To achieve high accuracy in high-dimensional spaces, they maximize the gap between the nearest data points of each class by identifying the ideal hyperplane that optimally divides the data into multiple classes.

KNN: A straightforward, non-parametric method for regression and classification is K-Nearest Neighbors. It is simple to use and efficient for small to medium-sized datasets since it uses distance metrics to detect similarity and assigns a class to a data point based on the majority class among its k-nearest neighbours.

3.3 Confusion Matrix

An essential tool for assessing how well categorization models work is a confusion matrix [15]. It provides a thorough examination of model accuracy by classifying predictions into four categories: True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). Key performance metrics derived from these values include sensitivity, precision, accuracy, and F1 Score [15].

The confusion matrix makes it easier to analyse the model's performance in-depth and points out areas that want improvement. It offers a lucid representation that facilitates comprehension and judgment in a range of categorization assignments.

True Positive: True Positives are instances correctly predicted as positive by the model. For example, correctly identified spam emails by a spam filter are TPs.

True Negative: True Negatives are instances accurately identified as negative by the model. In fraud detection, TNs represent transactions correctly classified as non-fraudulent.

False Positive: False Positives occur when the model incorrectly predicts a negative instance as positive, also known as a Type I error. For instance, a medical test falsely indicating the presence of a disease is an FP.

False Negative: False Negatives happen when the model incorrectly predicts a positive instance as negative, referred to as a Type II error. For example, a security system failing to detect a breach is an FN.

Sensitivity: Sensitivity measures the proportion of actual positives correctly identified by the model.

Precision: Precision assesses the accuracy of the model in predicting positive outcomes.

Accuracy: Accuracy evaluates the overall correctness of the model across all classes.

F1 Score: The F1 Score balances sensitivity and precision, especially useful in scenarios with imbalanced class distributions.

CHAPTER 4: METHODOLOGY

The flowchart depicted in Figure 4.1 illustrates the comprehensive process involved in data science and machine learning workflow. The process initiates with Data Collection, which is the foundational step of gathering relevant data for analysis. This is followed by Data Preprocessing, where data is cleaned by handling missing values and outliers and converting the data into a numeric form suitable for analysis. The next step, Feature Engineering and Ranking, involves creating new features or modifying existing ones, followed by ranking these features using appropriate methods to ensure the most significant attributes are utilized. Subsequently, Model Training and Selection is carried out, where various machine learning models are compared, and the best-performing model is selected. The chosen model undergoes Model Evaluation using appropriate metrics to assess its performance. Finally, the process culminates in App / Webtool Building, where the evaluated model is integrated into a practical application or web tool for end-user deployment. This systematic approach ensures a structured and efficient workflow from data acquisition to application development.

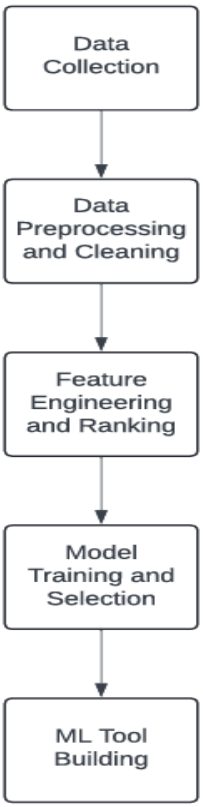


Fig 4.1 Basic Framework for Methodology

4.1 Data Collection

The data was collected from Lokopriya Gopinath Bordoloi Institute of Mental Health, Tezpur, Assam, India. The dataset has 248 attributes and initially had about 150 rows in total. The data includes personal details like child's name, age, gender, district of residence, family income, referral, the chief complaint, and information about their birth, such as weight, height, and head circumference. It includes medical information like information about their family's medical history, past treatments, medication, the dose limit, the total cost, and the doctor's diagnosis. There are also measures of how well they followed their treatment plan, like mean gap ratio, medication possession ratio, total follow ups, the frequency of follow ups, maximum compliance period and total duration of medication. There is a final compliance column telling us how well the child followed their treatment plan containing values – Good, Satisfactory and Poor. The final dataset comprises of 470 patient data with 248 attributes.

4.2 Data Preprocessing and Cleaning

To prepare the dataset for analysis, several preprocessing steps were undertaken. Initially, data attributes that primarily contained null or no values were eliminated. Attributes such as Name and Unique ID, which were not significantly relevant to the final compliance column, were also removed. For numerical attributes with missing values, these gaps were filled with zeros to ensure completeness. Textual attributes were converted into numerical values through a mapping process. This involved identifying unique names within each attribute and assigning each a specific value based on a separately created dictionary for each column. This approach was designed to address inconsistencies within attributes and effectively handle missing values, ensuring a robust and uniform dataset for subsequent analysis.

4.3 Feature Engineering

In our study, we expanded the scope of our dataset by introducing 14 new features, each representing distinct mental health states prevalent among individuals. These features were meticulously derived from the most frequently diagnosed mental illnesses, encompassing attributes such as irritability, anger, abnormal behaviour, restlessness, jerky movements, mood fluctuations, outbursts, eye contact patterns, cognitive impairments, anxiety levels, hallucinatory experiences, depressive symptoms, seizures, and attention-deficit/hyperactivity disorder (ADHD) tendencies.

To construct these features, we scrutinized the data's ten Axis Columns, which catalogued various medical diagnoses. If any diagnosis matching those encapsulated in our newly created feature columns was present in the Axis Columns, the corresponding feature was assigned a value of 1; otherwise, it remained 0. This method allowed us to effectively encode the presence or absence of specific mental health indicators within our dataset.

By focusing on these carefully selected mental health attributes, our goal was to capture a comprehensive spectrum of relevant psychological states prevalent in our sample population. Following the incorporation of these additional features and a rigorous process of attribute refinement, our dataset was distilled to a streamlined collection of 73 attributes. This meticulous creation of new features ensures a focused and efficient framework for subsequent analyses, ultimately enhancing the relevance and accuracy of our study's findings.

4.3.1 Proposed Method: Feature Ranking using average of RF, PC, and MI

Definition 1: Full feature set (F) is the set of all features in a dataset D, where $F=\{f1,f2,...,fn\}$, f_i denotes the i^{th} feature and n is the number of features.

Definition 2: Rank of a feature $f_i \in F$ (denoted as $R_{m_j}(i)$) w.r.t a feature selection model m_j is defined as the index of f_i in the sorted list of feature importance values R_{m_j}, obtained from the feature selection model m_j (where $j=1,2,3$). In our study, we have used j = 3 feature selection models viz. PC, MI and RF

Definition 3: Overall Rank of a feature f_i is the average of the ranks of f_i for each model m_j $(j=1,2,3)$

This facilitated the identification of features with the most significant impact on compliance. Subsequently, the aggregated ranks were sorted from lowest to highest. Notably, features with the lowest ranks were deemed to possess the most substantial compliance impact compared to their higher-ranked counterparts. Table 4.1 presents the top 32 features, of which we opted to select the top 30 based on our findings of optimal performance. Other characteristics beyond this selection were disregarded in our analysis.

Rank	Feature Name
1	Maximum period of compliance at LGB (in days) (longest streak of good compliance)
2	Total number of follow up at LGBRMIMH
3	Maximum duration of symptom free period (in days)
4	Days/freq
5	Total duration of medication treatment at LGB (in days) (from first consultation to last follow-up)
6	Off-medications duration (to add all such durations over follow-up in days)
7	Total duration of medication 2 (in days)
8	Mean gap ratio at LGB (total no of months of follow-up divided by no of follow-ups)
9	Medication possession ratios 1 (MPRs) in LGB
10	Total duration of medication 1 (in days)
11	Age at presentation (in yrs)
12	Continued medication 2/stopped/changed
13	School Adjustment
14	Total_days1
15	No of relapses/exacerbations
16	Max Duration of resolution of symptoms before recurrence/relapse (in days)
17	Age at onset (in years)
18	Maximum dose of medication 1 (in mg)
19	Rural/Urban
20	Age at last follow up
21	Medication possession ratios 2 (MPRs) in LGB
22	Outbursts
23	Continued medication 1/stopped/changed
24	Distance from LGBRMH (in KM)
25	Follow up diagnosis changed or not (yes/no)
26	Family environment
27	Weight (in Kg)
28	Time period between onset to first consultation at LGBRMIMH (DUI) (in days)
29	Adhd
30	Response to medication 1 (Good/partial/no)
31	Systematic Examination (abnormal/normal)
32	Response to medication 2 (Good/partial/no)

Table 4.1 Feature Ranking using overall rank from ranks using RF, PC and MI

4.3.2 Feature Ranking using RF

At each decision tree node, the effectiveness of each feature in helping to divide the data into more homogeneous groups is used to define the feature relevance in RFs. The method determines which feature gives the highest increase in purity or decrease in impurity of the subsequent leaf nodes for each tree in the forest at each split point. This impurity reduction metric, which is frequently computed using entropy, shows how well a feature divides the data into discrete classes or groups. Next, the significance of every feature of a single tree is totalled for all the trees in the forest. By averaging the importance scores determined by each individual tree, this aggregation can be accomplished. To provide a relative ranking of feature importance within the RF, these scores are finally normalised so that the sum of all feature importance scores equals 1. RFs can rank features according to their overall contribution to the prediction performance of the model by looking at how frequently and successfully each feature is employed across all trees in the ensemble [40]. Table 4.2 shows the top 32 features for feature ranking obtained with the help of RF.

Rank	Feature	Random Forest Importance
1	Off-medications duration (to add all such durations over follow-up in days)	0.1411981135413449
2	Maximum duration of symptom free period (in days)	0.12431624810669437
3	Maximum period of compliance at lgb (in days) (longest streak of good compliance)	0.06602369826781676
4	Total frequency	0.0517838366690728
5	Total duration of medication treatment at LGB (in days) (from first consultation to last follow-up)	0.05125419746268719
6	Mean gap ratio at lgb (total no of months of follow-up divided by no of follow-ups)	0.03678817275838798
7	Days/freq	0.03415253264063216
8	Continued medication 2/stopped/changed	0.03159955047408112
9	Total number of follow up at LGBRIMH	0.02641601904258999
10	Age at presentation (in yrs)	0.026211943197286612
11	Medication possession ratios 1 (MPRs) in lgb;x-syrup (total number of days when medications were taken divided by summation of total number of days when medications were taken with total off medication period)	0.025584796768803743
12	Distance from LGBRIMH (in KM)	0.024753239692615146
13	Total duration of medication 2 (in days)	0.02316438040148189
14	Age at last follow up	0.02238922928261885
15	Time period between onset to first consultation at LGBRIMH (DUI) (in days)	0.019199562921172093
16	No of relapses/exacerbations	0.018437110075078752
17	Total duration of medication 1 (in days)	0.017218406048751512
18	Max Duration of resolution of symptoms before recurrence/relapse (in days)	0.01698400239089499
19	Response to medication 2 (Good/partial/no)	0.016617095907808 32
20	weight (in Kg)	0.015265144841052245
21	Maximum dose of medication 1 (in mg)	0.014714196464566736
22	$\text{total}_{d}ays1$	0.014325717648138194
23	Age at onset (in years)	0.014314377827724583
24	Continued medication 1/stopped/changed	0.011645831255609895
25	Avg dose of medication 1 (Mode value of medication) (in mg)	0.011516864505050956
26	Response to medication 1 (Good/partial/no)	0.009850595733713495
27	School Adjustment	0.009822968315119293
28	Systemic examination (abnormal/normal)	0.008262147454874743
29	"If yes, after how many days from first presentation diagnosis changed (in days)"	0.00813615598242771
30	Past/Current medical conditions	0.006625903020411289
31	Religion	0.006085997928024703
32	significant psychosocial stressor	0.005964651005825684

Table 4.2 Feature Ranking with rank scores from results obtained using RF

4.3.3. Feature Ranking using PC

The features are ranked with respect to the target column using the formula (3.1). The values are then compared on basis of magnitude, where features with higher magnitude values get higher rank compared to features with lower magnitude values. Table 4.3 shows the top 32 features for feature ranking obtained with the help of PC.

Rank	Feature	Pearson Correlation
1	Maximum period of compliance at lgb (in days) (longest streak of good compliance)	0.29174415501789064
2	Maximum duration of symptom free period (in days)	0.28907008188694683
3	Total number of follow up at LGBRIMH	0.2838694431557081
4	Total frequency	0.27691896205943134
5	Max Duration of resolution of symptoms before recurrence/relapse (in days)	0.27608481799631185
6	Days/freq	0.23612793017593 04
7	Total duration of medication treatment at LGB (in days) (from first consultation to last follow-up)	0.20256315471696631
8	Total duration of medication 1 (in days)	0.19581960305468 65
9	Continued medication 2/stopped/changed	0.19396400471150815
10	Depression	0.1542450475388608
11	School Adjustment	0.13049360904668825
12	Outbursts	0.11869251807472944
13	Total duration of medication 2 (in days)	0.11381984234593333
14	Follow up diagnosis changed or not (yes/no)	0.10688142756902 67
15	Rural/Urban	0.09983022396532 85
16	Type of Family (Nuclear/Joint/single parent/orphan/foster family)	0.09508094653995165
17	Distance from LGBRIMH (in KM)	0.09080982607570 03
18	Age at onset (in years)	0.09072773769071 43
19	Age at presentation (in yrs)	0.09032203469111426
20	Continued medication 1/stopped/changed	0.08877473718585199
21	Low	0.08550728433105109
22	Off-medications duration (to add all such durations over follow-up in days)	0.08400450606981842
23	ADHD	0.08369454941149343
24	Number of In patient cares	0.08245465921621781
25	Past/Current medical conditions	0.08213997055454715
26	Medication possession ratios 1 (MPRs) in lgb	0.08200574009693978
27	Mean gap ratio at lgb (total no of months of follow-up divided by no of follow-ups)	0.07968451118414194
28	Systemic examination (abnormal/normal)	0.07468509004981613
29	Medication possession ratios 2 (MPRs) in lgb	0.07262398522397875
30	Total_days1	0.0720072875979988
31	Sex (m/f)	0.07106173422413431
32	No of relapses/exacerbations	0.07042271192097925

Table 4.3 Feature Ranking with rank scores from results obtained using PC

4.3.4 Feature Ranking using MI

The features are ranked with respect to the target column using the formula (3.2). The values are then compared on basis of magnitude, where features with higher magnitude values get higher rank compared to features with lower magnitude values. Table 4.4 shows the top 32 features for the feature ranking obtained with the help of MI.

Rank	Feature	Mutual Information
1	Off-medications duration (to add all such durations over follow-up in days)	0.15199382977124554
2	Total duration of medication 2 (in days)	0.11418086360904578
3	Outbursts	0.10202544215629405
4	Rural/Urban	0.08323417065300065
5	Eye contact	0.0828264874628033
6	Medication possession ratios 2 (MPRs) in lgb; x-syrup	0.07528851329080899
7	Maximum period of compliance at lgb (in days) (longest streak of good compliance)	0.07144738359691871
8	Age at last follow up	0.0709626816106268
9	Response to medication 1 (Good/partial/no)	0.07045454013349906
10	Total duration of medication treatment at LGB (in days) (from first consultation to last follow-up)	0.06466487948528066
11	Depression	0.06423525439763456
12	Age at presentation (in yrs)	0.05947054175456934
13	Cost of medication	0.05664344733033788
14	Total number of follow up at LGBRIMH	0.0563705399636647
15	Total frequency	0.05544669369186739
16	Restlessness	0.05531541675843599
17	Seizure	0.05377922764803511
18	Religion	0.050984523113708935
19	Continued medication 1/stopped/changed	0.04814664733310181
20	Weight (in Kg)	0.04482567710446972
21	Anxious	0.032978642280470094
22	Mean gap ratio at lgb (total no of months of follow-up divided by no of follow-ups)	0.030376647463925543
23	Total_days1	0.029392228730604764
24	Low	0.02604719712880188
25	Max Duration of resolution of symptoms before re-currence/relapse (in days)	0.023676729501421967
26	Maximum dose of medication 1 (in mg)	0.02275437750387077
27	Days/freq	0.022019259183145845
28	No of relapses/exacerbations	0.017687669855306343
29	Number of In patient cares	0.017441071976846967
30	School Adjustment	0.01642291955327524
31	Total duration of medication 1 (in days)	0.013806029062825687
32	Time period between onset to first consultation at LGBRIMH (DUI) (in days)	0.013653748847653358

Table 4.4 Feature Ranking with rank scores from results obtained using MI

4.4 Model Training

4.4.1 Model Training (Initial Real Data)

We conducted a comparative analysis of five prominent Supervised Machine Learning algorithms based on their accuracy metrics. The selected algorithms encompass AdaBoost, XGBoost, SVM, KNN, and RF. Notably, for the RF algorithm, we introduced a novel approach utilizing weighted entropy as a splitting criterion. Additionally, we employed an ensemble technique involving RF (meta-RF) to leverage majority voting for predictions.

Our study delved into an extensive parameter tuning exercise for each algorithm, exploring the impact of varying parameters on their respective accuracies. Through

systematic experimentation, we plotted graphs to visualize the performance trends and identify parameter configurations yielding the highest accuracies. Subsequently, for these optimal parameter values, we further investigated the influence of varying the number of features on accuracy metrics.

By meticulously analysing these parameters and feature variations, we aimed to pinpoint the optimal settings for each algorithm, thus facilitating informed decision-making in real-world applications. Initially, all tests were conducted on data collected from 150 individuals.

Proposed Method: Random Forest (using weighted entropy as a splitting criterion)

In this study, we propose a novel approach to enhance the performance of the RF classifier by incorporating a custom weighted entropy criterion. The primary motivation behind this method is to account for class imbalance and improve classification accuracy across different classes. We can also give more importance to classes. We utilize a custom information gain calculation that integrates class-specific weights into the entropy computation.

Weighted Entropy Calculation

To generalize the weighted entropy method for any number of classes *n*, let us define the entropy *H* for a dataset *S* with *n* classes as follows:

$$H(S) = -\sum_{i=1}^{n} p_i \log_2(p_i) \quad (4.1)$$

where P_i is the proportion of instances belonging to class *i*

In the weighted entropy method, each class *i* is assigned a weight w_i. The weighted entropy H_w for a dataset *S* is then calculated as:

$$H_w(S) = -\sum_{i=1}^{n} w_i p_i \log_2(p_i) \quad (4.2)$$

where w_i represents weight of class *i*

Information Gain from Weighted Entropy

The information gain *IG* for an attribute *A* with weighted entropy is calculated as follows. Let *S* be split into subsets S_v based on the values of attribute *A:*

$$IG_w(S, A) = H(S) - \sum_{v \subset values(A)} \frac{|S_v|}{|S|} H_w(S_v) \quad (4.3)$$

where $|S_v|$ is the number of instances in subset

S_v, and $H_w(S_v)$ is the weighted entropy of subset S_v

Class Weight Adjustment

In our proposed method, we dynamically adjust the class weights to handle class imbalance using a parameter termed as weight. Let n represent the total number of classes. The adjustment of weights is achieved through the following formula: the weight for the first class is set to 1/n + weight, while the weight for the last class is 1/n − weight. For the intermediate classes, their weights gradually decrease based on their position in the sequence, following the pattern of 1/n − i−1/n−1 · weight, where *i* denotes the position of the class. This approach ensures that the weights are appropriately adjusted to address the imbalance, with greater emphasis given to the minority classes while maintaining balance across the dataset. In our proposed method, the class weights are dynamically adjusted to handle class imbalance. Given the class weights *Wi* for each class *i*, we introduce a parameter weight that adjusts these weights. For instance, if there are three classes 'poor', 'satisfactory', and 'good', the weights can be adjusted as follows:

$$W_{poor} = \frac{1}{3} + \text{weight}$$
$$W_{satisfactory} = \frac{1}{3} \qquad (4.4)$$
$$W_{good} = \frac{1}{3} - \text{weight}$$

This ensures that the weight for the 'poor' class is increased by weight, while the weight for the 'good' class is decreased by weight, maintaining the weight for the 'satisfactory' class constant.

Example Calculation

Consider a dataset *S* with three classes 'poor', 'satisfactory', and 'good'. The weighted entropy $H_w(S)$ is calculated as:

$$H_w(S) = -\ (W_{poor}\ p_{poor}\ log_2(p_{poo}r) + W_{satisfactory}\ p_{satisfactory}\ log_2(p_{satisfactory}) + W_{good}\ p_{good}\ log_2(p_{good}) \qquad (4.5)$$

Meta Ensemble Random Forest

In this method we use an ensemble method of n Random Forests for prediction. The majority vote of all the n Random Forest is returned as a prediction. In case of a tie the class which appears first is returned. We may later delve into the Explainable AI methods to decode the black box of the RF and know about the features considered more important.

Explainable AI

Explainable AI (XAI) [22] aims to make the decisions of machine learning models more interpretable to humans. Explainable AI (XAI) addresses the challenge posed by black-box AI models, providing insights into how AI systems arrive at their decisions. This transparency enhances trust, accountability, and regulatory compliance. Explainable AI (XAI) techniques include model inspection, feature importance analysis, and explanation methods. These approaches enable users to understand the factors influencing AI model predictions and decisions. Two prominent methods in this domain are LIME and SHAP.

LIME: Local Interpretable Model-agnostic Explanations (LIME) [3] is a technique in Explainable AI (XAI) designed to elucidate how complex machine learning models make predictions. LIME creates simplified, interpretable models around specific instances by slightly tweaking the input data and observing changes in the model's predictions. This process reveals the decision-making logic that mimics the behavior of the black-box model for that example, highlighting which features the model considers important. In essence, LIME helps make the 'black box' of AI more transparent and understandable to humans.

SHAP: SHAP (Shapley Additive Explanations) [27] is a powerful tool in Explainable AI (XAI) that provides insights into the decision-making processes of machine learning models. By considering all possible combinations of features and their impact on predictions, SHAP identifies which features have the most influence on model outcomes. This makes the inner workings of black-box models more accessible, enabling users to make informed decisions and validate model behaviour effectively.

4.4.2 Model Training (using Final Real Data)

We employed the RF algorithm to conduct various tests on the dataset. Numerous parameters were adjusted to achieve the optimal accuracy, weighted precision,

weighted sensitivity and weighted F1 score. Graphs were plotted to examine the impact of different combinations of features and parameters. These tests were performed on the final dataset, which comprised data from 470 children.

Specifically, we introduced a new approach for the RF algorithm by incorporating weighted entropy as the splitting criterion. Additionally, we applied an ensemble method (Meta-RF) that uses majority voting to enhance the prediction accuracy of the RF model.

4.5 Model Selection

The model selected for developing our app and web tool is the RF model (using entropy as a splitting criterion), which was trained on the final real data, utilizing the number of decision trees = 75, incorporating the top 30 features as identified in Table 4.1. This configuration, with the specified parameters and features, yielded the optimal performance in terms of accuracy, precision, sensitivity, and F1 score.

4.6 App Tool Building

4.6.1 Requirements

i) Language Used: Python

ii) Package Used: Pandas, NumPy, Scikit-learn, Joblib, Kivy

iii)Software Used: VSCode

4.6.2 App Tool Building

We have developed our application using the Kivy package in Python, incorporating a RF model that employs entropy as the splitting criterion. This model utilizes 75 decision trees and incorporates the top 30 features as identified in Table 4.1. The model was trained in Jupyter Notebook and subsequently serialized as a .pkl (pickle) file for integration into our application.

The app presents users with a questionnaire designed to capture the values of the top 30 features used during the model training. To ensure data integrity, robust error-checking mechanisms have been implemented, preventing users from submitting incomplete responses or invalid entries.

Upon completing the questionnaire, users can click the "Predict" button, which triggers the model to process the input data. The application converts the user-provided responses into the appropriate numeric format and passes them to the pre-trained model stored in

the pickle file. The model then predicts the patient's compliance status, which is subsequently displayed on the screen.

This structured approach ensures that users receive accurate predictions based on their input, leveraging the power of machine learning to enhance the application's functionality. A few snapshots of the app are shown in Figure 4.1

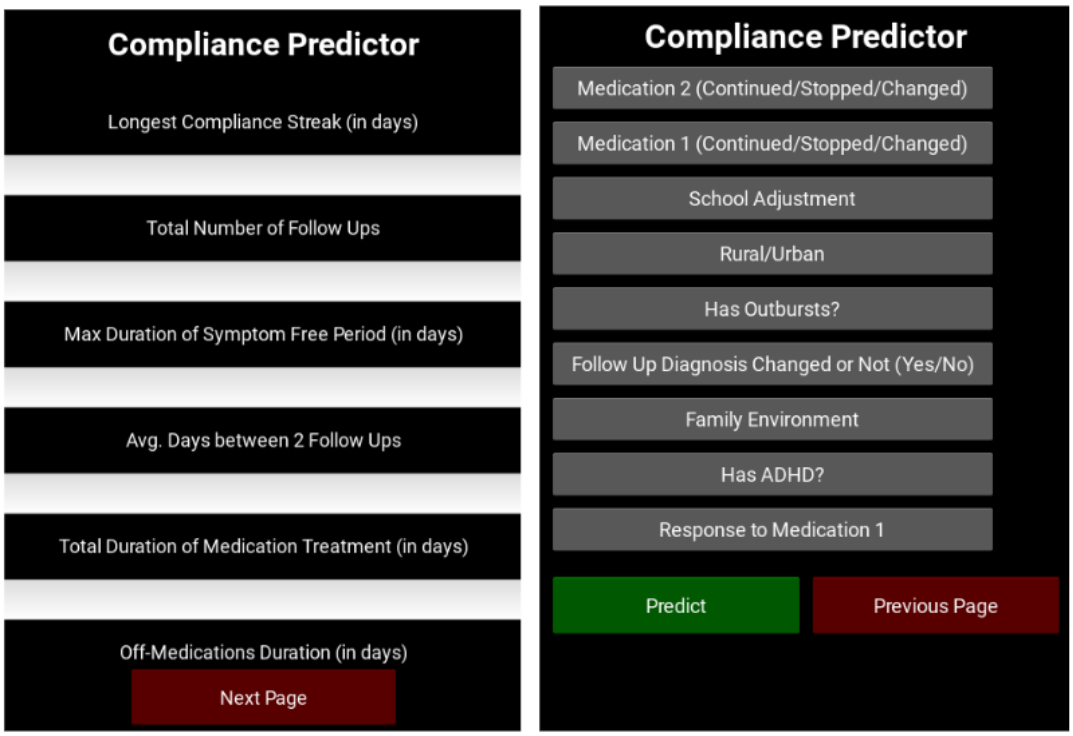


Fig 4.1 Snapshots of our App Tool

4.7 Web Tool Building

4.7.1 Requirements

i) Language Used: Python, HTML, CSS

ii) Package Used: NumPy, Scikit-learn, Pickle, Flask, Pandas

iii)OS used: Windows 10

iv)Software Used: Jupyter Notebook, VSCode

4.7.2 Web Tool Building

We have developed our web tool using the Flask package in Python, incorporating a RF model that employs entropy as the splitting criterion. This model utilizes 75 decision trees and incorporates the top 30 features as identified in Table 4.1. The model was trained in

Jupyter Notebook and subsequently serialized as a .pkl (pickle) file for integration into our application.

The app presents users with a questionnaire designed to capture the values of the top 30 features used during the model training. To ensure data integrity, robust error-checking mechanisms have been implemented, preventing users from submitting incomplete responses or invalid entries.

Upon completing the questionnaire, users can click the "Predict" button, which triggers the model to process the input data. The application converts the user-provided responses into the appropriate numeric format and passes them to the pre-trained model stored in the pickle file. The model then predicts the patient's compliance status, which is subsequently displayed on the screen.

This structured approach ensures that users receive accurate predictions based on their input, leveraging the power of machine learning to enhance the application's functionality. A few snapshots of the web tool are shown in Figure 4.2

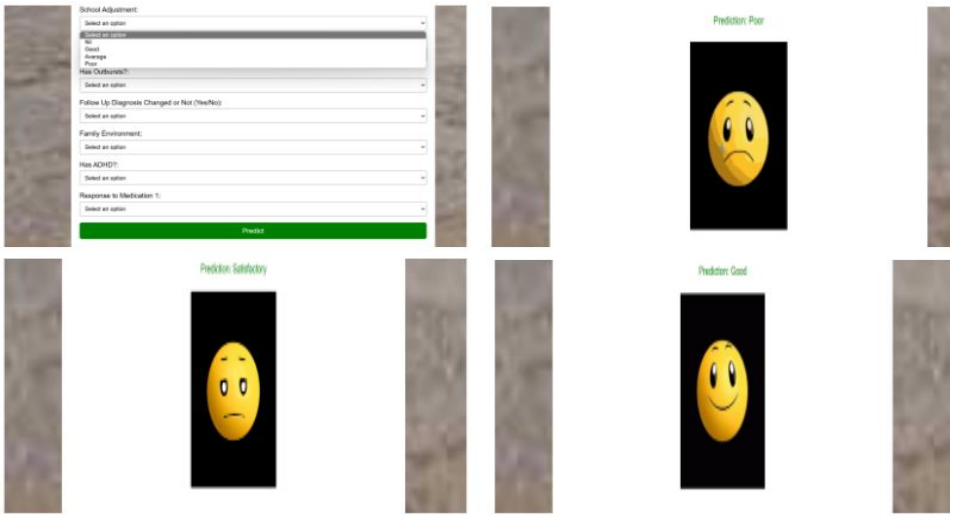


Fig 4.2 Snapshots of our Web Tool

CHAPTER 5: RESULTS AND DISCUSSION

5.1 Initial Real Data

The data was collected from Lokopriya Gopinath Bordoloi Institute of Mental Health, Tezpur, Assam, India. The dataset has 248 attributes and initially had about 150 rows in total.

5.1.1 Feature Ranking Comparison

The feature ranking was conducted using four distinct methods: RF, PC and MI and our proposed method of using ensemble of RF, PC and MI ranking features based on the average rank. The results obtained from these methods are presented in Tables 4.2, 4.3, 4.4 and 4.1, respectively.

To assess the effectiveness of our proposed ensemble method, we compared it with the individual methods. For this evaluation, we employed the RF algorithm with entropy as the splitting criterion. We set the number of decision trees to 80 and considered the top 30 features for each method.

Table 5.1 summarizes the performance metrics obtained from these comparisons, including accuracy, weighted precision, weighted sensitivity, and weighted F1 Score. Notably, the ensemble-based feature ranking yielded superior results across all metrics, demonstrating its effectiveness in identifying relevant features.

Method Used	Accuracy	Weighted Precision	Weighted Sensitivity	Weighted F1 Score
Pearson Correlation	0.800	0.812	0.800	0.802
Mutual Information	0.800	0.821	0.800	0.791
Random Forest	0.766	0.763	0.766	0.745
Ensemble Method	0.867	0.867	0.867	0.867

Table 5.1 Comparison of various feature selection methods

5.1.2 Interpretation of SHAP and LIME Feature Ranking Results

LIME

We have applied LIME algorithm on RF Model (using entropy as a splitting criterion) to rank the features, where we have used value of number of decision trees = 80 and applied LIME on the whole refined dataset containing 73 features to find out the average effect of each feature on the target column. This ranking is shown in Table 5.2.

Rank	Feature Name	Average Probabilities
1	jerky	0.004883
2	School Adjustment	0.004398
3	total frequency	0.004219
4	total number of follow up at LGBRIMH	0.004089
5	systemic examination (abnormal/normal)	0.003870
6	eye contact	0.003600
7	Follow up diagnosis changed or not (yes/no)	0.002786
8	depressed	0.002619
9	Mental status examination/Behavioral Observations	0.001802
10	significant psychosocial stressor	0.001609
11	Distance from LGBRIMH (in KM)	0.001529
12	Max Duration of resolution of symptoms before recurrence/relapse (in days)	0.001496
13	Continued medication 1/stopped/changed	0.001306
14	Socioeconomic status	0.001198
15	Age at last follow up	0.001066
16	mean gap ratio at lgb (total no of months of follow-up divided by no of follow-ups)	0.000968
17	Rural/Urban	0.000760
18	Medication possession ratios 1(MPRs) in lgb;x-syrup	0.000605
19	maximum period of compliance at lgb (in days) (longest streak of good compliance)	0.000489
20	Religion	0.000389
21	irritable	0.000373
22	Avg dose of medication 1 (Mode value of medication)	0.000145
23	seizure	0.000120
24	Time period between onset to first consultation	-0.000072
25	Academic performance	-0.000141
26	Age at presentation (in yrs)	-0.000170
27	Number of In patient cares	-0.000274
28	Total duration of medication 2 (in days)	-0.000320
29	poor	-0.000371
30	Response to medication 2 (Good/partial/no)	-0.000431
31	Medication possession ratios 2(MPRs) in lgb	-0.000671
32	Age at onset(in years)	-0.000733

Table 5.2 Feature Ranking using LIME

SHAP

We have applied SHAP algorithm on RF Model (using entropy as a splitting criterion) to rank the features, where we have used value of number of decision trees = 80 and applied SHAP on the whole refined dataset containing 73 features to find out the average effect of each feature on the target column. This ranking is shown in Table 5.3.

Comparing Table 5.2 which shows the feature ranking obtained from LIME, Table 5.3 which shows the feature ranking obtained from SHAP with Table 4.1 which is our ensemble-based feature ranking, we find that SHAP's global interpretability of the RF model aligns more closely with our ensemble-based feature ranking. In contrast, LIME's local interpretability results in a feature ranking that differs from our ensemble ranking, highlighting the distinction between SHAP's holistic approach and LIME's focus on individual predictions.

Rank	Label Name	Average Importance
1	Off-medications duration (to add all such durations over follow-up in days)	0.032107
2	total number of follow up at LGBRIMH	0.028937
3	maximum period of compliance at lgb (in days) (longest streak of good compliance)	0.024616
4	total_frequency	0.024439
5	days/freq	0.017878
6	Maximum duration of symptom free period (in days)	0.017711
7	Medication possession ratios 1(MPRs) in lgb	0.012903
8	School Adjustment	0.012857
9	No of relapses/exacerbations	0.012759
10	Continued medication 2/stopped/changed	0.011886
11	Max Duration of resolution of symptoms before recurrence/relapse (in days)	0.011326
12	Age at presentation (in yrs)	0.011321
13	Age at last follow up	0.010211
14	total_days1	0.009757
15	total duration of medication treatment at LGB(in days) (from first consultation to last follow-up)	0.008405
16	mean gap ratio at lgb (total no of months of follow-up divided by no of follow-ups)	0.008319
17	weight (in Kg)	0.007885
18	restless	0.007397
19	Total duration of medication 2(in days)	0.007348
20	adhd	0.007080
21	Response to medication 1 (Good/partial/no)	0.007025
22	Total duration of medication 1 (in days)	0.006820
23	Medication possession ratios 2(MPRs) in lgb;x-syrup	0.006715
24	Distance from LGBRIMH (in KM)	0.006207
25	Socioeconomic status	0.005854
26	Continued medication 1/stopped/changed	0.005535
27	Time period between onset to first consultation at LGBRIMH (DUI) (in days)	0.004776
28	Academic performance	0.004539
29	Maximum dose of medication 1 (in mg)	0.004465
30	Avg dose of medication 1 (Mode value of medication) (in mg)	0.004210
31	Age at onset(in years)	0.004056
32	Angry	0.004029

Table 5.3 Feature Ranking using SHAP

5.1.3 Effects of changing various feature subsets and parameters of various supervised ML algorithms on accuracy

We analysed five supervised ML algorithms (AdaBoost, XGBoost, SVM, KNN, and RF) and the effect on the accuracy of each algorithm on changing various parameters of each algorithm and their feature subsets.

AdaBoost

Figure 5.1 depicts the correlation between the number of decision trees and the accuracy of the AdaBoost model, leveraging the 30 top features outlined in Table 4.1. Additionally, Figure 5.2 illustrates how varying the number of top attributes, identified through our feature ranking in Table 4.1, influences the accuracy of the AdaBoost model. It's worth noting that these analyses were carried out while keeping the number of weak learner's constant at 15, a configuration that yielded the highest accuracy.

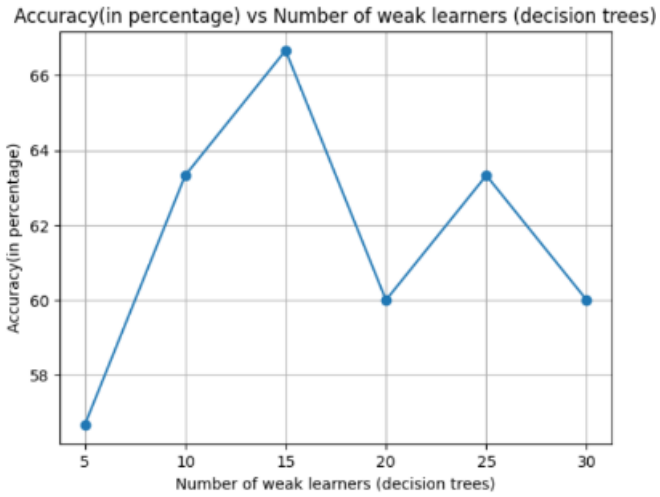


Fig 5.1 Accuracy (in percentage) vs number of weak learners (decision trees) used in AdaBoost (at constant value of 30 top features according to feature ranking obtained in Table 4.1)

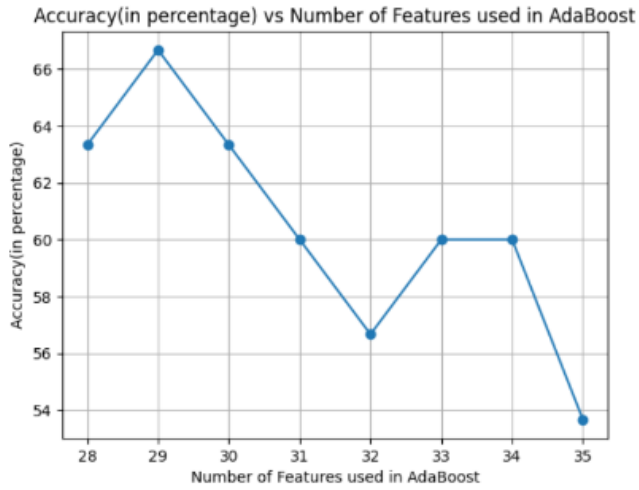


Fig 5.2 Accuracy (in percentage) vs number of top features used in AdaBoost (at constant value of 15 weak learners) according to feature ranking obtained in Table 4.1

XGBoost

Figure 5.3 depicts the correlation between the number of boosting rounds and the accuracy of the XGBoost model, leveraging the 30 top features outlined in Table 4.1. Additionally, Figure 5.4 illustrates how varying the number of top attributes, identified through our feature ranking in Table 4.1, influences the accuracy of the XGBoost model. It's worth noting that these analyses were carried out while keeping the number of weak learner's constant at 10, a configuration that yielded the highest accuracy.

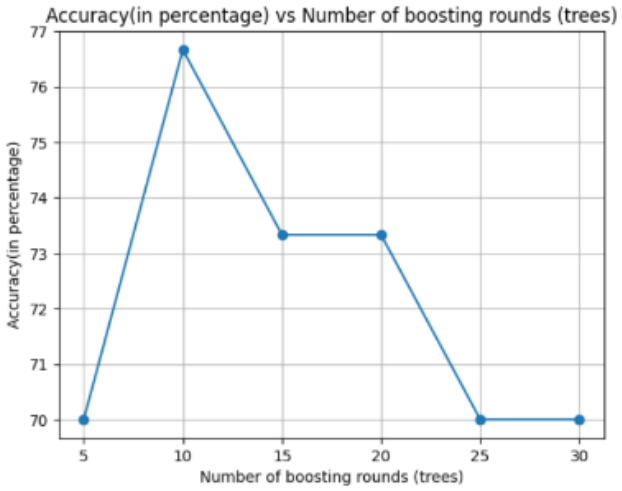


Fig 5.3 Accuracy (in percentage) vs number of boosting used in XGBoost (at constant value of 30 top features according to feature ranking obtained in Table 4.1)

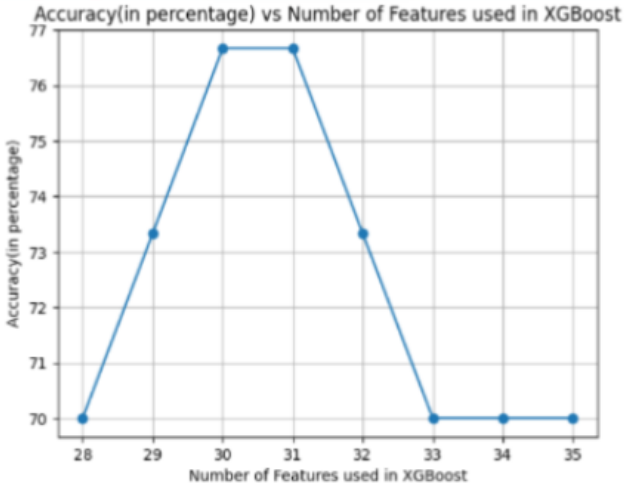


Fig 5.4 Accuracy (in percentage) vs number of top features used in XGBoost (at constant value of 10 boosting rounds) according to feature ranking obtained in Table 4.1

KNN

Figure 5.5 depicts the correlation between the value of k and the accuracy of the KNN model, leveraging the 30 top features outlined in Table 4.1. Additionally, Figure 5.6 illustrates how varying the number of top attributes, identified through our feature ranking in Table 4.1, influences the accuracy of the KNN model. It's worth noting that these analyses were carried out while keeping the value of k constant at 7, a configuration that yielded the highest accuracy.

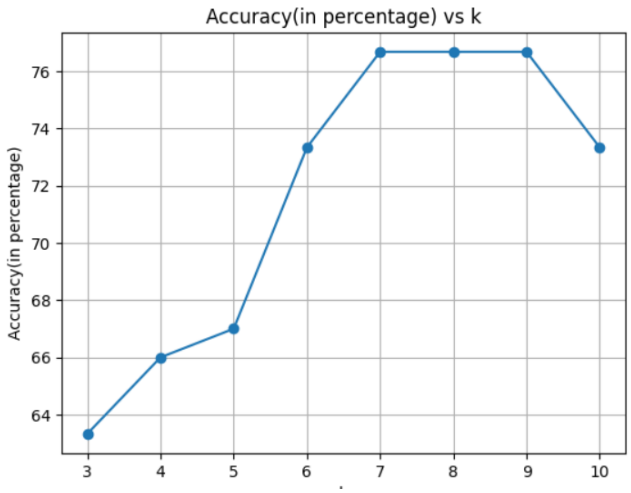


Fig 5.5 Accuracy (in percentage) vs value of k used in KNN (at constant value of 30 top features according to feature ranking obtained in Table 4.1)

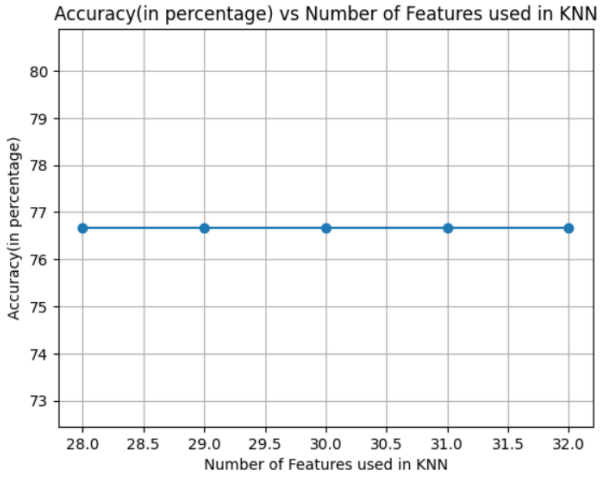


Fig 5.6 Accuracy (in percentage) vs number of top features used in KNN (at constant value of k = 7) according to feature ranking obtained in Table 4.1

SVM

Figure 5.7 depicts the correlation between the value of regularization parameter and the accuracy of the SVM model, leveraging the 30 top features outlined in Table 4.1. Additionally, Figure 5.8 illustrates how varying the number of top attributes, identified through our feature ranking in Table 4.1, influences the accuracy of the SVM model. It's worth noting that these analyses were carried out while keeping the value of regularization parameter constant at 20, a configuration that yielded the highest accuracy.

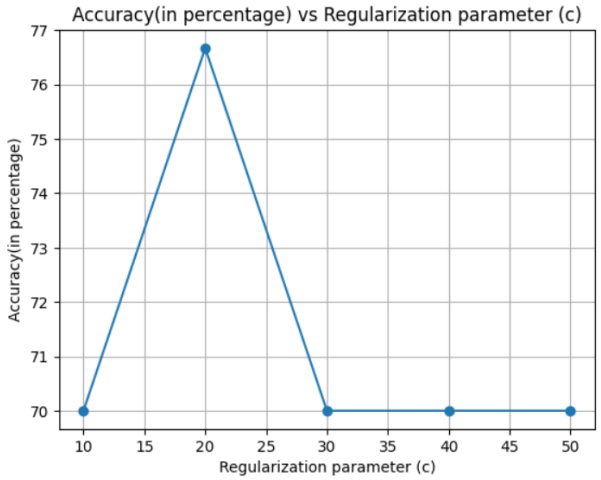


Fig 5.7 Accuracy (in percentage) vs value of c used in SVM (at constant value of 30 top features according to feature ranking obtained in Table 4.1)

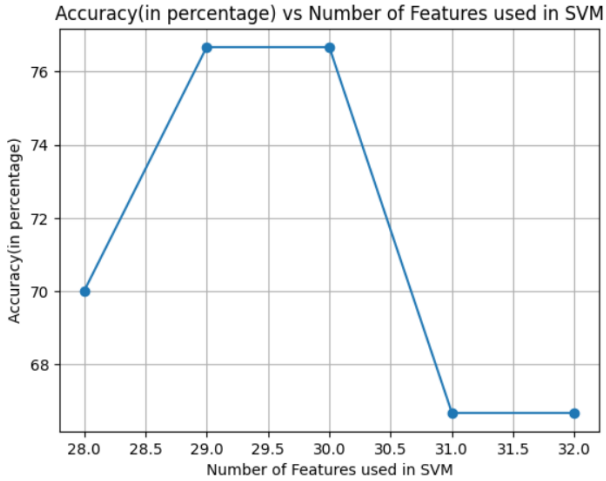


Fig 5.8 Accuracy (in percentage) vs number of top features used in SVM (at constant value of c = 20) according to feature ranking obtained in Table 4.1

Random Forest (using entropy as a splitting criterion)

Figure 5.9 depicts the correlation between the value of number of decision trees and the accuracy of the RF model (using entropy as a splitting criterion), leveraging the 30 top features outlined in Table 4.1. Additionally, Figure 5.10 illustrates how varying the number of top attributes, identified through our feature ranking in Table 4.1, influences the accuracy of the RF model (using entropy as a splitting criterion). It's worth noting that these analyses were carried out while keeping the value of number of decision trees constant at 80, a configuration that yielded the highest accuracy.

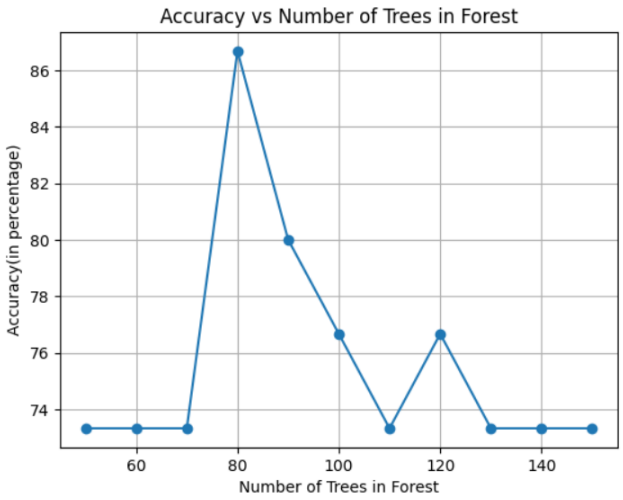


Fig 5.9 Accuracy (in percentage) vs number of decision trees used in RF model (using entropy as a splitting criterion at constant value of 30 top features according to feature ranking obtained in Table 4.1)

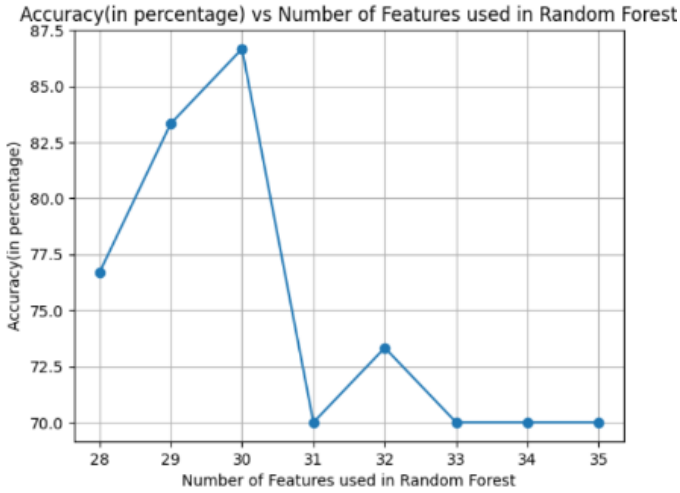


Fig 5.10 Accuracy (in percentage) vs number of top features used in RF model (using entropy as a splitting criterion at constant value of number of decision trees = 80 according to feature ranking obtained in Table 4.1)

Random Forest (using weighted entropy as a splitting criterion)

Figure 5.11 depicts the correlation between the value of weight parameter at constant number of decision trees = 80 and the accuracy of the RF model (using weighted entropy as a splitting criterion), leveraging the 30 top features outlined in Table 4.1. Additionally, Figure 5.12 illustrates how varying the number of top attributes, identified through our feature ranking in Table 4.1, influences the accuracy of the RF model (using weighted entropy as a splitting criterion). It's worth noting that these analyses were carried out

while keeping the value of number of decision trees constant at 80 and value of weight parameter = 0.3, a configuration that yielded the highest accuracy.

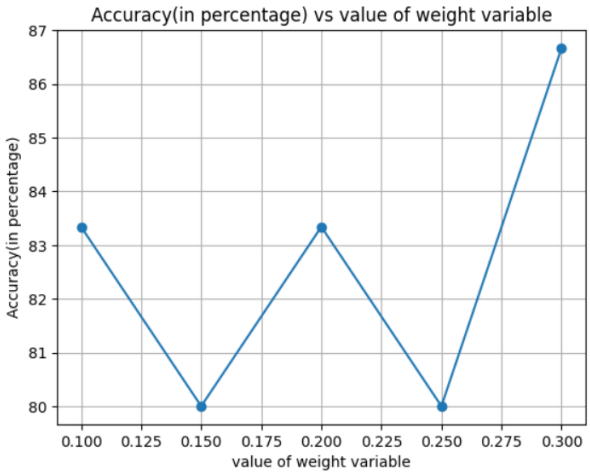


Fig 5.11 Accuracy (in percentage) vs value of weight parameter used in RF model (using weighted entropy as a splitting criterion at constant value of 30 top features according to feature ranking obtained in Table 4.1 and constant number of decision trees = 80)

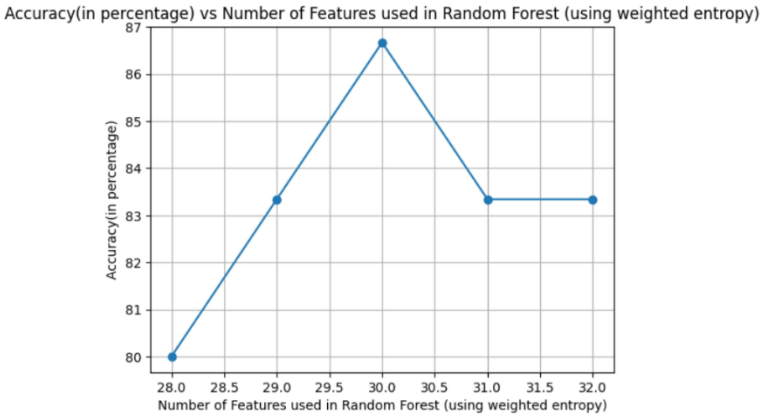


Fig 5.12 Accuracy (in percentage) vs number of top features used in RF model (using entropy as a splitting criterion at constant value of number of decision trees = 80 and constant weight = 0.3 according to feature ranking obtained in Table 4.1)

Meta Ensemble Random Forest

Figure 5.13 depicts the correlation between the value of the Meta-RF at constant number of decision trees = 75 and the accuracy of the RF model (using entropy as a splitting criterion), leveraging the 30 top features outlined in Table 4.1. Additionally, Figure 5.14 illustrates how varying the number of top attributes, identified through our feature ranking in Table 4.1, influences the accuracy of the meta-RF model (using entropy as a splitting

criterion). It's worth noting that these analyses were carried out while keeping the value of number of decision trees constant at 80, a configuration that yielded the highest accuracy.

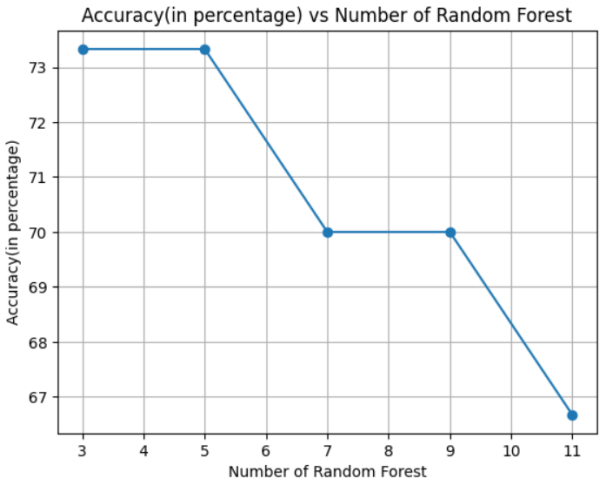


Fig 5.13 Accuracy (in percentage) vs number of ensemble of RF model (using entropy as a splitting criterion at constant value of 30 top features according to feature ranking obtained in Table 4.1 and constant number of decision trees = 80)

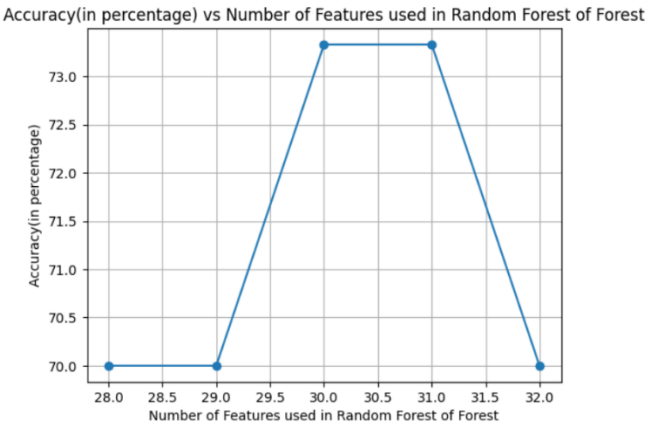


Fig 5.14 Accuracy (in percentage) vs number of top features used in meta-RF model (using entropy as a splitting criterion at constant value of number of decision trees = 80 according to feature ranking obtained in Table 4.1)

5.2 Final Real Data

The data was collected from Lokopriya Gopinath Bordoloi Institute of Mental Health, Tezpur, Assam, India. The dataset has 248 attributes and initially had about 470 rows in total.

5.2.1 Effects of changing various feature subsets and parameters of various supervised ML algorithms on accuracy

We analysed RF (using entropy as a splitting criterion) and its modifications which include using weighted entropy as a splitting criterion and the Meta-RF supervised ML algorithms and the effect on the accuracy of each algorithm on changing various parameters of each algorithm and their feature subsets.

Random Forest (using entropy as a splitting criterion)

Figure 5.15 depicts the correlation between the value of number of decision trees and the accuracy of the RF model (using entropy as a splitting criterion), leveraging the 30 top features outlined in Table 4.1. Additionally, Figure 5.16 illustrates how varying the number of top attributes, identified through our feature ranking in Table 4.1, influences the accuracy of the RF model (using entropy as a splitting criterion). It's worth noting that these analyses were carried out while keeping the value of number of decision trees constant at 75, a configuration that yielded the highest accuracy.

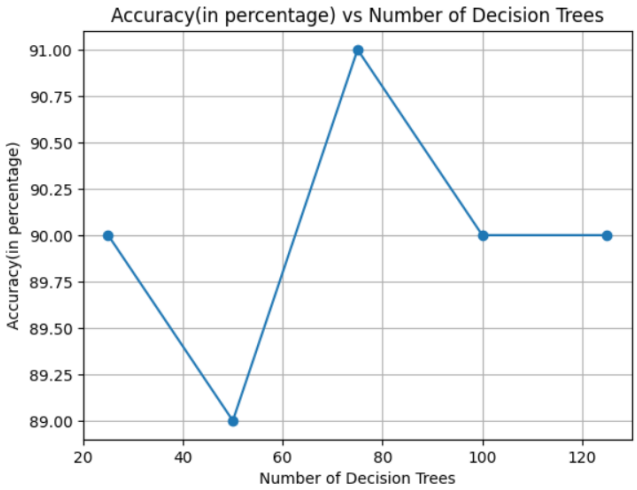


Fig 5.15 Accuracy (in percentage) vs number of decision trees used in RF model (using entropy as a splitting criterion at constant value of 30 top features according to feature ranking obtained in Table 4.1)

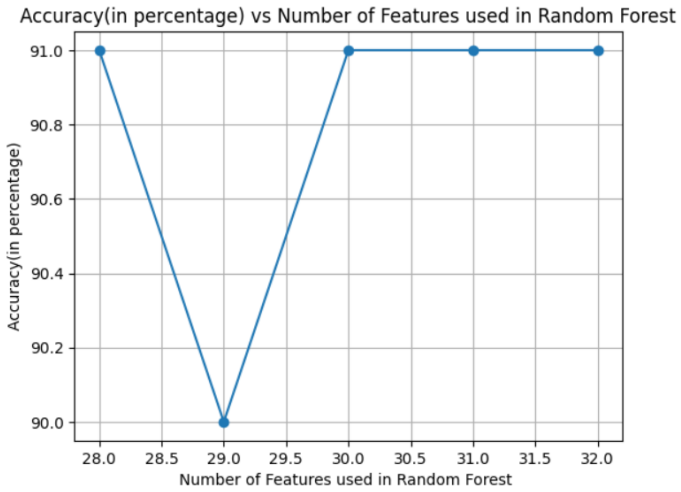


Fig 5.16 Accuracy (in percentage) vs number of top features used in RF model (using entropy as a splitting criterion at constant value of number of decision trees = 75 according to feature ranking obtained in Table 4.1)

Random Forest (using weighted entropy as a splitting criterion)

Figure 5.17 depicts the correlation between the value of weight parameter at constant number of decision trees = 75 and the accuracy of the RF model (using weighted entropy as a splitting criterion), leveraging the 30 top features outlined in Table 4.1. Additionally, Figure 5.18 illustrates how varying the number of top attributes, identified through our feature ranking in Table 4.1, influences the accuracy of the RF model (using weighted entropy as a splitting criterion). It's worth noting that these analyses were carried out while keeping the value of number of decision trees constant at 75 and value of weight parameter = 0.3, a configuration that yielded the highest accuracy.

Meta-RF (using entropy as a splitting criterion)

Figure 5.19 depicts the correlation between the number of ensembles of RFs at constant number of decision trees = 75 and the accuracy of the RF model (using entropy as a splitting criterion), leveraging the 30 top features outlined in Table 4.1. Additionally, Figure 5.20 illustrates how varying the number of top attributes, identified through our feature ranking in Table 4.1, influences the accuracy of the RF model (using entropy as a splitting criterion). It's worth noting that these analyses were carried out while keeping the value of number of decision trees constant at 75, a configuration that yielded the highest accuracy.

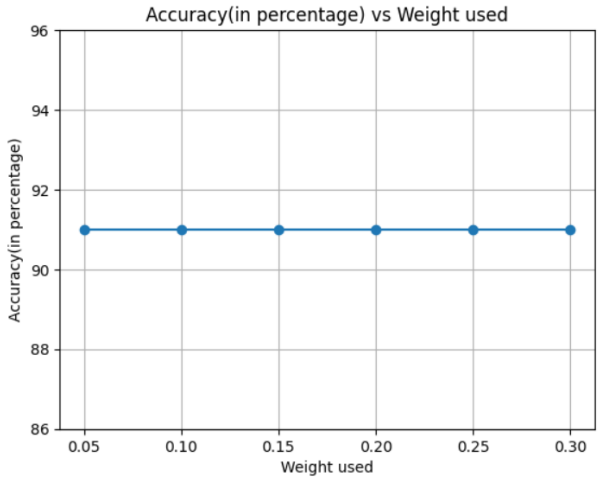


Fig 5.17 Accuracy (in percentage) vs value of weight used in RF model (using weighted entropy as a splitting criterion at constant value of 30 top features according to feature ranking obtained in Table 4.1 and constant number of decision trees = 75)

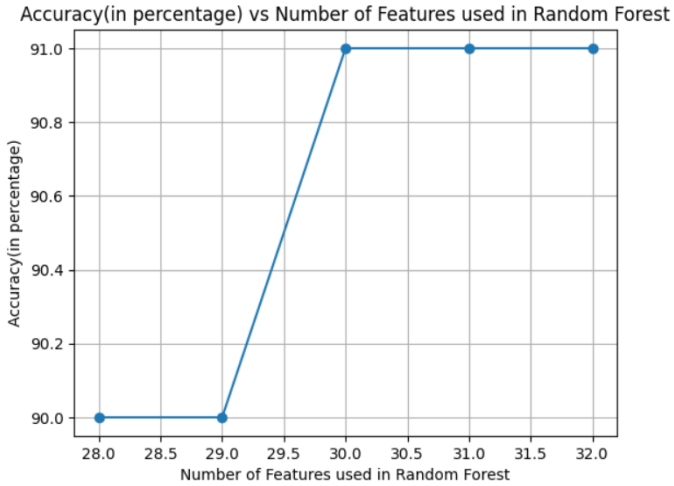


Fig 5.18 Accuracy (in percentage) vs number of top features used in RF model (using entropy as a splitting criterion at constant value of number of decision trees = 75 and constant weight = 0.3 according to feature ranking obtained in Table 4.1)

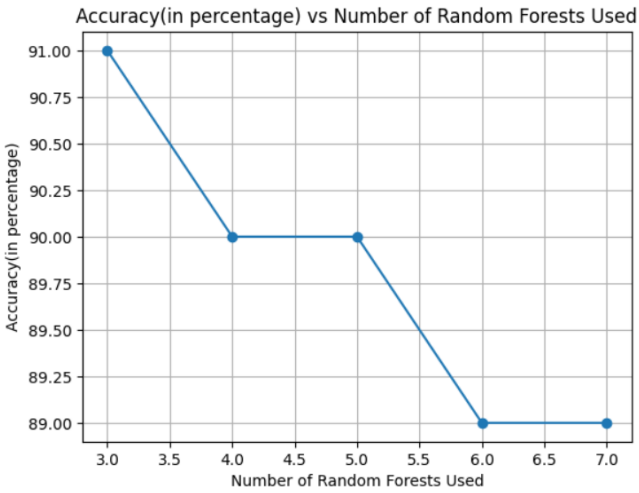


Fig 5.19 Accuracy (in percentage) vs number of ensemble of RF model (using entropy as a splitting criterion at constant value of 30 top features according to feature ranking obtained in Table 4.1 and constant number of decision trees = 75)

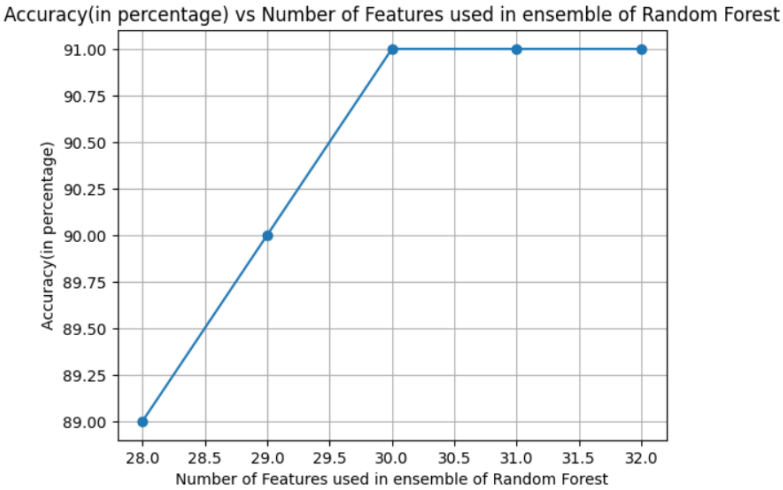


Fig 5.20 Accuracy (in percentage) vs number of top features used in meta-RF (using entropy as a splitting criterion at constant value of number of decision trees = 75 according to feature ranking obtained in Table 4.1)

5.3 Results

5.3.1 Initial Real Data

The best performance metrics for all the supervised learning algorithms we utilized and tested on the initial real data are summarized in Table 5.4.

Algorithm Name	Accuracy	Weighted Precision	Weighted Sensitivity	Weighted F1 Score
Meta Random Forest	0.733	0.819	0.733	0.683
Random Forest (using entropy)	0.867	0.866	0.867	0.870
Random Forest (using weighted entropy)	0.867	0.867	0.867	0.867
AdaBoost	0.689	0.715	0.689	0.691
XGBoost	0.767	0.779	0.767	0.764
SVM	0.767	0.803	0.767	0.753
KNN	0.767	0.826	0.767	0.781

Table 5.4 Summarized Performance Metrics of each Algorithm (Initial Data)

The RF algorithm achieved the highest accuracy of 86.67% when using entropy as the splitting criterion, with 80 decision trees and the top 30 features from Table 4.1. This model also achieved the highest sensitivity, F1 Score and second highest precision.

5.3.2 Final Real Data

The best performance metrics for all the supervised learning algorithms we utilized and tested on the final real data are summarized in Table 5.5.

Algorithm Name	Accuracy	Weighted Precision	Weighted Sensitivity	Weighted F1 Score
Meta Random Forest	0.906	0.888	0.906	0.891
Random Forest (using entropy)	0.906	0.909	0.906	0.891
Random Forest (using weighted entropy)	0.906	0.857	0.906	0.880

Table 5.5 Summarized Performance Metrics of each Algorithm (Final Data)

The RF algorithm achieved the highest accuracy of 90.68% for all its modifications. We observe that the Random Forest when using entropy as the splitting criterion, with 75 decision trees and the top 30 features identified in Table 4.1 resulted in the highest values for all the evaluation metrics used – accuracy, weighted precision, weighted sensitivity and weighted F1 Score.

CHAPTER 6: CONCLUSION

In conclusion, our study extensively investigated the performance of various supervised learning algorithms on both the initial real data and final real data, scrutinizing their accuracy and responsiveness to distinct parameter configurations and feature selections. We found that our proposed method of feature ranking using the ensemble of PC, MI and RF was superior compared to when these 3 methods are used individually to rank the features.

Notably, we observed that RF delivered the highest accuracy of 90.68% when employing entropy as the splitting criterion and trained on final real data comprising 330 patient data for training and 140 patient data for testing, with 75 decision trees and the top 30 features. Additionally, our proposed approach of utilizing weighted entropy in RF also yielded optimal accuracy, albeit with slight variations in precision and F1 score compared to the entropy-based method.

Furthermore, our exploration of Explainable AI techniques, such as LIME and SHAP, provided invaluable insights into the decision-making processes of our models, enriching our comprehension of the factors influencing predictions.

In summary, our findings underscore the critical significance of meticulous algorithm selection, parameter fine-tuning, and judicious feature selection in optimizing predictive performance. These outcomes significantly contribute to the ongoing endeavour of constructing robust and interpretable machine learning models for predictive analytics across diverse domains. Furthermore, our identification of the top 30 features as the most optimal set for predicting compliance validates their utility in developing our envisioned app and web tool. This selection underscores the importance of these features in enhancing the accuracy and effectiveness of our model.

REFERENCES

[1] Ang, J. C., Mirzal, A., Haron, H., & Hamed, H. N. A. (2016). Supervised, unsupervised, and semi-supervised feature selection: A review on gene selection. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 13(5), 971-989. https://doi.org/10.1109/TCBB.2015.2478454.

[2] Awan, S. E., Bennamoun, M., Sohel, F., Sanfilippo, F. M., & Dwivedi, G. (2019). Machine learning-based prediction of heart failure readmission or death: Implications of choosing the right model and the right metrics. ESC Heart Failure, 6(2), 428-435.

[3] Biecek, P., & Burzykowski, T. (2021). Local interpretable model-agnostic explanations (LIME). In Explanatory model analysis: Explore, explain, and examine predictive models (pp. 107–124).

[4] Birmaher, B., & Axelson, D. (2006). Course and outcome of bipolar spectrum disorder in children and adolescents: A review of the existing literature. Development and Psychopathology, 18(4), 1023–1035.

[5] Blum, A. L., & Langley, P. (1997). Selection of relevant features and examples in machine learning. Artificial Intelligence, 97, 245-271.

[6] Bohlmann, A., Mostafa, J., & Kumar, M. (2021). Machine learning and medication adherence: Scoping review. JMIRx Med, 2(4), e26993.

[7] Brezo, J., Paris, J., Barker, E. D., et al. (2007). Natural history of suicidal behaviors in a population-based sample of young adults. Psychological Medicine, 37(11), 1563–1574.

[8] Carlson, G. A., Kotov, R., & Chang, S. W., et al. (2012). Early determinants of four-year clinical outcomes in bipolar disorder with psychosis. Bipolar Disorders, 14(1), 19–30.

[9] Chaix, B., Bibault, J. E., Pienkowski, A., Delamon, G., Guillemassé, A., & Nectoux, P., et al. (2019). When chatbots meet patients: One-year prospective study of conversations between patients with breast cancer and a chatbot. JMIR Cancer, 5(1), e12856.

[10] Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785-794). ACM.

[11] Cramer, J. A., Roy, A., Burrell, A., Fairchild, C. J., Fuldeore, M. J., Ollendorf, D. A., & Wong, P. K. (2008). Medication compliance and persistence: Terminology and definitions. Value in Health, 11(1), 44–47.

[12] Evgeniou, T., & Pontil, M. (2001). Support vector machines: Theory and applications. Machine Learning and Its Applications: Advanced Lectures, 2049, 249-267. https://doi.org/10.1007/3-540-44673-7_12.

[13] Gardiner, P., & Dvorkin, L. (2006). Promoting medication adherence in children. American Family Physician, 74, 793-798, 800.

[14] Guo, G., Wang, H., Bell, D., Bi, Y., & Greer, K. (2003). KNN model-based approach in classification. In On the Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE (pp. 986-996). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-39964-3_62.

[15] Han, J., Kamber, M., & Pei, J. (2012). Data mining: Concepts and techniques. Elsevier.

[16] Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction (2nd ed.). Springer Science & Business Media.

[17] Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. Science, 349, 255–260. https://doi.org/10.1126/science.aaa8415.

[18] Jovel, J., & Greiner, R. (2021). An introduction to machine learning approaches for biomedical research. Frontiers in Medicine, 8, 771607. https://doi.org/10.3389/fmed.2021.771607.

[19] Krenn, M., Buffoni, L., & Coutinho, B., et al. (2023). Forecasting the future of artificial intelligence with machine learning-based link prediction in an exponentially growing knowledge network. Nature Machine Intelligence, 5, 1326–1335. https://doi.org/10.1038/s42256-023-00735-0.

[20] Lask, B. (2003). Motivating children and adolescents to improve adherence. Journal of Pediatrics, 143, 430-433.

[21] Lieberman, D. A. (2001). Management of chronic pediatric diseases with interactive health games: Theory and research findings. Journal of Ambulatory Care Management, 24, 26-38.

[22] Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2020). Explainable AI: A review of machine learning interpretability methods. Entropy, 23(1), 18. https://doi.org/10.3390/e23010018.

[23] Maan, C. G., Munnawar, H. M. S., Heramani, N., & Lenin, R. K. (2015). Factors affecting non-compliance among psychiatric patients in the Regional Institute of Medical Sciences, Imphal. Volume 5, Issue 1 (January 2015), PP. 01-07.

[24] McCorduck, P. (2004). Machines who think: A personal inquiry into the history and prospects of artificial intelligence. CRC Press. https://doi.org/10.1201/9780429258985.

[25] Müller, A. C., & Guido, S. (2016). Introduction to machine learning with Python: A guide for data scientists. O'Reilly Media.

[26] Nichols-English, G., & Poirier, S. (2000). Optimizing adherence to pharmaceutical care plans. Journal of the American Pharmaceutical Association, 40, 475-485.

[27] Nohara, Y., Matsumoto, K., Soejima, H., & Nakashima, N. (2022). Explanation of machine learning models using shapley additive explanation and application for real data in hospital. Computer Methods and Programs in Biomedicine, 214, 106584. https://doi.org/10.1016/j.cmpb.2021.106584.

[28] Olino, T. M., Seeley, J. R., & Lewinsohn, P. M. (2010). Conduct disorder and psychosocial outcomes at age 30: Early adult psychopathology as a potential mediator. Journal of Abnormal Child Psychology, 38(8), 1139–1149.

[29] Piette, J. D., Sussman, J. B., Pfeiffer, P. N., Silveira, M. J., Singh, S., & Lavieri, M. S. (2013). Maximizing the value of mobile health monitoring by avoiding redundant patient reports: Prediction of depression-related symptoms and adherence problems in automated health assessment services. Journal of Medical Internet Research, 15(7), e118.

[30] Rapoff, M. A. (2010). Consequences of nonadherence and correlates of adherence. In Roberts, M. (Ed.), Adherence to pediatric medical regimens, issues in clinical child psychology (pp. 33–46). Springer.

[31] Rashedi, E., Nezamabadi-pour, H., & Saryazdi, S. (2013). A simultaneous feature adaptation and feature selection method for content-based image retrieval systems. Knowledge-Based Systems, 39, 85-94.

[32] Reilly, E. L., Wilson, W. P., & McClinton, H. K. (1967). Clinical characteristics and medication history of schizophrenics readmitted to the hospital. International Journal of Neuropsychiatry, 39, 85-90.

[33] Rekha, R., Masroor, J., Sushma, K., & Prashant, K. C. (2005). Reasons for drug non-compliance of psychiatric patients: A centre-based study. Journal of the Indian Academy of Applied Psychology, 31, 24-28.

[34] Rodenhausser, P., Schwenker, C. E., & Khamis, H. J. (1987). Factors related to drug treatment refusal in a forensic hospital. Hospital Community Psychiatry, 38, 631-637.

[35] Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. SN Computer Science, 2(3), 160. https://doi.org/10.1007/s42979-021-00592-x.

[36] Selen, Y., Wertheimer, A., & Dublin, W. (2003). Demographical factors affecting patient compliance to medications in an outpatient psychiatric clinic: A preliminary study. FABAD J. Pharm. Sci., 28, 77-84.

[37] Tu, C., Liu, H., & Xu, B. (2017). AdaBoost typical algorithm and its application research. MATEC Web of Conferences, 139, 00222. https://doi.org/10.1051/matecconf/201713900222.

[38] Vidya, K., Anzar, A., Li, Z., Lei, G., Shaolei, F., Vijay, Y., & Isaac, R. (2020). Accuracy of machine learning-based prediction of medication adherence in clinical research. Psychiatry Research, 294, 11355.

[39] World Health Organization. (2003). Adherence to long-term therapies: Evidence for action. World Health Organization. https://apps.who.int/iris/handle/10665/42682.

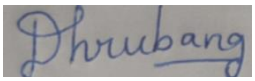
[40] YourDataTeacher. (2021). Feature selection with random forest. Retrieved from https://www.yourdatateacher.com/2021/10/11/feature-selection-with-random-forest/.

PLAGIARISM UNDERTAKING

I hereby declare that the project work presented in the dissertation report titled "Predicting Poor Compliance to Psychotropics using Machine Learning Approach" is solely my own work, with no significant contribution from any other person/report. Any small contributions or help received have been duly acknowledged. The complete report has been written by me.

I understand the zero-tolerance policy of Tezpur University towards plagiarism. Therefore, as the author of the above-titled thesis, I declare that no portion of my report has been plagiarized. Any material used as a reference is properly cited.

I undertake that if I am found guilty of any formal plagiarism in the above-titled report, even after the award of my B.Tech (CSE) degree, Tezpur University reserves the right to revoke my B.Tech (CSE) degree.

Student Signature: 

Name: Dhrubang Utpal Talukdar

Enrolment Number: CSB20049