A Mini Project report

on

PREDICTING POOR COMPLIANCE TO PSYCHOTROPICS USING MACHINE LEARNING

Of

Bachelor Of Technology

In

Computer Science and Engineering

Submitted 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
Lokopriya Gopinath Bordoloi Regional Institute of Mental Health

School Of Engineering
Department of Computer Science and Engineering
Tezpur University
Napaam - 784028, Assam, India

Dec 2023

Department of Computer Science and Engineering

Tezpur University

Certificate by the HoD

This is to certify that the dissertation entitled "**Predicting poor compliance to psychotropics using machine learning approach**" is submitted by **Dhrubang Utpal Talukdar** bearing Roll no: **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 & 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.

Date:

Head of the Department

Department of Computer Sc & Engineering

Place**:**

Tezpur University

Department of Computer Science and Engineering

Tezpur University

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 Roll no: **CSB20049** is carried out by him 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.

Date:

Dr.Rosy Sarmah

Associate Professor

Department of Computer Sc & Engineering

Place**:**

Tezpur University

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 Roll no: **CSB20049** is carried out by him 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.

Date:

Dr. Siddeswara BL

Assistant Professor

Department of Child & Adolescent Psychiatry

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

Department of Computer Science and Engineering

Tezpur University

DECLARATION

I hereby declare that the dissertation work titled "**Predicting poor compliance to psychotropics using machine learning approach**" submitted to the Department of Computer Science & Engineering. Tezpur University is prepared by me and was not submitted to any other institution for award of any other degree.

Date:

Dhrubang Utpal Talukdar

CSB20049

Department of Computer Sc & Engineering

Place**:** Tezpur University

ACKNOWLEDGEMENT

I would like to extend my heartful gratitude to my project guide Dr. Rosy Sarmah, Associate Professor, Dept. of CSE, Tezpur University, for giving us the opportunity to work under him and providing me ample guidance and support through the course of the project. I am highly indebted to Dr. Siddeswara BL, Assistant Professor, Department of Child & Adolescent Psychiatry, LGBRIMH, Tezpur for his helpful guidance as well as for providing necessary information regarding the project.

I would also like to thank Head of the department of Computer Science and Engineering department and all the faculty members of Dept. of CSE, Tezpur University for the valuable guidance and co-operation throughout the project.

My thanks and appreciations also go to all other people who have directly or indirectly helped me out with their abilities.

Dhrubang Utpal Talukdar

ABSTRACT

Background

Nonadherence to medications is a huge problem to national health systems. According to the World Health Organization, the primary way to improve overall health is by improving compliance to medical treatments, compared to improvement in medical treatments.

Good compliance is important for effective psychotropic treatments. According to recent research, the adherence of children to their treatment plans is poor when compared to adults. Special measures are needed to combat these problems for children, as non-compliance to treatment plans lead to adverse negative impacts, which include worsening of mental health conditions, prevalence of symptoms and recurrence of mental illnesses. These factors negatively influence their daily life activities, social relations, academic performances and their psychological well-being.

Machine learning approaches have played a pivotal role in the field of healthcare for diagnosis, prognosis and prediction of treatments. Machine learning algorithms can find patterns in huge datasets, analyse large datasets quickly and effectively, and draw key insights from datasets. In this work, we propose two clustering methods based on k-nearest neighbour (KNN) to group children reporting at LGBRIMH, Tezpur for various mental health problems. The clustering is based on 78 physiological, socio-demographic, economic and psychotropic treatment features. An ensemble feature selection method is used to obtain the reduced feature subset. Clustering is performed to obtain paediatric patient groups using this feature subset and form a hypothesis of medicine adherence among the children. The cluster results help to obtain a rough prediction of medicine adherence among children. Our methods have been compared with k-medoids, fuzzy c-means and DBSCAN using Adjusted Rand Index (ARI), Adjusted Mutual Information (AMI), Accuracy, Silhouette Index, Davies-Bouldin Index and Calinski-Harabasz Index. We observe that both our algorithms gave better result w.r.t. the above-mentioned validation indices. We have further developed a web tool for use by clinicians to observe the predictions in a user-friendly way.

Contents

List of Figures

List of Tables

CHAPTER 1

INTRODUCTION

1.1 Non-Compliance to psychotropic treatments in children

In medical literature and practices by clinicians, term the way of medicines being taken by patients as adherence or compliance. These two terms are often interchanged and used synonymously and thought to meant he same [1]. There is a slight difference between the two terms. Adherence to medications is described by the World Health Organization as **"the extent to which a person's behavior- taking medication, following a diet, and/or executing lifestyle changes- corresponds with the agreed recommendations from a healthcare provider"** [2]. Compliance on the other hand is defined as **"the extent to which a patient acts in accordance with the prescribed interval and dose of a dosing regimen"** [3].

Nonadherence indicates that a patient's self-management of their health and a medically prescribed regimen are not well aligned [4]. In today's medical treatments nonadherence to treatments by patients is commonly observed. Speaking only about children and adolescents, as much as 50-88 % of them do not show proper adherence to their treatments [5]. Rehospitalizations and morbidity could be increased because of nonadherent measures [6]. Age is a very crucial factor for this high rate for non-compliance as children are not cognitively and emotionally developed [7].

1.2 Factors Affecting Compliance

Family background too impacts compliance level. Lower income families are less compliant to their prescribed medications. With the illness getting older the compliance levels usually drop [20]. It was seen that families with incomes less than Rs. 10000 and with 1-2 years of illness showed low compliance levels [14].

The type of illness too impacted compliance. The most prevalent mental health condition associated with noncompliance was schizophrenia [19]. Another factor affecting compliance were the side effects observed due to medications. Most patients discontinued their medications on observance of side effects [15]. Hopelessness due to poor support too affected compliance

[16]. Females were usually found to have lower levels of compliance compared to males [17]. Children belonging to illiterate families were less compliant to medications [18]. The most frequent excuses for noncompliance include subjective well-being, drug paranoia, lack of understanding of the condition, adverse drug reactions, loss of hope for a solution, inadequate caregiving or support, and financial difficulties [14]

1.3 Effects of low psychotropic compliance in children

A more severe course of illness, morbidities like substance abuse and suicidal thoughts, and the presence of comorbidities and complications like poor performance at work and school, interpersonal conflicts, or legal issues have all been linked to many psychiatric disorders that present in youth at an early age [21-24]. Despite great efforts, treatment challenges are frequent for psychiatric diseases in young people, and illness-related effects, such as mortality, are highly prevalent [25].

1.4 Improving Non-Compliance in children

A regular update between the paediatrician and child is indeed needed to improve the compliance to treatments [8]. Paediatricians need to introduce a plan which is easily self-manageable without many problems. Sheets can be provided to the child and their family, which contain medicine, doses, schedule and common side-effects [9].

Proper communication is one of the vital ways to improve compliance in children. Language barriers could be one of the prime factors for low compliance levels of treatments. Measures like patient sheet translations can therefore improve the compliance in patients [10].

Technologies can be used to influence children. It is seen that self-care levels of children go up on playing video games related to health education and disease management [11]. Another example of advance technology which could be used in this field are electronic caps, which could reveal daily pattens of medications by recording information about the frequency of opening of medical caps [12].

Patients too need to be educated on this topic. Beforehand awareness on the benefits of medications and the possible observed side effects can increase the compliance levels [13]. Social support, positiveness and helpfulness are keys to high compliances.

1.5 Machine Learning

Machine learning in simple terms is finding patterns in huge datasets and convey meaning form them. It is a subset of artificial intelligence mostly used for prediction purposes. Thus, when a system learns information from data, it is called machine learning [26,27]. In classical terminologies the three types of machine learning are – Supervised learning, Unsupervised Learning and Reinforcement learning [28].

Machine learning is not a new concept.it is a century old idea [29]. Recent technological advancements have however hastened the pace for machine learning [30]. Now a days data can be collected quickly. Huge storage spaces in forms of clouds are available for data. This increased pace of data collection, capacity to store huge data volumes promote machine learning.

The emergence of major software corporations like Google, Facebook etc. has led to the development of the most well-known machine learning (ML) frameworks (libraries) in the community today. TensorFlow, Keras, PyTorch, Scikit-Learn, Caffe, CNTK, Lasagne, and Theano are a few of these libraries. Additionally, thanks to cloud computing firms like Google, Amazon, Microsoft, and IBM, clients may now obtain Artificial Intelligence algorithms as a Service (AIaaS). In recent years, several machine learning implementations have made it possible to solve problems in real life that were previously intractable [31].

Machine learning has crept its way to today's world hugely involved in various tasks like image/speech detection, analysis, fraud detection, research or medical diagnosis.

1.6 Supervised Learning

Assume that the real estate company wishes to calculate a home's price depending on specific qualities. Initially, the company would first gather a dataset including lots of examples [32-34]. Each case is a distinct analysis of a home and its surrounds. Features are the documented attributes of a house that might be useful in pricing estimation, such as total square footage, number of stories, and yard space [32-34]. The aim is the feature to be expected, in this example the home price. In general, datasets are categorized into three groups: testing, validation, and training. The data used to train models gives them the greatest results [32,33].

Using patterns identified in the training dataset, features are mapped to the target in supervised learning, allowing an algorithm to predict home prices on subsequent datasets. Because the

model infers an algorithm from feature-target pairs and the target provides feedback to the model regarding the accuracy of its predictions, this approach is supervised.

The tests are made on an unseen dataset called the test dataset on which the model tries to make its assumptions and check for its correctness [32,33]. Put another way, by mastering the mapping function, f, features, x, are mapped to the target, Y, making it possible to estimate future home prices using the algorithm $Y = f(x)$.

The two most common supervised learning tasks are regression and classification. Regression analysis is used to predict numerical data, such as test or lab findings, or item prices, like in the example of housing prices. However, categorization entails figuring out which group an example belongs to. Classifications is a method or discrete values while regression is a method for continuous values. Some popular supervised learning algorithms include Random Forest, Decision Tree, Ridge Regression and Support Vector Machine etc.

1.7 Unsupervised Learning

Unlike supervised learning, which assigns a category to each instance of a dataset, unsupervised learning searches for patterns in the dataset [32,33,35]. Since the algorithm finds patterns in a dataset without consulting a target, these methods are considered unsupervised. The most widely used unsupervised learning tasks are clustering, association, and anomaly detection [32,33,35].

As the name suggests, clustering separates instances within a collection based on specific feature combinations into many clusters [32,33,35]. Assume the real estate company applies a clustering technique to their dataset and finds three distinct clusters. Upon closer inspection, it might become clear that the three distinct architects that designed the houses in their dataset are represented by the clusters, which weren't included in the training dataset.

Main aim of unsupervised learning is to find trends in the data to segregate the data into different groups. Some popular unsupervised learning algorithms include KNN, k-means etc.

1.8 Machine Learning in psychotropics

Health care costs and complexities is predicted to keep rising, unless future-based machine learning solutions are implemented [36]. Machine learning could automate the health care systems and cut costs for medical treatments.

Recent published reviews discuss how effective machine learning solutions are to improve adherence [37]. Machine learning can track down some strong predictors, which largely influence the predictions [36]. With the course of time, we would find the future medical systems including psychotropics reliant on machine learning approaches, whether it would be for grouping tasks or prediction tasks etc.

Our work has the following objectives: (i) Considering the significance of predicting adherence in children to their undergoing psychotropic treatments, we propose to use unsupervised learning to form a hypothesis to roughly predict compliance in children using our own clustering algorithms – Hierarchical KNN and Association KNN. We have tested our algorithms on a dataset of children, suffering from psychotropic conditions to group them based on their adherence to treatment plans and (ii) develop a web tool to predict compliance of children in advance.

Our algorithms have outperformed other algorithms like k-medoids, fuzzy c-means and DBSCAN with respect to Adjusted Rand Index, Adjusted Mutual Index, Accuracy, Silhouette Index, Davies-Bouldin Index and Calinski-Harabasz Index.

We developed a web tool which would show the compliance of the patient based on the entered details. We have used Hierarchical KNN (k – 50, threshold – 5, outlier threshold – 10) as it gave the best evaluation metrics. In future we will also integrate other clustering methods as well as Association KNN in the web tool. We plan to use supervised methods of machine learning to improve the accuracy of the model.

CHAPTER **2**

RELATED WORKS

The use of machine learning algorithms for predicting adherence have already been proposed. For prediction of adherence, it is necessary to use strong predictors which highly influence our results [38].

Proposals to predict depressions using machine learning among patients have also been made [39]. Interactive voice response tests were used to monitor the medication adherence of depression patients both prospectively and currently [39]. The data for this study came from 208 patient assessments of their interactive voice answers. These assessments were then used to create projections. The model's effectiveness was enhanced by age, baseline physical functioning, and past medication adherence, among these variables. The only method used to predict future drug adherence was logistic regression.

The next study used computer vision and face recognition software to monitor the drug adherence of 53 patients with schizophrenia [40]. The drug ingestion records that the study participants recorded using their smartphones' cameras were sent to the research team. Following that, face recognition and computer vision algorithms were used to alert patients displaying suspicious behaviour, indicating a low possibility of long-term adherence. Furthermore, this method reminded patients to take their medications at a specific time each day. The main limitation of the research was the patients' capacity to choose between app-based monitoring and in-person observation, which led to bias.

Using the adherence categorization, four distinct models were developed to predict a desired outcome at different phases of a clinical trial. As dynamic features, the daily values of adherence, adjusted adherence, number of interventions, dosage delay, and dose length were used at varying intervals for the model. The remaining features, which included condition, trial length, and micro reimbursements, functioned as static predictors in all model types. Sub-datasets with daily and common attributes were created based on the specified intervals. XGBoost classifier was used to predict the drug compliance based on the dynamic and static feature which were recorded videos of patients on their smartphones [41].

CHAPTER 3

BACKGROUND

3.1 Feature Selection

Massive volumes of data, including text, voice, video, photo, and social media data, as well as data gathered from sources including the Cloud computing and the Internet of Things, have been produced at a rate never seen before thanks to the amazing new computer and internet applications that have emerged with the rapid growth of modern technology. Due to the enormous dimensionality of these data, data analysis and decision-making are frequently quite difficult. It has been demonstrated that feature selection works well in both in practical and theoretical measures to analyse data up to massive dimensions and improve efficiency for learning [56] [57] [58].

The process involving the extraction of a subset of relevant features from the original set containing features applies a predetermined feature selection criterion to pertinent features found in the dataset is known as feature selection. By eliminating aspects that are superfluous and unnecessary, it contributes to a reduction in the volume of data processing. Learning objectives can be made simpler, learning times can be shortened, and learning accuracy can be increased [59] [60] [61]. These techniques can be used to pre-process learning algorithms.

Three types of feature selection are: -

1)Supervised Feature Learning

2)Unsupervised Feature Learning

3)Semi-Supervised Learning

3.1.1 Supervised Feature Learning

Supervised method of feature selection is based on fundamental idea of the correlation, similarity and relevance between the features of our dataset with the class labels. It usually focuses on issues related to classification. The importance of the features can be determined by metrics for significance. The main aim of this type of feature learning is to select the subset of features which would maximise the accuracy of the model [62].

e.g. Support Vector Machine, Random Forest

3.1.2 Unsupervised Feature Learning

Unsupervised feature selection techniques deal mostly in classification of data naturally. It strives on to improve the accuracy while performing clustering by selection of a subset containing set of features, which are to be evaluated using various criteria for evaluation. Methods under this can be classified into 2 types - i) unsupervised filter ii) wrapper feature selection methods which differ on basis whether they make use of a cluster algorithm [62].

Based on the properties of the data, unsupervised filter feature selection techniques are used to select the features. Since we do not apply any clustering algorithm or use any of the learning algorithms while selection of features, we see a decreased algorithmic and time complexity for clustering. We may extend to using statistical data obtained during model training for evaluating the results

e.g. PCA, Pearson correlation

3.1.3 Semi - supervised Feature Learning

For training a model we use only the features and exclude the target. A training model uses datasets divided into two parts – the features and the target. Semi - supervised method of learning is based on the idea of training a model on a dataset with the labels as well as on a dataset without labels. Semi supervised feature selection methods, particularly filter models, are central to semi-supervised learning.

Score functions, consist of 4 categorical measurers —Fisher Score [63–68], Laplacian Score [69–72], Constraint Score [73–75] and Variance Score [76]. These are used in most semi-supervised feature selection methods.

e.g. Mutual Information, Label Propagation

In our project we have used 3 feature selection techniques to rank the features. We have used Random Forest, Pearson correlation and Mutual Information.

3.1.4 Random Forest [77]

A potent supervised learning algorithm that is excellent at selecting features is called Random Forest. By determining how each feature affects the predictive accuracy of multiple decision trees, it assesses the significance of each feature separately. The algorithm rates the features according to how well they contribute to the overall performance of the model by assigning an importance score to each feature. Feature scores are increased for consistently significant features across different tree ensembles to aid in effective dimensionality reduction. Since Random Forest can compute quickly, find important predictors, and improve the interpretability of the model, it is a useful tool in many machine learning applications.

3.1.5 Pearson Correlation [78]

The Pearson coefficient expresses the magnitude and direction of the linear relationship between two variables. Positive correlations are indicated by positive coefficients, and negative correlations are indicated by negative coefficients, with a range of -1 to 1. Strong linear relationships are indicated by coefficients near 1 or -1, whereas values close to 0 suggest weak correlations. When selecting features and evaluating redundancy in data analysis, Pearson correlation—a popular statistical tool—is especially good at spotting linear relationships between features.

$$r = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2 \sum_{i=1}^{n}(Y_i-\bar{Y})^2}}$$

Here:

- X_i and Y_i are the individual data points for variables X and Y,
- $\bar{X}$ and $\bar{Y}$ are the means of X and Y respectively,
- n is the number of data points.

3.1.6 Mutual Information [79]

Mutual information, or MI, quantifies the amount of information shared by two random variables, X and Y. Mutual information is used to assess a feature's dependence on the target variables during the feature selection process. A high MI value suggests strong dependency, whereas a low value suggests independence. Because MI is so good at capturing non-linear

relationships, it is widely used for informative feature selection in machine learning tasks like text mining and image analysis.

$$MI(X,Y) = \sum_{x \in X} \sum_{y \in Y} P(x,y) \log \left(\frac{P(x,y)}{P(x)P(y)} \right)$$

Here:

- $P(x,y)$ is the joint probability distribution of X and Y,
- $P(x)$ and $P(y)$ are the marginal probabilities of X and Y respectively.

3.2 Clustering

The most significant unsupervised learning problem, clustering, deals with data structure partitioning in unknown domains and establishes the foundation for additional learning [42]. Clustering has no proper definition; however, we can use a few guidelines to describe it classically [43]:

1. Items belonging to the same cluster should be similar.
2. Items in one cluster should be dissimilar to items in other clusters.
3. Appropriate distance measures should be used to evaluate the items in the cluster.

In our project we have incorporated algorithms like DBSCAN [44], K-Medoids [45] and Fuzzy C-Means [47] to compare their performance to the two algorithms we have proposed – Association KNN and Hierarchical KNN.

3.2.1 DBSCAN [44]

In machine learning and data mining, the DBSCAN clustering algorithm is widely utilized to identify groups of data points based on their density distribution. Unlike traditional clustering algorithms, which assume that clusters are spherical, DBSCAN can find clusters of any shape.

Classifying each data point as a noise, border, or core point is the fundamental idea behind DBSCAN. Core points are those that are surrounded by a specific number of other points and lie inside a predefined radius. Border points are near a core point, even though they don't have enough neighbours to be considered core points. Noise points are outliers that do not fit into any cluster.

DBSCAN works well with datasets that have varying cluster densities and shapes. It is particularly useful when the structure of the data is unknown ahead of time because it is noise-

resistant and can automatically detect the number of clusters. On the other hand, DBSCAN performance can be affected by the choice of parameters, such as the radius and minimum number of points required to form a dense region.

3.2.2 K-Medoids [45]

K-Medoids is a clustering algorithm that splits a dataset into a fixed number of clusters (k), much like K-Means. However, K-Medoids does not use cluster centroids; rather, it uses medoids, which are the most centrally located data points within each cluster.

The algorithm iteratively optimizes cluster assignments and medoids to minimize the total dissimilarity between data points and their corresponding medoids. K-Medoids are robust in scenarios where the data may have uneven distributions or outliers because they can handle non-Euclidean distance metrics and are less sensitive to outliers than K-Means.

Numerous fields, such as biology, pattern recognition, and image analysis, have applications for K-Medoids [46]. Because of its ability to identify robust cluster representatives, it is helpful in situations where the mean (as used in K-Means) may be impacted by outliers or may not accurately reflect the cluster's central tendency.

3.2.3 Fuzzy C-Means [47]

Fuzzy C-Means is a clustering algorithm that builds upon the traditional K-Means algorithm to handle fuzzy memberships. Every data point in FCM is linked to every cluster and has a degree of membership that ranges from 0 to 1. Unlike K-Means, which assigns a single cluster to each data point, FCM allows for a more intricate representation of the data's concurrent membership in multiple clusters.

The algorithm iteratively updates the cluster centres and membership degrees until convergence by minimizing the sum of the weighted distances between the assigned cluster centres and the data points. FCM provides a more adaptable and practical method of clustering, and it is particularly useful when data points exhibit partial membership to more than one cluster.

One benefit of FCM is its adaptability to various cluster sizes and shapes. Nonetheless, the system's performance might be impacted by the fuzziness parameter and the choice of the initial membership value. Applications for FCM include pattern recognition, image segmentation, and data clustering tasks where conventional crisp clustering methods may not be adequate.

After performing clustering, we have evaluated our clusters using internal cluster validation metrics, external cluster validation metrics and accuracy.

3.3 External Cluster Validation Metrics

External cluster validation metrics evaluate the performance of clustering algorithms by comparing the results to a ground truth [48]. To quantify the similarities between the ground truth clusters and the original clusters for comparison with ground truth, we try to use different similarity scores. The similarity scores we have used are Adjusted Rand Index (ARI) [49] and Adjusted Mutual Information (AMI) [50].

3.3.1 Adjusted Mutual Information (AMI) [50]:

AMI assesses the agreement between two clustering by considering both homogeneity and completeness. A score of 1 indicates perfect agreement, while a score of 0 indicates no agreement at all. An AMI score that is higher indicates better clustering similarity because it considers both the purity and completeness of each individual cluster.

$$AMI(U,V) = \frac{MI(U,V) - E[MI(U,V)]}{\max(H(U),H(V)) - E[MI(U,V)]}$$

- U and V are two clusterings being compared.
- $MI(U,V)$ is the Mutual Information between U and V.
- $H(U)$ and $H(V)$ are the entropies of U and V.
- $E[MI(U,V)]$ is the expected Mutual Information under a random model.
  - Meaning: AMI measures the normalized agreement between two clusterings, accounting for both completeness and homogeneity.

3.3.2 Adjusted Rand Index (ARI) [49]

To assess how similar two clustering are, ARI compares agreements and disagreements while accounting for chance. The range is from -1 to 1. A score of one indicates perfect agreement, a score of zero indicates random clustering, and a score of negative indicates disagreement. ARI is useful for assessing the accuracy of clustering solutions in the presence of chance.

$$ARI(U,V) = \frac{\sum_{ij}\binom{n_{ij}}{2} - [\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}]/\binom{n}{2}}{\frac{1}{2}[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}] - [\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}]/\binom{n}{2}}$$

- U and V are two clusterings being compared.
- n_{ij} is the number of data points that are in both cluster i of U and cluster j of V.
- a_i is the number of data points in cluster i of U.
- b_j is the number of data points in cluster j of V.
- n is the total number of data points.
  - Meaning: ARI quantifies the similarity between two clusterings by comparing agreements and disagreements, adjusting for chance.

3.4 Internal Cluster Validation Metrics

Without any external information the cluster evaluation could be done through internal cluster evaluation metrics [52]. The goodness of clusters is evaluated through only the information in the data [51]. The indexes we have used under internal cluster evaluation are Silhouette index [53], Davies-Bouldin index [54] and Calinski-Harabasz index [55].

3.4.1 Silhouette Index [53]

This measure assesses how compact and far apart clusters are from one another. It lies in the range of -1 to 1. Scores near 1 for silhouettes suggest well-defined clusters, scores near 0 suggest overlapping clusters, and negative values suggest incorrectly assigned data points. A higher silhouette score indicates an improvement in the quality of the clustering solution.

$$\text{Silhouette Index} = \frac{b_i - a_i}{\max(a_i, b_i)}$$

- a_i is the average distance from the i-th data point to other data points in the same cluster.
- b_i is the average distance from the i-th data point to data points in the nearest cluster (excluding its own cluster).

3.4.2 Davies-Bouldin Index [54]

This index measures a cluster's separation and compactness at the same time. A lower Davies-Bouldin Index indicates a higher quality clustering solution. On a range of 0 to positive infinity, compact and well-separated clusters are suggested, with 0 representing the optimal clustering condition. Higher values signify increased inter-cluster overlap and less distinctly defined cluster boundaries. The Davies-Bouldin Index is a particularly useful tool for evaluating the trade-off between cluster compactness and separation.

$$\text{Davies-Bouldin Index} = \frac{1}{k}\sum_{i=1}^{k}\max_{j\neq i}\left(\frac{\text{avg\_diam}(C_i)+\text{avg\_diam}(C_j)}{\text{distance}(c_i,c_j)}\right)$$

- k is the number of clusters.
- C_i and C_j are clusters.
- c_i and c_j are the centroids of clusters C_i and C_j.
- $\text{avg\_diam}(C_i)$ is the average distance between all pairs of points in cluster C_i.
- $\text{distance}(c_i, c_j)$ is the distance between the centroids c_i and c_j.

3.4.3 Calinski-Harabasz Index [55]

Measures the ratio of variance within a cluster to variance between clusters, which is known as the Calinski-Harabasz Index. Increased Calinski-Harabasz Index values are indicative of better-defined clusters. The range is 0 to positive infinity; higher values indicate more distinct clusters. This index is sensitive to the size, density, and shape of the clusters and serves as a thorough indicator of the quality of clustering. The Calinski-Harabasz Index is especially useful for simultaneously assessing the compactness and separation of clusters.

$$\text{Calinski-Harabasz Index} = \frac{\text{trace}(B_k)}{\text{trace}(W_k)\times(N-k)} \times \frac{N-k}{k-1}$$

- B_k is the between-cluster scatter matrix.
- W_k is the within-cluster scatter matrix.
- N is the total number of data points.
- k is the number of clusters.
  - Meaning: The Calinski-Harabasz Index evaluates the ratio of between-cluster variance to within-cluster variance.

3.5 Accuracy

The proportion of correctly classified data points in a clustering task is known as accuracy. On a range from 0 to 1, 1 represents absolute accuracy. Accuracy is a straightforward metric, but it should be used carefully because it doesn't always indicate successful clustering, especially when working with unbalanced datasets.

$$Accuracy = \frac{\text{Number of Correct Classifications}}{\text{Total Number of Data Points}}$$

3.6 Confusion Matrix [46]

An essential tool for assessing a classification model's efficacy is a confusion matrix. True positives (TP), false negatives (FN), false positives (FP), and true negatives (TN) are the four categories into which the model's predictions are divided for a thorough analysis.

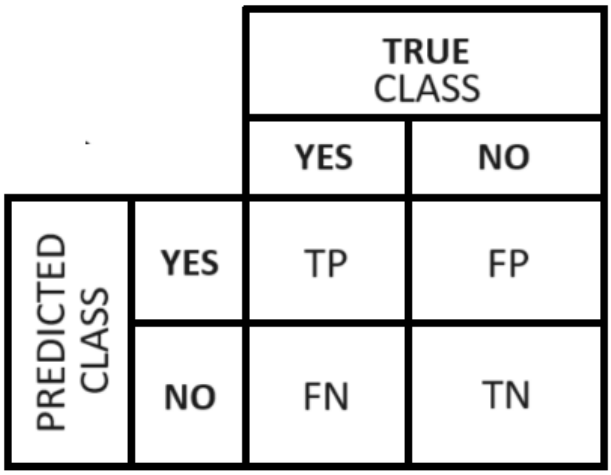


Fig 1. Confusion Matrix

These metrics and the confusion matrix show the model's performance and highlight potential areas for development. In a variety of classification scenarios, the confusion matrix visualization expedites interpretation and decision-making. Confusion Matrix gives 4 values – True Positive [46], True Negative [46], False Positive [46] and False Negative [46].

3.6.1 True Positive (TP) [46]

True Positives are cases that the model correctly identified as positive. When it comes to medical diagnosis, TP would be the instances in which the model accurately determines the ailment of a patient.

For example, emails that a spam detection system correctly identifies as spam are linked to TP.

3.6.2 True Negative (TN) [46]

True Negatives are instances that the model correctly classifies as negative. In the context of binary classification, TN represents the quantity of instances in which the model accurately detects the absence of the target condition.

For instance, transactions that in a fraud detection system are correctly classified as non-fraudulent despite having TN attached to them.

3.6.3 False Positive (FP) [46]

When a model mistakenly labels some events as positive when they are negative, this is known as a false positive. Another name for this is a Type I error.

For instance, FP will apply to medical tests if the results erroneously suggest the existence of a disease.

3.6.4 False Negative (FN) [46]

When a model mistakenly labels some occurrences as negative when they are positive, this is known as a false negative. This is also known as a Type II error.

Example: Occasionally, a security system may fail to identify a security breach.

The metrics calculated using the above parameter are sensitivity [46], precision [46], accuracy [46] and F1 Score [46].

3.6.5 Recall Sensitivity or True Positive Rate [46]

The percentage of actual positive examples that the model correctly detects is measured by sensitivity.

3.6.6 Precision, or positive predictive value [46]

The precision of a model is its ability to predict successful outcomes.

3.6.7 Accuracy [46]

Accuracy measures how well the model predicts things overall across all classes.

3.6.8 F1 Score [46]

The F1 Score is a helpful metric that finds a balance between sensitivity and precision in situations with unequal class sizes.

CHAPTER **4**

METHODOLOGY

The basic framework starts by generating the synthetic dataset according to the requirements provided by the clinician. Distance matrix is calculated for the top features after assigning weights to the features accordingly. This distance matrix is used to get the cluster of patients which are evaluated using the various evaluation metrics. The best cluster of patients is used, and the compliance is predicted by assigning the input features for the patient to the nearest cluster based on the Euclidean distance.

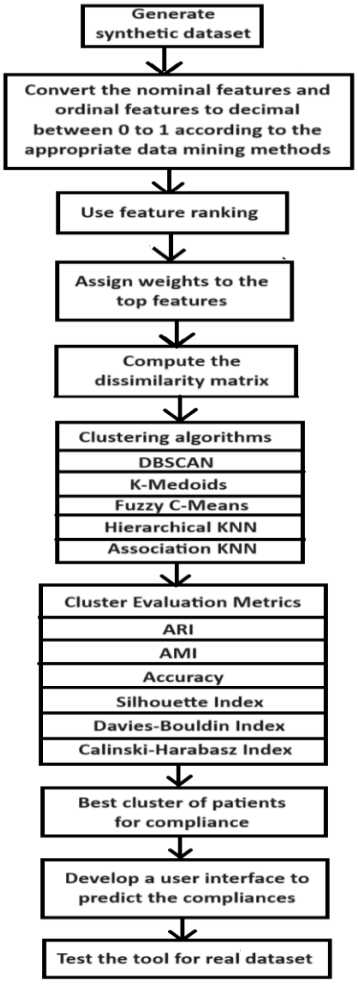


Fig 2. Basic framework for methodology

4.1 Generate Synthetic Dataset

The requirement for the dataset includes 79 columns in total where the first 78 columns include the personal and medical details of the patient. Personal details include the 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. Medical details include 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, average follow up interval, total follow ups, the frequency of follow ups, maximum compliance period and total duration of medication. The final compliance column tells us about how well the patients followed their treatment plan. It contains three results – Good, Satisfactory and Bad.

A python code was written to generate these requirements. Lists and dictionaries were used where a value was randomly selected, and the rows were filled consistently for each entry. The dataset was repeatedly checked and corrected until it was approved by clinicians.

4.2 Conversion of Dataset into Machine Readable Form

4.2.1 Nominal Variables [46]

For nominal variables the conversion to decimal value is done using the ratio of mismatches [46]. It is calculated using the formula:

$$d(i,j) = \frac{p - m}{p}$$

where m is the number of matches and p is the total number of attributes describing the objects.

4.2.2 Ordinal Variables [46]

For ordinal variables the conversion to decimal value is done using the rank method [46]. It is calculated using the formula:

$$z_{if} = \frac{r_{if} - 1}{M_f - 1}.$$

Where r is the rank assigned to each value in the column and M is the maximum rank in the column.

4.2.3 Numeric Variables [46]

The numerical values were left unaltered.

4.2.4 Ratios [46]

The ratios were left unaltered.

4.3 Normalization [46]

Normalisation ensures that numerical features are on a uniform scale, which is an essential component in preparing data for machine learning. Two often used methods are Z-score normalisation [46] and Min-Max normalisation [46]. Restricting features with higher proportions from controlling model training is intended to enhance equity. Normalisation improves the overall performance of the model by quickening the convergence of algorithms that are sensitive to feature scale. Normalisation is not necessary for all algorithms, though, and its use should be evaluated considering the machine learning technique being used.

4.3.1 Min-Max Normalisation [46]

Min-max normalisation is used to scale data to a certain range, typically [0, 1]. Each data point is transformed by putting the below formula on it.

$$X_{\text{norm}} = \frac{X_i - \min(X)}{\max(X) - \min(X)}$$

This keeps the dataset within a uniform scale and inhibits the dominance of characteristics with greater values.

4.3.2 Z-Score Normalisation [46]

In Z-Score normalisation scaling is done according to the standard deviation (σ) and the data is centred around the mean (μ). The formula for Z-Score normalisation is given below: -

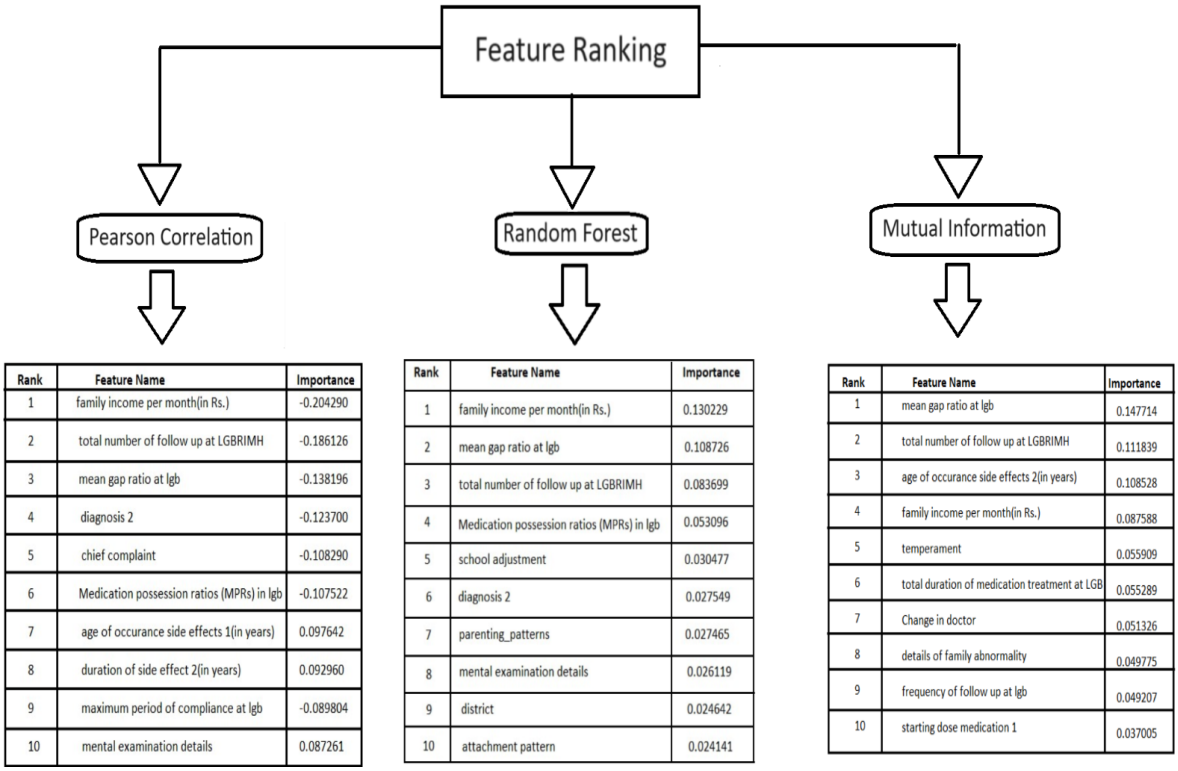
$$X_{\text{norm}} = \frac{X_i - \mu}{\sigma}$$

This procedure produces a distribution with a mean of 0 and a standard deviation of 1, which reduces the influence of outliers and facilitates feature comparison.

The dataset was then normalized to values between 0 and 1 using min-max normalization [46].

4.4 Feature Selection

Three feature selection measures – Random Forest [77], Pearson correlation [78] and Mutual Information [79] are used to rank the features of the dataset. For each similarity measure we get the corresponding rank for the feature. An average of all these ranks is taken. The features having a lower average rank is given a higher importance. The top 17 features are selected while the rest are dropped.



Rank	Feature Name	Importance
1	family income per month(in Rs.)	-0.204290
2	total number of follow up at LGBRIMH	-0.186126
3	mean gap ratio at lgb	-0.138196
4	diagnosis 2	-0.123700
5	chief complaint	-0.108290
6	Medication possession ratios (MPRs) in lgb	-0.107522
7	age of occurance side effects 1(in years)	0.097642
8	duration of side effect 2(in years)	0.092960
9	maximum period of compliance at lgb	-0.089804
10	mental examination details	0.087261

Rank	Feature Name	Importance
1	family income per month(in Rs.)	0.130229
2	mean gap ratio at lgb	0.108726
3	total number of follow up at LGBRIMH	0.083699
4	Medication possession ratios (MPRs) in lgb	0.053096
5	school adjustment	0.030477
6	diagnosis 2	0.027549
7	parenting_patterns	0.027465
8	mental examination details	0.026119
9	district	0.024642
10	attachment pattern	0.024141

Rank	Feature Name	Importance
1	mean gap ratio at lgb	0.147714
2	total number of follow up at LGBRIMH	0.111839
3	age of occurance side effects 2(in years)	0.108528
4	family income per month(in Rs.)	0.087588
5	temperament	0.055909
6	total duration of medication treatment at LGB	0.055289
7	Change in doctor	0.051326
8	details of family abnormality	0.049775
9	frequency of follow up at lgb	0.049207
10	starting dose medication 1	0.037005

Fig 3. Feature rank by Random Forest, Pearson Correlation and Mutual Information

Rank	Feature	Average_score
1	family income per month(in Rs.)	2
2	mean gap ratio at lgb	2
3	total number of follow up at LGBRIMH	2.333333333
4	Medication possession ratios (MPRs) in lgb	10
5	diagnosis 2	12.66666667
6	parenting_patterns	13.33333333
7	details of family abnormality	15
8	school adjustment	15
9	mental examination details	16.66666667
10	past medical history	17.33333333
11	total duration of medication treatment at LGB(in years)	18.33333333
12	temperament	18.66666667
13	frequency of follow up at lgb	22.33333333
14	chief complaint	22.66666667
15	maximum period of compliance at lgb	22.66666667
16	age of occurance side effects 2(in years)	23.33333333
17	cost of medication	23.33333333

Fig 4. Feature rank

4.5 Assignment of Weight to Features.

Using hit and trial method a weight is assigned to the features. Below are the tables for weight for all the algorithms we have used.

Rank	Feature	Weight
1	family income per month(in Rs.)	30
2	mean gap ratio at lgb	12
3	total number of follow up at LGBRIMH	15
4	Medication possession ratios (MPRs) in lgb	15
5	diagnosis 2	1
6	parenting_patterns	2
7	details of family abnormality	1
8	school adjustment	1
9	mental examination details	2
10	past medical history	1
11	total duration of medication treatment at LGB(in years)	1
12	temperament	1
13	frequency of follow up at lgb	1
14	chief complaint	0
15	maximum period of compliance at lgb	1
16	age of occurance side effects 2(in years)	1
17	cost of medication	1

Fig 5. Weights of feature for Hierarchical KNN

Rank	Feature	Weight
1	family income per month(in Rs.)	39
2	mean gap ratio at lgb	28
3	total number of follow up at LGBRIMH	23
4	Medication possession ratios (MPRs) in lgb	20
5	diagnosis 2	14
6	parenting_patterns	3
7	details of family abnormality	1
8	school adjustment	4
9	mental examination details	1
10	past medical history	3
11	total duration of medication treatment at LGB(in years)	3
12	temperament	1
13	frequency of follow up at lgb	1
14	chief complaint	0
15	maximum period of compliance at lgb	1
16	age of occurance side effects 2(in years)	1
17	cost of medication	1

Fig 6. Weights of feature for Association KNN

Rank	Feature	Weight
1	family income per month(in Rs.)	82
2	mean gap ratio at lgb	70
3	total number of follow up at LGBRIMH	70
4	Medication possession ratios (MPRs) in lgb	64
5	diagnosis 2	15
6	parenting_patterns	10
7	details of family abnormality	9
8	school adjustment	8
9	mental examination details	12
10	past medical history	0
11	total duration of medication treatment at LGB(in years)	6
12	temperament	15
13	frequency of follow up at lgb	2
14	chief complaint	0
15	maximum period of compliance at lgb	2
16	age of occurance side effects 2(in years)	0
17	cost of medication	3

Fig 7. Weights of features for DBSCAN

Rank	Feature	Weight
1	family income per month(in Rs.)	39
2	mean gap ratio at lgb	28
3	total number of follow up at LGBRIMH	23
4	Medication possession ratios (MPRs) in lgb	20
5	diagnosis 2	14
6	parenting_patterns	3
7	details of family abnormality	1
8	school adjustment	4
9	mental examination details	1
10	past medical history	3
11	total duration of medication treatment at LGB(in years)	3
12	temperament	1
13	frequency of follow up at lgb	1
14	chief complaint	0
15	maximum period of compliance at lgb	1
16	age of occurance side effects 2(in years)	1
17	cost of medication	1

Fig 8. Weights of features for K-Medoids

Rank	Feature	Weight
1	family income per month(in Rs.)	39
2	mean gap ratio at lgb	28
3	total number of follow up at LGBRIMH	23
4	Medication possession ratios (MPRs) in lgb	20
5	diagnosis 2	14
6	parenting_patterns	3
7	details of family abnormality	1
8	school adjustment	4
9	mental examination details	1
10	past medical history	3
11	total duration of medication treatment at LGB(in years)	3
12	temperament	1
13	frequency of follow up at lgb	1
14	chief complaint	0
15	maximum period of compliance at lgb	1
16	age of occurance side effects 2(in years)	1
17	cost of medication	1

Fig 9. Weights of features for Fuzzy C-Means

4.6 Calculate Distance Matrix

For each algorithm the values of each feature given in a column are multiplied by the weights assigned to the feature for that algorithm. Individual dissimilarity matrix for each feature and for each algorithm is calculated using Manhattan distance [46]. The final dissimilarity matrix for each algorithm is calculated by summing up all the dissimilarity matrix for each feature for that algorithm.

4.7 Clustering

For clustering we use the distance matrix calculated, as our input. We have performed clustering using a total of 5 algorithms - DBSCAN [44], K-Medoids [45], Fuzzy C-Means [47], Association KNN and Hierarchical KNN. We have proposed two algorithms – Association KNN and Hierarchical KNN to perform clustering and compared the performance using Silhouette index [53], Davies-Bouldin index [54], Calinski-Harabasz index [55], Adjusted Rand Index (ARI) [49] and Adjusted Mutual Information (AMI) [50] and accuracy to the performance of the former three clustering algorithms.

4.7.1 Hierarchical-KNN

The first algorithm, Hierarchical KNN algorithm is based on the concept of KNN [80] and agglomerative clustering [46]. The Hierarchical KNN, operates by considering the k nearest points for each individual point within the same cluster. It follows an agglomerative approach, merging clusters until the average distance between two clusters remains below or equal to a specified threshold. This algorithm employs a hard clustering strategy, assigning each point to a single cluster.

4.7.1.1 Algorithm

1) Input the value of k, threshold and outlier threshold

2) Compute the distance matrix

3) Store the computed distance matrix in another variable

4) Sort each row of the distance matrix in ascending order with indexes

5) Calculate the average cluster distance for the points with its k nearest neighbour, which are the first k values of the sorted distance matrix and store it

6) Sort the average cluster distances with its position

7) Create a dummy list which consists of all the k nearest points along with the points it is near to

8) Remove similar looking clusters

9) Calculate the average cluster distance between 2 clusters and store it in a 2d list

10) Get the row number and column number from the distance matrix where value is below threshold

11) If no such value found which is below threshold, then go to step 15

12) Merge both clusters and remove similar points

13) Remove the clusters which were merged

14) Go to step 8

15) Sort the clusters we get to start from largest cluster

16) Assign clusterid to points according to cluster

17)If a cluster contains a smaller number of points than the outlier threshold then labels the cluster as an outlier

18) Exit

4.7.1.2 Time Complexity

1)Computing the distance matrix: The nested for loops iterate over the points in the dataset to compute the Euclidean distance between each pair of points. The time complexity of this step is O(n^2), where n is the number of points in the dataset.

2)Sorting the distance matrix: Each row is sorted which takes n log n time and we do this for all n points and thus the total time complexity comes out to be O (n^2 log n)

3)Calculating average cluster distances: The algorithm calculates the average cluster distance for each point based on the k nearest neighbours. This step involves iterating over each point and summing up the distances, resulting in a time complexity of O (n * k).

4)Sorting the average cluster distances: The algorithm sorts the average cluster distances in ascending order. Sorting a list of n elements takes O (n log n) time.

5)Calculate average cluster distance: This takes around O(l*n^4) time complexity where l is number of iterations of while loop

6)To check for the threshold condition: This takes around (l * n^2) time complexity where l is the number of iterations of while loop.

7)To remove the clusters and insert new cluster: This takes around (l * n) time complexity where l is the number of iterations of while loop.

Thus, overall time complexity according to dominant terms is O(l*n^4)

Thus, hierarchical knn takes too much time complexity and this is one of its disadvantages. Thus, in our future work we propose to reduce its time complexity.

4.7.2 Association-KNN

The Association KNN algorithm is based on the concept of KNN and the associations/connections which the point currently under consideration and the points which belong in a cluster [80,81] have. This algorithm analyses the k nearest points for every point within a cluster. It initiates by assigning points to clusters based on the cluster with the lowest average intracluster distance. Subsequently, it incorporates additional points into the cluster if they satisfy two conditions: they fall within a defined threshold distance and have not been assigned to any other cluster. This process continues until all points are allocated to a cluster. Like Hierarchical KNN, the Association KNN algorithm is also a hard clustering technique where each point is assigned exclusively to one cluster.

4.7.2.1 Algorithm

1) Input the value of k and threshold

2) Compute the distance matrix

3) Store the computed distance matrix in another variable

4) Sort each row of the distance matrix in ascending order with indexes

5) Calculate the average cluster distance for the points with its k nearest neighbour, which are the first k values of the sorted distance matrix and store it

6) Sort the average cluster distances with its position

7) Assign default clusterid of 1 to all the points which form the smallest cluster

8) Initialize clusternumber to 1

9) Repeat step 10 to step 15 till clusterid != clusterid1

10) Copy clusterid values to clusterid1

11) Make a cluster for the current clusternumber under observation using clusterid

12) Calculate the average cluster distance of the point with the current cluster

13) If the point under observation is within the threshold value and not assigned to any other cluster then append point to the current cluster under observation using clusterid1

14)Exchange the values of clusterid and clusterid1

15) Go to step 9

16) Go to step 20 if all points have been checked once

17) Now move to next smallest cluster and continue till we find an unassigned point

18) Make the next cluster to start observation with which is the current chosen point and all its k nearest point which are unassigned

19) Go to step 9

20)If a cluster contains a smaller number of points than the outlier threshold then labels the cluster as an outlier

21) Exit

4.7.2.2 Time Complexity

The time complexity of the provided algorithm can be analysed as follows:

1) Computing the distance matrix: The nested for loops iterate over the points in the dataset to compute the Euclidean distance between each pair of points. The time complexity of this step is O(n^2), where n is the number of points in the dataset.

2) Sorting the distance matrix: Each row is sorted which takes n log n time and we do this for all n points and thus the total time complexity comes out to be O (n^2 log n)

3) Calculating average cluster distances: The algorithm calculates the average cluster distance for each point based on the k nearest neighbours. This step involves iterating over each point and summing up the distances, resulting in a time complexity of O(n*k)

4) Sorting the average cluster distances: The algorithm sorts the average cluster distances in ascending order. Sorting a list of n elements takes O (n log n) time.

5) Assign default value 1 to the clusterid - Takes O(k) time

6) Appending points in cluster to checklist assigned to current cluster takes O(n*l*m) where l is number of loop iterations for inner loop and m is the number of iterations for outside loop

7) Calculate average cluster distance of all points to the current cluster - Takes O (n^2 * l * m) where l is number of loop iterations for inner loop and m is the number of iterations for outside loop

8) To assign points less than threshold and not assigned to any cluster to cluster is O(n*l*m) time where l is number of loop iterations for inner loop and m is the number of iterations for outside loop

9) To iterate till all points are traversed-Takes O(n) time

10) Make the next cluster to start observation with which is the current chosen point and all its k nearest point which are unassigned takes O(k*m) time complexity where m is number of iterations of outer loop.

Thus, worst case time complexity is O(n^2*l*m) + O (n^2 * log n)

Thus, improving time complexity and making the algorithm better can be reserved for future work.

CHAPTER 5

DEVELOPMENT OF WEBTOOL

5.1 Requirements

Software Requirements

i) Language Used: Python,HTML,CSS,JavaScript

ii) Package Used: Pandas (Appendix - 2), NumPy (Appendix - 3), SciPy (Appendix - 4), Flask (Appendix - 1)

iii)OS used: Windows 10

iv)Software Used: Jupyter Notebook, VSCode

5.2 Method for development

Hierarchical KNN (k – 50, threshold – 5, outlier threshold – 10) is used to get the cluster of patients. Each cluster is assigned a label which is the maximum number of compliance values included in the cluster. For Hierarchical KNN we get a total of 13 clusters.

The user is asked to enter the various parameters. Each entered parameter is converted to a similar form which we obtained for calculating the individual distance matrix for the feature. Euclidean distance [46] of the entered parameters is calculated to all the 13 clusters. The new point is assigned to the cluster having the minimum cluster distance to the point.

The prediction is returned as the label which is assigned to the cluster.

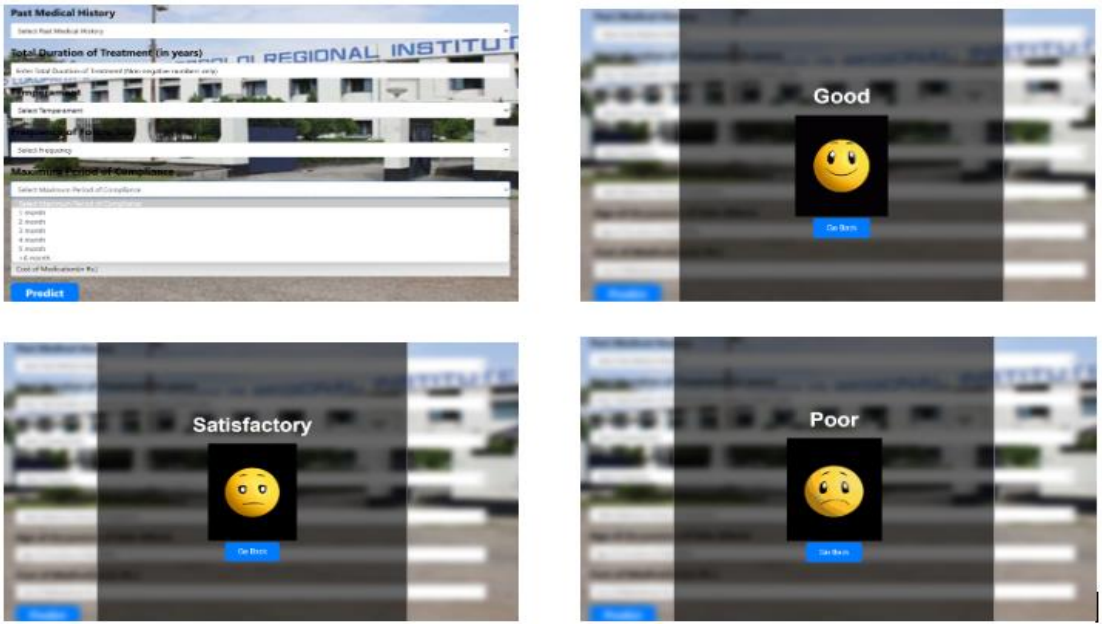


Fig 10. Screenshots of webtool

CHAPTER **6**

RESULTS AND DISCUSSIONS

6.1 Results on Synthetic Datasets

ALGO	K	THRES HOLD	EPS	MIN POINT	OUTLIER THRESHOLD	ACC	ARI	AMI	SH	DB	CH
H-KNN	50	5	-	-	10	68.3%	0.24	0.26	0.48	3.7	60.2
A-KNN	50	0.5	-	-	5	65.4%	0.23	0.25	0.08	2.3	47.23
K-MEDOID	3	-	-	-	-	58%	0.23	0.30	0.11	2.18	60.01
FUZZY C-MEAN	3	-	-	-	-	45%	0.1	0.1	0.05	3.23	23.78
DBSCAN	-	-	30	4	4	49.7%	0.2	0.11	0.1	3.05	24.79

Fig 11. Results for clustering on synthetic dataset

Abbreviations,

ACC – Accuracy

ARI – Adjusted Rand Index

AMI – Adjusted Mutual Information

SH – Silhouette Index

DB – Davies-Bouldin Index

CH – Calinski-Harabasz Index

Hierarchical KNN (k – 50, threshold – 5, outlier threshold – 10) was chosen as the clustering algorithm for implementing the web tool, as it was giving the highest Silhouette index [53], Davies-Bouldin index [54], Calinski-Harabasz index [55], accuracy and Adjusted Rand Index (ARI) [49]. It was giving second highest value for Adjusted Mutual Information (AMI) [50] compared to all other clustering algorithms we had used.

DBSCAN was found to be giving low accuracy as it only divided the datapoints into 2 clusters, while our final compliance column required at least three clusters for – Good, Satisfactory and Poor. The density-based algorithms are found to give low values compared to non-density-based algorithms.

On increasing the outlier threshold, we could see an increase in accuracy and all other indexes, but the outlier count was not raised beyond 10, as higher value eliminates a greater number of clusters, and we were left with only a few datapoints.

6.1.1 Confusion Matrix [46]

		Actual		
		Good	Satisfactory	Poor
Predicted	Good	49	35	0
	Satisfactory	28	118	26
	Poor	0	25	79

Fig 12. Confusion Matrix for Synthetic Data

6.1.1.1 Confusion Matrix for Good Class

		TRUE GOOD	
		YES	NO
PREDICTED GOOD	YES	49	35
	NO	28	248

Fig 13. Confusion Matrix for Good Class for Synthetic Data

The values for different validity measures for synthetic data for class good are

Sensitivity: 0.6364, Precision: 0.5833, Accuracy: 0.8250 and F1 Score: 0.6087

6.1.1.2 Confusion Matrix for Satisfactory Class

		TRUE SATISFACTORY	
		YES	NO
PREDICTED SATISFACTORY	YES	118	54
	NO	60	128

Fig 14. Confusion Matrix for Satisfactory Class for Synthetic Data

The values for different validity measures for synthetic data for class satisfactory are Sensitivity: 0.6629, Precision: 0.6860, Accuracy: 0.6833 and F1 Score: 0.6743

6.1.1.3 Confusion Matrix for Poor Class

		TRUE POOR	
		YES	NO
PREDICTED POOR	YES	79	25
	NO	26	230

Fig 15. Confusion Matrix for Poor Class for Synthetic Data

The values for different validity measures for synthetic data for class bad are

Sensitivity: 0.7524, Precision: 0.7596, Accuracy: 0.8583 and F1 Score: 0.7560

6.1.1.4 Conclusion for Confusion Matrix for Synthetic Data

The values for different validity measures for synthetic data are

Weighted Sensitivity:0.6825, Weighted Precision:0.6832, Weighted Accuracy: 0.7669 and Weighted F1 Score: 0.6825

A good, weighted accuracy suggests that the predictions made by the model are often correct. We also see high values for Good and Poor Class, stating that the model is better for classifying these instances as compared to Satisfactory cases, For the other metrics we see moderate values stating moderate performances by model.

6.2 Results on Actual Datasets

6.2.1 Confusion Matrix [46]

		Actual		
		Good	Satisfactory	Poor
Predicted	Good	8	0	0
	Satisfactory	5	7	5
	Poor	1	0	2

Fig 16. Confusion Matrix for Actual Data

6.2.1.1 Confusion Matrix for Good Class

		TRUE GOOD	
		YES	NO
PREDICTED GOOD	YES	8	0
	NO	6	14

Fig 17. Confusion Matrix for Good Class for Actual Data

The values for different validity measures for actual data for class good are

Sensitivity: 0.5714, Precision: 1.0000, Accuracy: 0.7857 and F1 Score: 0.7273

6.2.1.2 Confusion Matrix for Satisfactory Class

		TRUE SATISFACTORY	
		YES	NO
PREDICTED SATISFACTORY	YES	7	10
	NO	0	11

Fig 18. Confusion Matrix for Satisfactory Class for Actual Data

The values for different validity measures for actual data for class satisfactory are

Sensitivity: 1.0000, Precision: 0.4118, Accuracy: 0.6429 and F1 Score: 0.5833

6.2.1.3 Confusion Matrix for Poor Class

		TRUE POOR	
		YES	NO
PREDICTED POOR	YES	2	1
	NO	5	20

Fig 19. Confusion Matrix for Poor Class for Actual Data

The values for different validity measures for actual data for class poor are

Sensitivity: 0.2857, Precision: 0.6667, Accuracy: 0.7857 and F1 Score: 0.4000

6.2.1.4 Conclusion for Confusion Matrix for Actual Data

The values for different validity measures for actual data are

Weighted Sensitivity:0.8010, Weighted Precision:0.6071, Weighted Accuracy: 0.6990 and Weighted F1 Score: 0.6048

A moderate, weighted accuracy suggests moderate predictions made by the model. An overall good value for sensitivity suggests that the model is very good for identifying the actual positive instances as positive. We observe poor sensitivity for poor class stating that the model performs poorly for predicting poor class. A perfect value for sensitivity for satisfactory class tells that the model does not misclassify any true satisfactory instance. A low value for precision for satisfactory class states that when model predicts satisfactory it is mostly false. A perfect value for precision for good compliance means when the model predicts good it is always correct. Other parameters being quite moderate suggest the model being average for predictions.

CHAPTER 7

CONCLUSION AND FUTURE WORK

7.1 Conclusion

7.1 .1 Clustering on synthetic data

Thus, we can finally infer that Hierarchical KNN gave the best results overall among all the algorithms we had used. The results were better compared to Association KNN, DBSCAN, k medoids and fuzzy c means. The parameters we had used for Hierarchical KNN were k – 50, threshold – 5 and outlier threshold – 10. It was giving the highest Silhouette index, Davies-Bouldin index, Calinski-Harabasz index, Adjusted Rand Index (ARI) and accuracy. It was giving second highest values for Adjusted Mutual Information (AMI).

Association KNN was giving overall the second-best results among all the algorithms we had used which included Hierarchical KNN, DBSCAN, k medoids and fuzzy c means. The parameters we had used for Association KNN were k – 50, threshold – 0.5 and outlier threshold – 5. It was giving the second highest accuracy and Adjusted Rand Index (ARI). It was giving the third highest Calinski-Harabasz index and Adjusted Mutual Information (AMI). It was giving the fourth highest Silhouette index and Davies-Bouldin index.

7.1.2 Advantages of using Hierarchical KNN

The highest value for accuracy on using Hierarchical KNN (k – 50, threshold – 5 and outlier threshold – 10) suggests that Hierarchical KNN is the best algorithm we have used among Association KNN, DBSCAN, k medoids and fuzzy c means to correctly identify the compliance, as the predictions made through it were most correct of all. The highest values for Silhouette index, Davies-Bouldin index, Calinski-Harabasz index, Adjusted Rand Index (ARI) and second highest values for Adjusted Mutual Information suggests the best cluster quality and separation out of all algorithms we have used.

7.1.2 Disadvantages of using Hierarchical KNN

It is dependent on the features used, weight for the features used and the parameters used for Hierarchical KNN. The major disadvantage was that hierarchical KNN had the highest time complexity. As the algorithm was dependent on the various parameters, the performance and

the results were thus completely dependent on weight, feature subset and parameters for Hierarchical KNN. The results of all the indices used which are Silhouette index, Davies-Bouldin index, Calinski-Harabasz index, Adjusted Rand Index (ARI) and Adjusted Mutual Information (AMI) were quite moderate, although best among all algorithms we had used. The accuracy too was moderate. This suggests that we have scope for improvement in future.

7.2 Future Work

In future we plan to incorporate other algorithms in our webtool which would include DBSCAN, Association KNN, k medoids and fuzzy c means. We would also include other new algorithms which would test in future. We will use supervised methods of learning and compare the various evaluation metrics for all algorithms and find the best result.

We plan to extend our development to a mobile app. This app would be provided to clinicians to aid their work. We would work on actual data which would be provided by the clinicians at LGBRIMH.

CHAPTER **8**

ETHICAL CLEARANCE

The project is under scientific and ethical review at LGBRIMH, Tezpur and clearance will be soon granted.

REFERENCES

[1] De las Cuevas, C. (2011). "Towards a clarification of terminology in medicine taking behavior: Compliance, adherence and concordance are related although different terms with different uses." *Current Clinical Pharmacology,6*(2), 74–77.

[2] World Health Organization. (2003). "Adherence to long-term therapies."

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

[4] DiMatteo, M. R., & Lepper, H. S., and Croghan, T. W. (2000). "Depression is a risk factor for noncompliance with medical treatment." *160*, 2101.

[5] 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.* New York: Springer, 2010, pp. 33–46.

[6] Berg, J.S., Dischler, J., Wagner, D.J., Raia, J.J., Palmer-Shevlin, N. "Medication compliance: A healthcare problem." *Ann Pharmacother, 27*, S1–S24, 1993.[7] El-Rachidi, S., LaRochelle, J.M., Morgan, J.A. "Pharmacists and Pediatric Medication Adherence: Bridging the Gap." *Hosp Pharm, 52*(2), 124-131, 2017. doi: 10.1310/hpj5202-124. PMID: 28321139; PMCID: PMC5345910.

[7] El-Rachidi, S., LaRochelle, J. M., & Morgan, J. A. (2017). Pharmacists and Pediatric Medication Adherence: Bridging the Gap. *Hosp Pharm*, 52(2), 124-131. https://doi.org/10.1310/hpj5202-124

[8] Dawood, T., Mohamed, I., Mohamed, I. and Palaian, S. (2010). "Medication compliance among children." *World Journal of Pediatrics : WJP, 6*, 200-2.10.1007/s12519-010-0218-8.

[9] Paula Gardiner, M.D. and Lana, D. "Promoting medication adherence in children." *Am Fam Physician, 74*, 793-798, 800, 2006.

[10] Lask, B. "Motivating children and adolescents to improve adherence." *J Pediatr, 143*, 430-433, 2003.

[11] Lieberman, D.A. "Management of chronic pediatric diseases with interactive health games: theory and research findings." *J Ambul Care Manage, 24*, 26-38, 2001.

[12] Rapoff, M.A., Belmont, J.M., Lindsley, C.B. and Olson, N.Y. "Electronically monitored adherence to medications by newly diagnosed patients with juvenile rheumatoid arthritis." *Arthritis Rheum, 53*, 905-910, 2005.

[13] Side1, V., Berger, J.L., Lisi-Fazio, D., Kleinman, K., Wenston, J. and Thomas, C. et al. "Controlled study of the impact of educational home visits by pharmacists to high-risk older patients." *Community Health, 15*, 163-174, 1990.

[14] Maan, C.G., Munnawar H.M.S., Heramani, N. and Lenin, R.K. "Factors Affecting Non-Compliance among Psychiatric Patients in the Regional Institute of Medical Sciences, Imphal." *Volume 5, Issue 1 (January 2015), PP. 01-07.*

[15] Reilly, E.L.; Wilson, W.P.; and McClinton, H.K. "Clinical characteristics and medication history of schizophrenics readmitted to the hospital." *International Journal of Neuropsychiatry, 1967; 39*: 85-90.

[16] Victoria, O., Mohsen, Y., Mohammad, Y. and Mahshid, N. "Noncompliance and its causes resulting in psychiatric readmissions." *Iran J Psychiatry, 2008; 3*: 37-42.

[17] Selen, Y., Wertheimer, A. and Dublin, W. "Demographical factors affecting patient compliance to medications in an outpatient psychiatric clinic: A preliminary study." *FABAD J. Pharm. Sci., 2003; 28*: 77-84.

[18] Nichols-English, G. and Poirier, S. "Optimizing adherence to pharmaceutical care plans." *J Am Pharm Assoc., 2000; 40*: 475-85.

[19] Rodenhausser, P., Schwenker, C.E., and Khamis, H.J. "Factors related to drug treatment refusal in a forensic hospital." *Hospital Community Psychiatry, 1987; 38*: 631-637.

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

[21] Carlson, G.A., Kotov, R. and Chang, S.W. et al. "Early determinants of four-year clinical outcomes in bipolar disorder with psychosis." *Bipolar Disord. 2012;14(1):19–30.* [PMC free article] [PubMed] [Google Scholar]

[22] Olino, T.M., Seeley, J.R. and Lewinsohn, P.M. "Conduct disorder and psychosocial outcomes at age 30: early adult psychopathology as a potential mediator." *J Abnorm Child Psychol. 2010;38(8):1139–49.* [PMC free article] [PubMed] [Google Scholar]

[23] Birmaher, B. and Axelson, D. "Course and outcome of bipolar spectrum disorder in children and adolescents: a review of the existing literature." *Dev Psychopathol. 2006;18(4):1023–35.* [PubMed] [Google Scholar]

[24] Klein, D.N., Shankman, S.A. and Rose, S. "Dysthymic disorder and double depression: prediction of 10-year course trajectories and outcomes." *J Psychiatr Res. 2008;42(5):408–15.* [PMC free article] [PubMed] [Google Scholar]

[25] Brezo, J., Paris, J. and Barker, E.D. et al. "Natural history of suicidal behaviors in a population-based sample of young adults." *Psychol Med. 2007;37(11):1563–74.* [PubMed] [Google Scholar]

[26] Hastie, T., Tibshirani, R. and Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition.* New York, NY: Springer Science & Business Media (2009).

[27] Jordan, M.I. and Mitchell, T.M. "Machine learning: Trends, perspectives, and prospects." *Science. (2015) 349:255–60.* doi: 10.1126/science.aaa8415

[28] Müller, A.C. and Guido, S. *Introduction to Machine Learning with Python: A Guide for Data Scientists.* Sebastopol, CA: O'Reilly Media, Inc. (2016).

[29] McCorduck, P. *Machines Who Think: A Personal Inquiry into the History and Prospects of Artificial Intelligence.* Canada: CRC Press (2004). doi: 10.1201/9780429258985

[30] Sarker, I.H., Machine Learning: Algorithms, Real-World Applications and Research Directions. SN Computer Sci. 2021;2(3):160. doi: 10.1007/s42979-021-00592-x. Epub 2021 Mar 22. PMID: 33778771; PMCID: PMC7983091.

[31] Juan, J. and Russell, G. "An Introduction to Machine Learning Approaches for Biomedical Research." *Front. Med., 16 December 2021.* Volume 8 - 2021 | https://doi.org/10.3389/fmed.2021.771607

[32] James, G., Witten, D., Hastie, T. and Tibshirani R., eds. *An Introduction to Statistical Learning: With Applications in R.* New York: Springer; 2013.

[33] Hastie, T., Tibshirani, R. and Friedman, J.H. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* 2nd ed. New York, NY: Springer; 2009.

[34] Molnar, C. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable.* Morrisville, North Carolina: Lulu Press, Inc.; 2019. https://christophm.github.io/interpretableml-book/

[35] NVIDIA Blog: "Supervised Vs. Unsupervised Learning." *The Official NVIDIA Blog.* https://blogs.nvidia.com/blog/2018/08/02/supervised-unsupervised-learning/. Published August 2, 2018. Accessed October 24, 2019.

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

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

[38] Bohlmann, A., Mostafa, J. and Kumar M. "Machine Learning and Medication Adherence: Scoping Review." *JMIRx Med 2021;2(4):e26993.*

[39] Piette, J.D., Sussman, J.B., Pfeiffer, P.N., Silveira, M.J., Singh, S. and Lavieri, M.S. "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." *J Med Internet Res 2013 Jul 05;15(7):e118.*

[40] Chaix, B., Bibault, J., Pienkowski, A., Delamon, G., Guillemassé, A. and Nectoux, P. et al. "When chatbots meet patients: one-year prospective study of conversations between patients with breast cancer and a chatbot." *JMIR Cancer 2019 May 02;5(1):e12856.*

[41] Vidya, K., Anzar, A., Li, Z., Lei, G., Shaolei, F., Vijay, Y. and Isaac, R. "Accuracy of machine learning-based prediction of medication adherence in clinical research." *Psychiatry Research, Volume 294, 2020, 113558.*

[42] Xu, D. and Tian, Y. "A Comprehensive Survey of Clustering Algorithms." *Ann. Data. Sci. 2, 165–193 (2015).* https://doi.org/10.1007/s40745-015-0040-1

[43] Jain, A. and Dubes, R. (1988) "Algorithms for clustering data." *Prentice-Hall, Inc, Upper Saddle River.*

[44] Martin, E., Hans-Peter, K., Jörg, S. and Xiaowei, X. "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise." *Institute for Computer Science, University of Munich.*

[45] Jin, X. and Han, J. (2011). "K-Medoids Clustering." *In: Sammut, C., Webb, G.I. (eds) Encyclopedia of Machine Learning. Springer, Boston, MA.* https://doi.org/10.1007/978-0-387-30164-8_426

[46] Han, J., Kamber, M. and Pei, J. "Data mining: Concepts and techniques." *Data mining: concepts and techniques (2012), 10.1016/C2009-0-61819-5.*

[47] James, C., Bezdek, Robert, Ehrlich and William, Full. "FCM: The fuzzy c-means clustering algorithm." *Computers & Geosciences, Volume 10, Issues 2–3, 1984, Pages 191-203.* https://doi.org/10.1016/0098-3004(84)90020-7

[48] Yang, L., James, C. Bezdek, Romano, S., Xuan, V., Jeffrey, C. and James, B. "Ground truth bias in external cluster validity indices." *Pattern Recognition, Volume 65, 2017, Pages 58-70.* https://doi.org/10.1016/j.patcog.2016.12.003

[49] Warrens, M.J. and van der Hoef, H. "Understanding the Adjusted Rand Index and Other Partition Comparison Indices Based on Counting Object Pairs." *J Classif 39, 487–509 (2022).*

[50] Nguyen, X.V. , Epps, J., and Bailey, J. "Information theoretic measures for clusterings comparison: is a correction for chance necessary?" *In ICML, 2009.*

[51] Liu, Y., Li, Z., Xiong, H., Gao, X. and Wu, J. "Understanding of Internal Clustering Validation Measures." *2010 IEEE International Conference on Data Mining, Sydney, NSW, Australia, 2010, pp. 911-916.* doi: 10.1109/ICDM.2010.35

[52] Tan, P.N., Steinbach, M., and Kumar, V. *Introduction to Data Mining.* USA: Addison-Wesley Longman, Inc., 2005.

[53] Rousseeuw, P. "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *J. Computer. Appl. Math., vol. 20, no. 1, pp. 53–65, 1987.*

[54] Davies, D. and Bouldin, D. "A cluster separation measure," *IEEE PAMI, vol. 1, no. 2, pp. 224–227, 1979.*

[55] Calinski, T. and Harabasz, J. "A dendrite method for cluster analysis," *Comm. in Statistics, vol. 3, no. 1, pp. 1–27, 1974.*

[56] Blum, A.L. et al,. "Selection of relevant features and examples in machine learning." *Artif. Intell. (1997).*

[57] Sun, Z.L. et al. "Extracting nonlinear features for multispectral images by FCMC and KPCA." *Digit. Signal Process. (2005).*

[58] Khotanzad, A. et al. "Rotation invariant image recognition using features selected via a systematic method." *Pattern Recognit. (1990).*

[59] Goltsev, A. et al. "Investigation of efficient features for image recognition by neural networks." *Neural Netw. (2012).*

[60] Rashedi, E. et al. "A simultaneous feature adaptation and feature selection method for content-based image retrieval systems." *Knowl.-Based Syst. (2013).*

[61] Amiri, F. et al. "Mutual information-based feature selection for intrusion detection systems." *J. Netw. Comput. Appl. (2011).*

[62] Ang, J.C. ,Mirzal, A., Haron, H. and Hamed, H. "Supervised, Unsupervised, and Semi-Supervised Feature Selection: A Review on Gene Selection." *IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 13, no. 5, pp. 971-989, 1 September 2016.* doi: 10.1109/TCBB.2015.2478454

[63] Chen, L., Huang, R. And Huang, W. "Graph-based semi-supervised weighted band selection for classification of hyperspectral data." *International Conference on Audio Language and Image Processing (ICALIP), 2010, pp. 1123-1126.*

[64] Yang, M., Chen, Y.J. and Ji, G.L. "Semi_Fisher Score: A semi-supervised method for feature selection." *International Conference on Machine Learning and Cybernetics (ICMLC), 2010, pp. 527-532.*

[65] Sunzhong, L., Jiang, H., Zhao, L., Wang, D. and Fan, M. "Manifold based Fisher method for semi-supervised feature selection." *International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), 2013, pp. 664-668.*

[66] Yang, W., Hou, C. and Wu, Y. "A semi-supervised method for feature selection." *International Conference on Computational and Information Sciences (ICCIS), 2011, pp. 329-332.*

[67] Liu, Y., Nie, F., Wu, J. and Chen, L. "Efficient semi-supervised feature selection with noise insensitive trace ratio criterion." *Neurocomputing 105 (2013) 12-18.*

[68] Liu,Y., Nie,F., Wu,J. and Chen,L. "Semi-supervised feature selection based on label propagation and subset selection." *International Conference on Computer and Information Application (ICCIA), 2010, pp. 293-296.*

[69] Cheng, H., Deng, W., Fu, W., Wang, Y. and Qin, Y. "Graph-based semi-supervised feature selection with application to automatic spam image identification." *Computer Science for Environmental Engineering and EcoInformatics (2011) 259-264.*

[70] Zhao, J., Lu, K. and He, X. "Locality sensitive semi-supervised feature selection." *Neurocomputing 71 (2008) 1842-1849.*

[71] Doquire, G. and Verleysen, M. "Graph Laplacian for semi-supervised feature selection in regression problems." *International Work-Conference on Artificial Neural Networks, 2011, pp. 248-255.*

[72] Doquire, G. and Verleysen, M. "A graph Laplacian based approach to semi-supervised feature selection for regression problems." *Neurocomputing 121 (2013) 5-13.*

[73] Kalakech, M., Biela, P., Macaire, L. and Hamad, D. "Constraint scores for semi-supervised feature selection: A comparative study." *Pattern Recognition Letters 32 (2011) 656-665.*

[74] Benabdeslem, K. and Hindawi, M. "Constrained laplacian score for semi-supervised feature selection." *Machine Learning and Knowledge Discovery in Databases (2011) 204-218.*

[75] Zhang, D., Chen, S. and Zhou, Z.H. "Constraint Score: A new filter method for feature selection with pairwise constraints." *Pattern Recognition 41 (2008) 1440-1451.*

[76] Bishop, C.M. *Neural networks for pattern recognition, Oxford university press, 1995.*

[77] University of Liège, Faculty of Applied Sciences, Department of Electrical Engineering & Computer Science. *PhD dissertation: UNDERSTANDING RANDOM FORESTS: from theory to practice by Gilles Louppe.*

[78] Samuels, P. & Gilchrist, M. (2014). *Pearson Correlation.*

[79] McClure, S. "A Deep Conceptual Guide to Mutual Information Embracing the "Correlation of the 21st Century."" *The Startup.*

[80] Fix, E. and Hodges, J.L. (1951): *an important contribution to nonparametric discriminant analysis and density estimations.*

[81] Bora, G. and Sarmah, R. *FLBC: A Fuzzy Link based Clustering approach for gene expression data.*

CHAPTER 10

APPENDIX

1) Flask: A popular Python web framework known for its versatility and ease of use is called Flask. It simplifies the handling of user requests, routing, and templating, which facilitates the creation of online applications. Flask is a fantastic tool for developing small-to medium-sized web apps and RESTful APIs.

2) Pandas: Pandas is a powerful Python data analysis and manipulation toolbox. It provides data structures like Data Frames, which simplify working with structured data. Pandas is an excellent tool for jobs involving data transformation, cleaning, and exploration. It offers a wide range of ways for handling and analysing tabular data. Because of its effectiveness and adaptability, it is a favoured tool in workflows related to data science and analysis.

3) NumPy: A core library for Python numerical computing is called NumPy. It presents strong array objects to facilitate effective operations on big datasets. For scientific and mathematical computations, NumPy is indispensable. It provides an extensive set of functions for linear algebra, statistical analysis, array manipulation, and other uses. The speed and effectiveness of Python's numerical operations are facilitated by its high-performance capabilities.

4) SciPy: An extension package called SciPy gives scientific computing users access to additional features and tools. It includes modules for signal and image processing, optimization, statistical analysis, and more. SciPy is extensively used in scientific research and engineering applications because it offers a large range of functions for scientific computing in Python and supports NumPy with a strong ecosystem.

Project report check 2

ORIGINALITY REPORT

9%	6%	6%	2%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.ncbi.nlm.nih.gov Internet Source	1%
2	Vidya Koesmahargyo, Anzar Abbas, Li Zhang, Lei Guan, Shaolei Feng, Vijay Yadav, Isaac Galatzer-Levy. "Accuracy of machine learning-based prediction of medication adherence in clinical research", Cold Spring Harbor Laboratory, 2020 Publication	1%
3	Nelson E. Ordoñez-Guillen, JoseLuis Gonzalez-Compean, Ivan Lopez-Arevalo, Miguel Contreras-Murillo, Edwin Aldana-Bobadilla. "Machine learning based study for the classification of Type 2 diabetes mellitus subtypes", BioData Mining, 2023 Publication	1%
4	"Verification, Model Checking, and Abstract Interpretation", Springer Science and Business Media LLC, 2011 Publication	<1%

5	Juan Jovel, Russell Greiner. "An Introduction to Machine Learning Approaches for Biomedical Research", Frontiers in Medicine, 2021 Publication	<1 %
6	Xiaoke Shang, Gehui Li, Zhiying Jiang, Shaomin Zhang, Nai Ding, Jinyuan Liu. "Holistic Dynamic Frequency Transformer for image fusion and exposure correction", Information Fusion, 2024 Publication	<1 %
7	xmed.jmir.org Internet Source	<1 %
8	link.springer.com Internet Source	<1 %
9	isip.piconepress.com Internet Source	<1 %
10	www.mdpi.com Internet Source	<1 %
11	www.biorxiv.org Internet Source	<1 %
12	www.cs.umd.edu Internet Source	<1 %
13	www.ijaerd.co.in Internet Source	<1 %

14	Submitted to Harrisburg University of Science and Technology Student Paper	<1%
15	Submitted to Brunel University Student Paper	<1%
16	Manpreet K. Singh, Kiki D. Chang. "The Neural Effects of Psychotropic Medications in Children and Adolescents", Child and Adolescent Psychiatric Clinics of North America, 2012 Publication	<1%
17	Yuwang Miao, Jizhong Zhu, Hanjiang Dong, Ziyu Chen, Shenglin Li, Xiyu Wen. "Short-term Load Forecasting Based on Echo State Network and LightGBM", 2023 IEEE International Conference on Predictive Control of Electrical Drives and Power Electronics (PRECEDE), 2023 Publication	<1%
18	insideaiml.com Internet Source	<1%
19	www.internationalscienceindex.org Internet Source	<1%
20	El Bilali, L.. "Role of sediment composition in trace metal distribution in lake sediments", Applied Geochemistry, 200209 Publication	<1%

21	Submitted to University of Aberdeen Student Paper	<1%
22	Submitted to St. Andrew's International School Student Paper	<1%
23	mdpi.com Internet Source	<1%
24	Yunfei Ding, Robert F. Harrison. "Relational visual cluster validity (RVCV)", Pattern Recognition Letters, 2007 Publication	<1%
25	dspace.lib.cranfield.ac.uk Internet Source	<1%
26	Submitted to Chandigarh Group of Colleges Student Paper	<1%
27	www.researchgate.net Internet Source	<1%
28	Submitted to ASA Institute Student Paper	<1%
29	Ani Harish, A. Prince, M. V. Jayan. "Fault Detection and Classification for Wide Area Backup Protection of Power Transmission Lines Using Weighted Extreme Learning Machine", IEEE Access, 2022 Publication	<1%

30	Sijin Yang, Lei Zhuang, Julong Lan, Jianhui Zhang, Bingkui Li. "Reuse-based online joint routing and scheduling optimization mechanism in deterministic networks", Computer Networks, 2024 Publication	<1 %
31	Soyeong Kim, Jinsu Ha, Kichun Jo. "Semantic Point Cloud-Based Adaptive Multiple Object Detection and Tracking for Autonomous Vehicles", IEEE Access, 2021 Publication	<1 %
32	blog.cureatr.com Internet Source	<1 %
33	dokumen.pub Internet Source	<1 %
34	liu.diva-portal.org Internet Source	<1 %
35	www.arxiv-vanity.com Internet Source	<1 %
36	www.azorobotics.com Internet Source	<1 %
37	Inzimam Ul Hassan, Zeeshan Ahmad Lone, Swati Swati, Aya Gamal. "chapter 4 Forecasting Weather and Water Management Through Machine Learning", IGI Global, 2023 Publication	<1 %

38 Neeta Chavan, Saakshi Karkera, Aishvarya Birambole, Isha Chavan, Risa Samanta. "Comparative Study of Machine Learning Algorithms for Prediction Of Polycystic Ovary Syndrome", 2023 9th International Conference on Advanced Computing and Communication Systems (ICACCS), 2023

Publication

<1%

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		