Babylonian Journal of Machine Learning
Vol.2024, **pp**. 69–79
DOI: https://doi.org/10.58496/BJML/2024/007; ISSN: 3006–5429
https://mesopotamian.press/journals/index.php/BJML

Review Article

Random Forest Algorithm Overview

Hasan Ahmed Salman [1,*], Ali Kalakech [2], Amani Steiti [3]

[1] *Computer Science Department , University Arts, Sciences and Technology, Beirut, Lebanon.*

[2] *Computer And Communication Engineering Department, School of Engineering Lebanese International University, Beirut, Lebanon.*

[3] *Department of Computer Systems And Networks, University Tishreen, faculty of information engineering, Latakia, Syria.*

ARTICLE INFO

Article History
Received 10 Mar 2024
Revised 07 Apr 2024
Accepted 14 May 2024
Published 08 Jun 2024

Keywords
Machine Learning
Decision Tree
Random Forest
Deep Learning

ABSTRACT

A random forest is a machine learning model utilized in classification and forecasting. To train machine learning algorithms and artificial intelligence models, it is crucial to have a substantial amount of high-quality data for effective data collecting. System performance data is essential for refining algorithms, enhancing the efficiency of software and hardware, evaluating user behavior, enabling pattern identification, decision-making, predictive modeling, and problem-solving, ultimately resulting in improved effectiveness and accuracy. The integration of diverse data collecting and processing methods enhances precision and innovation in problem-solving. Utilizing diverse methodologies in interdisciplinary research streamlines the research process, fosters innovation, and enables the application of data analysis findings to pattern recognition, decision-making, predictive modeling, and problem-solving. This approach also encourages innovation in interdisciplinary research. This technique utilizes the concept of decision trees, constructing a collection of decision trees and aggregating their outcomes to generate the ultimate prediction. Every decision tree inside a random forest is constructed using random subsets of data, and each individual tree is trained on a portion of the whole dataset. Subsequently, the outcomes of all decision trees are amalgamated to derive the ultimate forecast. One of the benefits of random forests is their capacity to handle unbalanced data and variables with missing values. Additionally, it mitigates the issue of arbitrary variable selection seen by certain alternative models. Furthermore, random forests mitigate the issue of overfitting by training several decision trees on random subsets of data, hence enhancing their ability to generalize to novel data. Random forests are highly regarded as one of the most efficient and potent techniques in the domain of machine learning. They find extensive use in various applications such as automatic categorization, data forecasting, and supervisory learning.

1. INTRODUCTION

Random Forest (RF) is a popular machine learning technique in the field of data mining [1]. It operates under the supervision of a group and has received significant recognition. Data mining can be categorized into two primary types: descriptive and predictive. Descriptive data mining is primarily concerned with providing detailed descriptions and summaries of data. On the other hand, predictive data mining involves studying historical data to identify patterns and trends that can be used to make predictions about the future. Metadata Mining is the process of describing and summarizing data, uncovering patterns and relationships within the data, and using historical data to make predictions about future trends. Predictive models are constructed by analyzing the features of predictive factors to provide hypotheses that assist in making future decisions. Predictive models are constructed by analyzing the characteristics of variables used for forecasting, and the results are hypotheses that can be empirically examined. The precision of such models relies on error estimating techniques. Metadata mining often employs unsupervised machine learning methods, whereas predictive data mining use supervised machine learning methods. Random forests are created by generating several decision trees. This is done by gathering random samples of data using Bootstrap samples and randomly selecting input features. Each decision tree is considered a simple decision tree [2]. One advantage of random forests is their high accuracy compared to other approaches like as bagging and boosting. They also function effectively on huge databases and can accommodate many variables, allowing us to analyze thousands of input variables without the need to delete any of them. In order to balance the category error in unbalanced data sets and assess the importance of variables, an unbiased estimation of the configuration error is necessary[18]. The forest

*Corresponding author. Email: has062@live.aul.edu.lb

demonstrates its power and efficiency through these advantages. Randomization is a widely used technique in machine learning and data mining due to its effectiveness in several practical applications [2][19].

2. MACHINE LEARNING

This algorithm involves a target or result variable, also known as the dependent variable, which may be predicted based on a specific collection of predictors, also known as independent variables. With this set of variables, we construct a function that maps inputs to the intended output. The training procedure persists until the model attains the desired degree of accuracy in the training data. The supervised learning techniques commonly used for regression analysis include decision tree, random forest, closest neighbor (KN), and logistic regression [4].

Unsupervised learning is used when there is data accessible solely in the input form and no corresponding output variable. These algorithms utilize statistical models to analyze the fundamental patterns inside the data in order to gain a deeper understanding of its characteristics.

Clustering is a prominent category of unsupervised algorithms. This technique involves the identification of intrinsic groupings within the data, which are then utilized to forecast the output of hidden inputs. A notable example of this methodology would be forecasting customer purchasing patterns [3].

This algorithm lacks both a target variable and a prediction or estimate outcome. It is utilized to categorize the population into distinct categories, a practice commonly employed to segment consumers for targeted services. Unsupervised learning examples include the prior-ISM method [5].

Reinforcement learning is utilized when the objective is to make a sequence of choices that lead to a final reward. Throughout the learning process, the artificial worker is given either rewards or penalties based on the acts it carries out. The objective is to optimize the overall reward. Examinations encompass the acquisition of knowledge related to playing computer games or carrying out activities involving robots, with the overarching objective [3].

By employing this method, the machine undergoes training to produce precise determinations. An instance involves subjecting the device to a setting where it consistently improves its performance by undergoing repeated trials and adjustments. This machine employs experiential learning to acquire and assimilate optimal knowledge in order to make precise business judgments. An illustration of reinforcement learning: A Markov Decision Process [4] is a mathematical framework used to model decision-making in situations where outcomes are uncertain and influenced by previous actions.

3. DECISION TREE

Decision Trees are a technique utilized in the fields of Statistics, data mining, and machine learning. It falls within the category of supervised machine learning. Data analysis technology categorizes data into different entities that are potentially associated with a specific procedure. There are two categories of such entities: contract and paper. The supervised learning approach employs the decision tree as a prediction model to examine the observations of an item in the branches and deduce the target value of the item in the leaves. The decision nodes symbolize the segmentation of data, while the sheets symbolize the outcomes.

A decision tree typically embodies cognitive processes resembling human thinking in order to facilitate informed decision making. Therefore, decision trees are highly comprehensible. Furthermore, there exists a hierarchical structure known as a decision tree, which greatly facilitates the comprehension of the underlying reasoning. For clarity, the following is the operational process of the terminal for your attention.

The root node: the starting point of the decision tree. The root node symbolizes the complete dataset, which is partitioned into two or more groups that can be compared.

Leaf node: The leaf nodes correspond to the ultimate result. The algorithm is unable to further divide the tree once it reaches the leaf node.

Splitting refers to the process of separating the root node or decision node into distinct sub-nodes based on specific parameters.

Pruning is the act of removing superfluous branches from a tree. By eliminating irrelevant branches, one can reach a conclusion much more quickly.

The parent node is the root node of the tree, while the other nodes are its children. A subtree is created when the primary tree is divided, resulting in new subtrees and branches. Machine learning encompasses two primary categories of decision trees, which are distinguished by the goal variable.

1. Categorization of trees
2. Trees used for regression analysis

3.1 Classification Trees

Classification trees are a specific kind of decision tree that can handle discrete values for the target variable. The decision tree is based on a categorical variable, with the possible value being "yes" at the first position. It eliminates the outcome following the evaluation of the provided data. A classification or judgment tree is an example of a tree structure that has two or more branches, each dependent on a distinct set of values.

An illustration of a rating tree may be seen in a dataset that assesses the decision of playing golf based on the prevailing weather conditions. The forecast serves as a decision node, which is then categorized into three branches: sunny, overcast, and rainy. In this context, a leaf node is activated when specific circumstances are met, and it determines whether the answer is "Yes" or "No" by traversing the tree.

3.2 Regression Trees

Regression trees are a specific kind of Decision Tree that can handle target variables with continuous values. The decision tree is designed to handle continuous variables, where the values can be represented as real numbers on a scale similar to cylinders. The outcome of the tree can consist of numerical values, such as 246. Typically, the creation of a regression tree includes utilizing inputs that consist of a mixture of continuous and discrete variables. Every decision node examines the input to evaluate the value of the variable [3]. The regression tree utilizes a binary recursive division method. At each iteration, the data is divided into parts, which are then further divided into smaller groups when ascending the branch.

A regression tree is capable of predicting the sale values of houses. The outcome of this example is a continuous dependent variable, where multiple constraining factors can include the size of the house in square feet, which remains constant. Additionally, there may exist categorical variables, such as the architectural style of the house, geographical region or location, and other others [3].

4. CLASSIFICATION AND REGRESSION TREE (CART)

Classification and regression trees are a data driven classification strategy that use historical data to create Decision Trees. Subsequently, decision trees are employed to categorize the novel data. Prior knowledge of the quantity of items is required in order to utilize the shopping cart. The cart methodology was abolished in the 1980s by Berman, Friedman, Olshan, and Stone in their publication "Classification and Regression Trees" (1984). In order to construct decision trees, the CART algorithm utilizes a sample learning approach, which involves using a collection of historical data where each observation is already allocated to a specific class. For instance, a training dataset for a credit scoring system might consist of fundamental data on past loans (variants) correlated with the corresponding repayment outcomes (categories).

Decision trees are characterized by a collection of inquiries that partition the training data into progressively smaller segments. The basket exclusively poses queries that may be answered with a simple "yes" or "no". A such inquiry could be: "Is the age greater than 50?" or "Is the sex male?" The shopping cart method aims to identify the optimal split, which is the question that splits the data into two sections with the highest level of homogeneity. To do this, the algorithm exhaustively searches for all potential variables and their corresponding values. Subsequently, the procedure is iterated for every individual data fragment that arises [6].

The shopping cart is a nonparametric method that selects the most significant factors and their interactions from a wide pool, which ultimately determine the outcome variable to be explained [7]. Cart is an independent program that can be executed on any operating system, including tutorials and Windows platforms.

The stroller excels in two specific areas. In practical terms, the installation and operation are straightforward. After the database is established, a straightforward application generates the results in a user friendly format. The user's text is [7].

1. The shopping cart does not make any assumptions about the distribution of dependent or pending variables. There should be no shopping cart variable that adheres to any statistical distribution.
2. The explanatory variables in the shopping cart can consist of a combination of category, comma, and trick variables.
3. The shopping cart is equipped with an integrated algorithm to handle missing values of a variable in most cases, except when employing a linear combination of variables as a division method.

4. The shopping cart is completely immune to outliers, interconnecting lines, varying elasticity, or distributive error structures that impact parametric actions. Outliers are data points that are distinct from the rest and do not influence the division of data. In contrast to parametric modeling, the shopping cart employs linear variables in the "alternative" division(s).
5. The cart has the ability to identify and expose the relationships inside the dataset.
6. The cart remains unchanged when the independent variables are transformed in a monotonous manner. This means that transforming the explanatory variables into logarithms, squares, or square roots does not have any impact on the resulting tree.

The shopping cart efficiently handles higher dimensions, meaning it may generate valuable findings by analyzing a huge number of factors while focusing on only a few variables that are immediately relevant.

A significant drawback of the cart is its lack of reliance on a probabilistic model. Predictions made using a cart tree to classify fresh data do not have an associated probability level or confidence interval. The level of trust an analyst can have in interpreting the results generated by a certain model, such as a tree, is determined exclusively by its historical accuracy. This refers to the amount to which the model accurately anticipated the desired outcome in previous situations [7].

5. RANDOM FOREST ALGORITHM

Random forests can be used either for a categorical response variable, such as "classification", or a continuous response, referred to as "regression". Similarly, expectation variables can be either categorical or continuous [8].
From a computational point of view, random forests are attractive because they:

- Naturally cope with both regression and classification (multilayer).
- Relatively quick to train and predict.
- Only depend on one or two adjusting parameters.
- You have a built-in Estimate of the generalization error.
- The following can be used directly for high dimensional problems.
- They can be easily implemented in parallel.

Statistically, random forests are attractive due to the additional features they provide, such as:

- Metrics of variable importance.
- Weighting the differential layer.
- Calculation of the missing value.
- The perception

5.1 How Random Forest Work?

Random forest is a machine learning technique that combines numerous decision trees to reduce the correlation among feature data. Simultaneously, the computational cost of RF is O(n) (where n represents the number of samples) when dealing with huge amounts of data. Additionally, the technique can be executed in parallel due to this integration, resulting in increased speed.

Random Forest (RF) mitigates the correlation between decision trees by employing a random selection of samples and features. Initially, an equivalent quantity of data is randomly chosen from the training sample in the original training data. Furthermore, a subset of the features is chosen at random to construct the decision tree. The utilization of these two forms of randomization results in a decrease in the correlation between each decision tree, hence mitigating the potential mistake caused by overfitting and enhancing the model's accuracy [8].A graph within a graph is an “inset”, not an “insert”. The word alternatively is preferred to the word “alternately” (unless you really mean something that alternates).

5.2 The Basic Algorithm

Figure 1 provides an intuitive understanding of radio frequencies. It has been observed that the daily charge quantity of different charging stations exhibits a distinct characteristic, namely, the charge quantity is significantly scattered. Thus, the process of dividing the original data into steps is deemed necessary. The range coverage intervals are selected based on the provided shipping quantity data value range. This approach enhances the efficacy and precision of the RF prediction algorithm by eliminating minor interference. There are two primary principles for dividing periods:

1. The shipping quantity range is separated into equal time periods to ensure equal allocation of time. This guarantees that the data is organized in a consistent manner, facilitating the identification of trends and patterns with greater ease.
2. The periods are separated according to the strength of the charge quantity. In this scenario, the time intervals vary and are contingent upon the concentration of data in specific regions within the range. This sort of segmentation enables concentration on the most crucial sections of the data that include more intricate and statistically meaningful information.

Data partitioning has several advantages. One of the key benefits is the reduction of interference. By splitting the data into particular periods, the interference between the data is minimized. This reduction in interference ultimately leads to improved accuracy in forecasting.

Data segmentation enhances the efficacy of algorithms employed in data analysis and RF forecasting, hence boosting their overall performance. The ease of analysis is enhanced when data is adequately segmented, as this allows for easier recognition of patterns and trends. By applying these concepts to divide time intervals, algorithms can be enhanced, leading to more precise and effective predicted outcomes. This, in turn, facilitates a more accurate comprehension and prediction of RF frequencies.

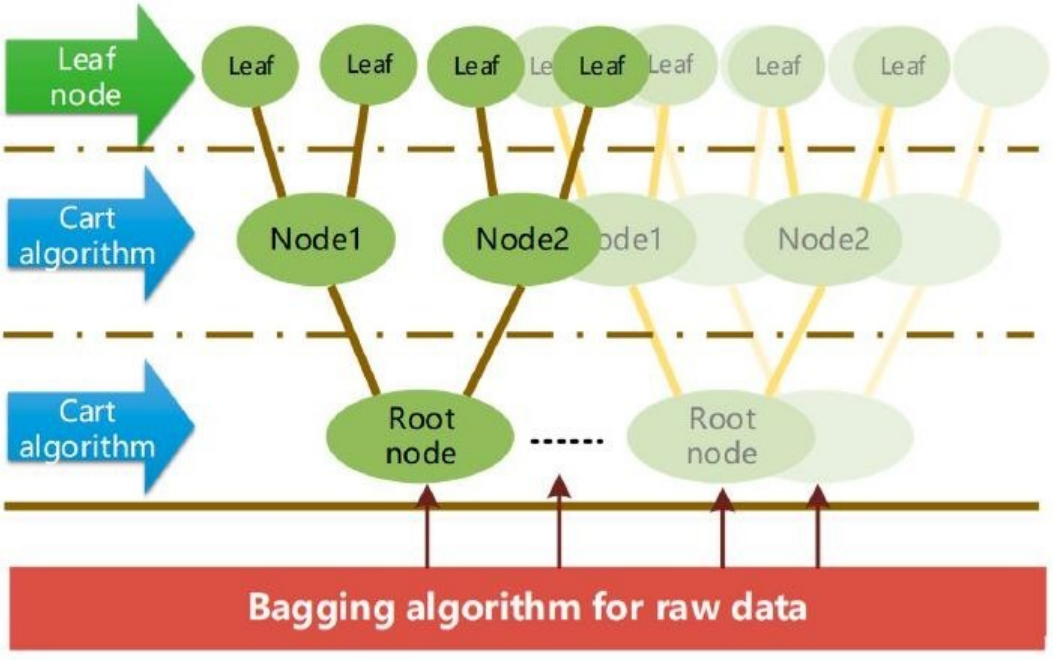

Fig. 1. Schematic diagram of random forest algorithm

5.3 Benefits and Challenges of Random Forest

One advantage of the random forest algorithm is its versatility. This approach is applicable to both regression and classification issues. The alga rhythm might be deemed advantageous as it yields superior outcomes even in the absence of hyperparameter adjustment. Furthermore, they possess a high level of clarity, making them easily comprehensible. Additionally, their number is quite limited.

Overfitting is a significant issue in machine learning. We must develop a universal model capable of achieving satisfactory outcomes on the test data. The random forest overcomes this dilemma by aggregating multiple decision trees, resulting in reduced bias and volatility.

The primary constraint of random forests is the extensive number of trees, resulting in a prolonged training time that renders it sluggish and inefficient for real-time predictions. Typically, these algorithms exhibit rapid training speed but significantly slower prediction speed after training. While the random forest algorithm is often efficient in real-world applications, there may be situations where runtime performance is crucial and alternative approaches may be more desirable [10].

5.4 Random Forest Applications

Some of the applications of the random forest may include [10]:

- Banking: Random forest is used in banking to predict the creditworthiness of a loan applicant. This helps the lending institution make a good decision on whether to give the customer the loan or not. Banks also use the random forest algorithm to detect fraudsters.
- Health care: Health professionals use random forest systems to diagnose patients. Patients are diagnosed by assessing their previous medical history. Past medical records are reviewed to establish the right dosage for the patients.
- Stock market: Financial analysts use it to identify potential markets for stocks. It also enables them to identify the behavior of stocks.
- Ecommerce: Through rainforest algorithms, ecommerce vendors can predict the preferences of customers based on past consumption behavior.

5.5 Classification and Regression in Random Forest

The classification process in random forests use an ensemble approach to achieve the final result. The training data is utilized to train multiple decision trees. This dataset comprises randomly selected observations and attributes that will be used for node splitting. A rainforest ecosystem depends on a multitude of decision trees. A decision tree is composed of decision nodes, leaf nodes, and a root node. The terminal node of each tree represents the ultimate result generated by that particular decision tree. The final outcome is determined using a majority-voting procedure. Here, the final output of the rainforest system is determined by the output selected by the majority of the decision trees. The provided diagram illustrates a basic random forest classifier [10].

Random forest algorithms also perform regression tasks. A random forest regression is based on the principles of simple regression. The random forest model receives the values of both dependent (features) and independent variables. Random forest regressions can be executed in different programming languages, including SAS, R, and Python. Each tree in a random forest regression generates a distinct prediction. The average forecast of the individual trees corresponds to the result of the regression. Unlike random forest classification, which determines its output based on the mode of the decision trees' class. While both random forest regression and linear regression have a common notion, they diverge in terms of their functions. Linear regression is a mathematical model represented by the equation $y = bx + c$. In this equation, y represents the dependent variable, x represents the independent variable, b is the parameter used for estimate, and c is a constant value. A complicated random forest regression can be likened to a black box, as it operates in a similar manner [10].

5.6 Bootstrap Aggregation (Bagging)

Bagging, or Bootstrap-Aggregating, entails creating K new training datasets. The process of selecting a fresh training data set involves randomly choosing a subset of information points, with the potential of picking the same data point many times (referred as bootstrap samples), from the initial set of data.

Sampling with replacement refers to the process of selecting data points from a dataset, where certain points may be duplicated in each fresh training dataset. The K models are trained using K bootstrap samples, which are then combined to make predictions. The predictions are obtained by averaging the results of all trees for regression tasks or by voting for classification tasks.

Bagging generally leads to enhanced accuracy compared to single tree prediction. However, the model that is produced can pose challenges when it comes to interpretation. The benefit of decision networks is their high interpretability. When we gather a substantial quantity of trees, it becomes impractical to convey the resultant statistical learning process using just one tree. Additionally, it becomes uncertain which variables hold the most significance. Bagging enhances the accuracy of predictions but sacrifices interpretability [11].

5.7 Properties and features

1. Feature selection

Feature selection is a procedure that involves using a set of rules to determine the relative importance and ranking of the attributes in a dataset. Feature selection techniques are commonly employed to modify the data in the classification analysis

in order to enhance the precision of categorization. Typically, machine learning feature selection strategies can be categorized into three primary groups: filter method (Filter), wrapper method (Wrapping), and the integration method [13].

This is the introductory section of the filtering technique. The filtration method involves using statistical analysis to assign weights to different characteristics. These weights are then used to rank the features. By applying certain rules and setting a threshold, features with weights above the threshold are kept, while those below it is removed. The feature selection process of the filtering approach is conducted based on the characteristics of the dataset, regardless of the particular classification algorithm being used. Several widely used filtering methods include Fisher ratio, information gain, Relief, T-test, and variance analysis. The subsequent section will provide a concise introduction to variance analysis [12].

2. Out of bag (OOB) estimates

Random Forest (RF) is a collection of regression or classification trees that was initially proposed by Bierman. One of the two stochastic elements in a random forest (RF) pertains to the selection of variables utilized for partitioning. For each division in a tree, the optimal variable for division is chosen from a random subset of predictors. If the selected entry number is too little, it is possible that none of the factors that make up the subset are significant, and that insignificant variables are frequently chosen for a split. The resultant trees exhibit little predictive capacity. If the subset consists of a significant number of predictors, it is probable that the same factors, specifically those with the greatest impact, are frequently picked for a split, whereas variables with lesser impacts have minimal chances of being selected. Hence, it is imperative to regard mtry as a tuning parameter [13].

Another stochastic element in Random Forest (RF) pertains to the selection of training data for each tree. Every tree in the algorithm known as Random Forests (RF) is constructed using a randomly chosen portion of the data. Typically, this refers to a bootstrap sample or a group of samples that is 0.632 times the size of the original sample (n). Hence, only select observations are utilized in the construction of a particular tree. Observations that are not utilized in the construction of a tree are referred to as out of bag (OOB) observations. Within a Random Forest (RF), every tree is constructed using a distinct subset of the original data, resulting in certain observations being excluded from some of the trees. The forecast for an observation can be obtained by utilizing only those trees that were not constructed using the observation.

By following this method, a classification is assigned to each observation, and the error rate can be determined based on these predictions. The error rate that is obtained as a result is commonly known as the out of bag (OOB) error. The Bierman Method, initially described by Bierman, has become a well-established technique for estimating errors in RF [13].

Out of Bag is synonymous with validation or test data. Random forests do not require a distinct testing dataset for result validation. The calculation is performed internally, within the algorithm's execution, using the following method.

Since the forest is constructed using training data, each tree is evaluated using 36.8% of the samples that were not used to create that specific tree. This is comparable to the validation data set.

The out of bag error estimate, often referred to as the internal error estimate, is a measure of the error in a random forest model while it is being built [14].

3. Variable importance measure (VIM) of Random Forest

The random forest model often uses the variable relevance measure to choose features across different categories. Díaz-Urartu and De Andres did a study on the utilization of random forests to identify a set of informative genes. The study showed that the random forest model has a similar level of prediction accuracy as the k-NN, support vector machine (SVM), and Diagonal Linear Discriminant Analysis (DLDA) models. The authors showed the importance of random forest variable relevance in discerning useful variables. Methods could be employed to identify a limited set of genes while maintaining predicting accuracy.

Through both simulation and empirical data, the model exhibits strong resilience when it comes to the parameters of node size and tree structure. The parameters determine the number of features that are available for each partitioning, the minimum number of samples required to reach a node before halting, and the number of trees utilized in the forest. The study demonstrated that increasing the size of the tree marginally enhanced the stability of the variable significance measurements. When the ratio of useful variables to the total number of variables is minimal, adding an entry results in a slight improvement in prediction accuracy. Finally, it was observed that the parameter node size, which determines the minimum size of the terminal nodes, had limited impact on prediction accuracy. The study delves deeper into the issue of identifying predictors, specifically the multiplicity problem. This problem arises when there are multiple subsets of predictor variables that yield the same level of prediction accuracy. Multiplicity is a frequently encountered difficulty in situations when the number of variables p is significantly greater than the number of observed cases n. This topic is extensively studied in the statistical literature. According to the author, solving a multitude of difficulties can be challenging in both small and large contexts. One possible approach to address this problem is to employ a range of methodologies and see if there is a specific group of variables that are consistently chosen by the majority of the models [14].

Random forests can be employed to assess the importance of variables in regression or classification problems using two significance measures. The first method, known as Mean Decrease Impurity (MDI), calculates the overall reduction in impurity of nodes when splitting on a particular variable. This calculation is then averaged over all trees. The second method is known as Mean Decrease Accuracy (MDA) [14].

4. Proximity measures

Accessing the original data in large population epidemiological research might be challenging due to privacy concerns associated with genetic data. Nevertheless, the study's summary statistics might be openly accessible and utilized by external entities for additional studies, without any worries regarding privacy concerns. The summary data may consist of individual association coefficients (regression betas) and probability values (p-values) indicating the relationship between genetic factors and phenotypic traits. By doing a comparison of the association patterns between numerous variables for one genetic variant and another, we may potentially get new insights regarding the functional similarities of these two variants. If two variations exhibit a comparable association pattern with a given collection of phenotypes, it is probable that both genes are functionally interconnected in some manner. Conversely, if their association patterns differ significantly, it is likely that they are functionally independent [15].

Varying approaches. Several procedures utilize machine learning techniques, including kernel methods, graph-based methods, Markov random field, and others.

Random Forests have been utilized in various genetic analysis projects in recent years due to their exceptional effectiveness in high dimensional analysis. Random Forests have the capability to provide many measures in addition to the classification model. These measurements include the proximity matrix, feature importance values, and the local importance matrix. After training the Random Forest, the proximity matrix displays the similarity between the samples in the Out of Bag (OOB) set. The OOB set is an internal validation set of the RF algorithm used to obtain performance measurements and the proximity matrix. The proximity between two samples is determined by counting the occurrences when these two instances end up in the same terminal node of the same tree in the Random Forest (RF) and then dividing it by the total number of trees in the forest [15].

5. Missing data

Random Forest is an ensemble method that combines many trees constructed from bootstrap samples of the original data. Random Forest is used for both classification and regression and provides many advantages such as having high accuracy, calculating a generalization error, determining the important variables and outliers, performing supervised and unsupervised learning, and imputing missing values with an algorithm based on proximity matrix [16].

Random forest (RF) missing data algorithms are an attractive approach for dealing with missing data. They have the desirable properties of being able to handle mixed types of missing data, they are adaptive to interactions and nonlinearity, and they have the potential to scale to big data settings. Currently, there are many different RF imputation algorithms but relatively little guidance about their efficacy, which motivated us to study their performance. Using a large, diverse collection of data sets, the performance of various RF algorithms was assessed under different missing data mechanisms. Algorithms included proximity imputation.

- **Approaches are missing the calculation of value.**

The first method: The algorithm for calculating the missing value of radio frequencies that was based on the proximity Matrix RF calculates the (n*n) proximity matrix to evaluate the similarity of observations. Elements outside the diagonal of the Matrix give the similarity of two different observations [11]. Based on these proximity values, the RF performs an iterative process of inclusion by following the following steps: First, an initial forest is created after using the mean embedding and then an approximation is calculated. The New calculated values are calculated by a weighted average based on proximity. With this updated data set, a new forest is created and by doing so a new affinity and calculated values are obtained. It was found that after performing 5 or 6 repetitions.

The second approach: the k-nearest neighbor (KN) embedding method, was applied to the data set before fitting the RF. In the KNN imputation method [15].

- first, the neighbors are determined by calculating the distance measures between observations. These measures are obtained through Makowski, Manhattan, or Euclidean functions.
- Because of is the most popular one amongst the others, the Euclidean distance function was used, and imputations were done based on weighted mean values of k nearest neighbors.

- The weights are inversely proportional to the distance measures. Not only different distance functions but also different algorithms of KNN can be seen. Some of them do not permit the neighbor values to contain missing values. this might cause the method to give less efficient results [16].

6. OPTIMIZATION OF THE RANDOM FOREST ALGORITHM

Like any other machine learning algorithm, Random Forest too, comes with some hyper-parameters to be optimized. And hyper-parameter tuning along with different cross validation techniques is what makes the results comparatively better [17].

A hyperparameter tuning itself won't get you anywhere if you haven't selected the best features during your data analysis and cleansing phase.

When tuning a Random Forest, you have to:

- Select the most influential parameters
- Understand how exactly, they will influence the training process
- Tune them manually or automatically way These parameters can be divided into: Category-A.
- Increase them → Underfitting to Overfitting
- Higher values → More powerful Model Category-B.
- Increase them → Overfitting to Underfitting
- Higher values → Less powerful Model.

6.1 Using classification and regression tasks learning technique random forest algorithm

1. Learn the band: A random forest is a group learning method, which means that it builds multiple decision trees during training and combines them to obtain a more accurate and stable prediction.
2. Decision trees: Each decision tree in the random forest is trained on a different subset of the training data and selects the best split at each node based on a random subset of features. Such randomness helps to decorative connect trees and prevent overfitting.
3. Boot Assembly (packing): A random forest uses a technique called Bootstrap aggregation or packing, in which each tree is trained on the bootstrap sample (a randomly sampled subset with substitution) of the training data. This adds more randomness to the model and helps to improve generalization.
4. Voting or average: For classification tasks, a random forest combines the predictions of individual trees by majority voting. For regression tasks, it averages the predictions of individual trees to obtain the final prediction.
5. Feature importance: A random forest provides a measure of feature importance based on how much each feature reduces impurities across all decision trees. This can be useful for feature selection and understanding the underlying patterns in the data.
6. Durability and scalability: Random forests are known for the power of noisy data and outliers. It also handles highdimensional data well and is relatively insensitive to hyperparameters, making it easy to use and less prone to over-processing compared to single decision trees.
7. Applications: The random forest is widely used in various fields, including, but not limited to:
    - Rating and regression problems in finance, such as credit scoring and stock price forecasting.
    - Healthcare for diagnosing disease and predicting patient outcomes.
    - Remote sensing for land cover classification and vegetation mapping.
    - Marketing to segment customers and predict wildly.
    - Natural language processing for text classification and sentiment analysis.

Random forest is a versatile and efficient algorithm that can produce high-quality predictions across a wide range of tasks and datasets [12].

7. RESULTS

Random forests are represented by a set of benefits and features that make them a popular choice in various fields of machine learning and predictive analysis. Here are some of the main consequences of using random forests:

1. High accuracy: random forests are powerful and reliable algorithms in many tasks, as they provide high accuracy in classification and forecasting.
2. Reduced variation: thanks to the assembly Learning technique, random forests can reduce mutation and increase stability in predictions compared to individual classification models.
3. Reduce the problem of overfitting: Random forests use the technique of building trees on different and independent sub-models, which reduces the problem of overfitting and improves the ability to generalize.
4. Improved performance in data Miscellaneous (performance): random forests work well on a variety of large and complex data and show a good ability to deal with lost data and jamming.
5. Provide an estimate of the importance of features (Feature Importance Estimation): Random forests provide an estimate of the importance of features, enabling users to identify the most important features of the model and understand the relative impact of each feature on the results.
6. Ease of Use and Integration: Random forests can be easily implemented using software libraries such as Sickie learn in Python, they are compatible with most software work environments and allow seamless integration with other technologies in the field of machine learning.

In general, random forests offer a set of significant benefits and positive results that make them a popular choice for solving a wide range of problems in the field of predictive analysis and machine learning.

8. CONCLUSIONS

Random forests are a highly efficient tool for making accurate predictions. As a result of the law of big numbers, they do not exceed their capacity. By including appropriate levels of unpredictability, they become precise classifiers and regressions. Furthermore, the framework elucidates the predictability of a random forest by examining the potency of individual predictors and their interconnections [13].

The random forest method is a user friendly and adaptable machine learning technique. Group learning is utilized to address regression and classification problems within businesses. This technique is optimal for developers as it effectively addresses the issue of excessive data processing. This method is highly valuable for generating precise forecasts that are essential for strategic decision-making in companies [14].

Aggregation approaches seek to enhance the accuracy of classification by combining predictions from many classifiers. The more the diversity and weaker the connections between the basic classifiers, the higher the accuracy of the set will be.

The random forest method is employed.

1) The process of selecting a subset of examples/cases, such as in packing, is referred to as sub-sampling.
2) When a subset of features is selected, it is called Feature Selection.

Both of these tactics are employed in random forests to incorporate randomization and attain diversity.

Conflict Of Interest

None.

Funding

None.

Acknowledgment

The author would like to thank the institution for creating an enabling environment that fostered the development of this research.

References

[1] Y. K. Rushall and P. K. Sinha, "Random Forest Classifiers: A Survey and Future Research Directions," in Int. J. Adv. Comput., vol. 36, no. 1, 2013.

[2] B. Leo, "Random Forests," in Machine Learning, vol. 45, pp. 5-32, 2001.

[3] S. Sah, "Machine Learning: A Review of Learning Types," doi:10.20944/preprints202007.0230.v1, 2020.

[4] A. Abdi, "Three types of Machine Learning Algorithms," DOI: 10.13140/RG.2.2.26209.10088, 2016.

[5] "Decision Trees in Machine Learning," [Online]. Available: https://pdf.co/blog/decision-trees-in-machine-learning.

[6] R. Timofeev, "Classification and Regression Trees (CART) Theory and Applications," Master's Thesis, Center of Applied Statistics and Economics, Humboldt University, Berlin, 2004.

[7] Y. Yohannes and J. Rhodiot, "Classification and Regression Trees: Introduction," Int. Food Policy Res. Inst., USA, 2006.

[8] A. Cutler, D. R. Cutler, and J. R. Stevens, "Random forest," in Machine Learning Mag., DOI: 10.1007/978-1-4419-9326-7_5, Jan. 2011.

[9] Y. Lu et al., "The Application of Improved Random Forest Algorithm on the Prediction of Electric Vehicle Charging Load," Energies, doi:10.3390/en11113207, 2018.

[10] "An Introduction to Random Forest Algorithm for Beginners," [Online]. Available: https://www.analyticsvidhya.com/blog/2021/10/an-introduction-to-random-forest-algorithm-for-beginners/.

[11] "Bagging: 25 Questions to Test Your Skills on Random Forest Algorithm," [Online]. Available: https://www.analyticsvidhya.com/blog/2021/05/bagging-25-questions-to-test-your-skills-on-random-forest-algorithm/.

[12] G. Biau, "A Random Forest Guided Tour," Sorbonne Universités, UPMC Univ Paris 06, F-75005, Paris, France & Institut Universitaire de France, 2010.

[13] S. Janitza and R. Hornung, "On the overestimation of random forest's out-of-bag error," Y-h. Taguchi, Chuo University, Japan, 2018.

[14] A. Hjerpe, "Computing Random Forests Variable Importance Measures (VIM) on Mixed Continuous and Categorical Data," Kth Royal Institute of Technology, School of Computer Science and Communication, Stockholm, Sweden, 2016.

[15] J. A. Seoane, I. N. M. Day, et al., "A Random Forest proximity matrix as a new measure for gene annotation," European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, Bruges, Belgium, i6doc.com publ., ISBN 978-287419095-7, 2014.

[16] H. Ozen and C. Bal, "A Study on Missing Data Problem in Random Forest," Osmangazi J. Med., doi: 10.20515/otd.496524, 2019.

[17] "Optimizing a Random Forest," [Online]. Available: https://medium.datadriveninvestor.com/optimizing-a-random-forest.

[18] T. Al-Quraishi, C. Keong NG, O. A. Mahdi, A. Gyasi, and N. Al-Quraishi , Trans., “Advanced Ensemble Classifier Techniques for Predicting Tumor Viability in Osteosarcoma Histological Slide Images”, Applied Data Science and Analysis, vol. 2024, pp. 52–68, May 2024, doi: 10.58496/ADSA/2024/006.

[19] S. Gupta, S. Roy, and K. Maji , Trans., “Integrated Learning Paradigm for Ecological Predictive Modeling”, Babylonian Journal of Artificial Intelligence, vol. 2023, pp. 64–73, Oct. 2023, doi: 10.58496/BJAI/2023/010.