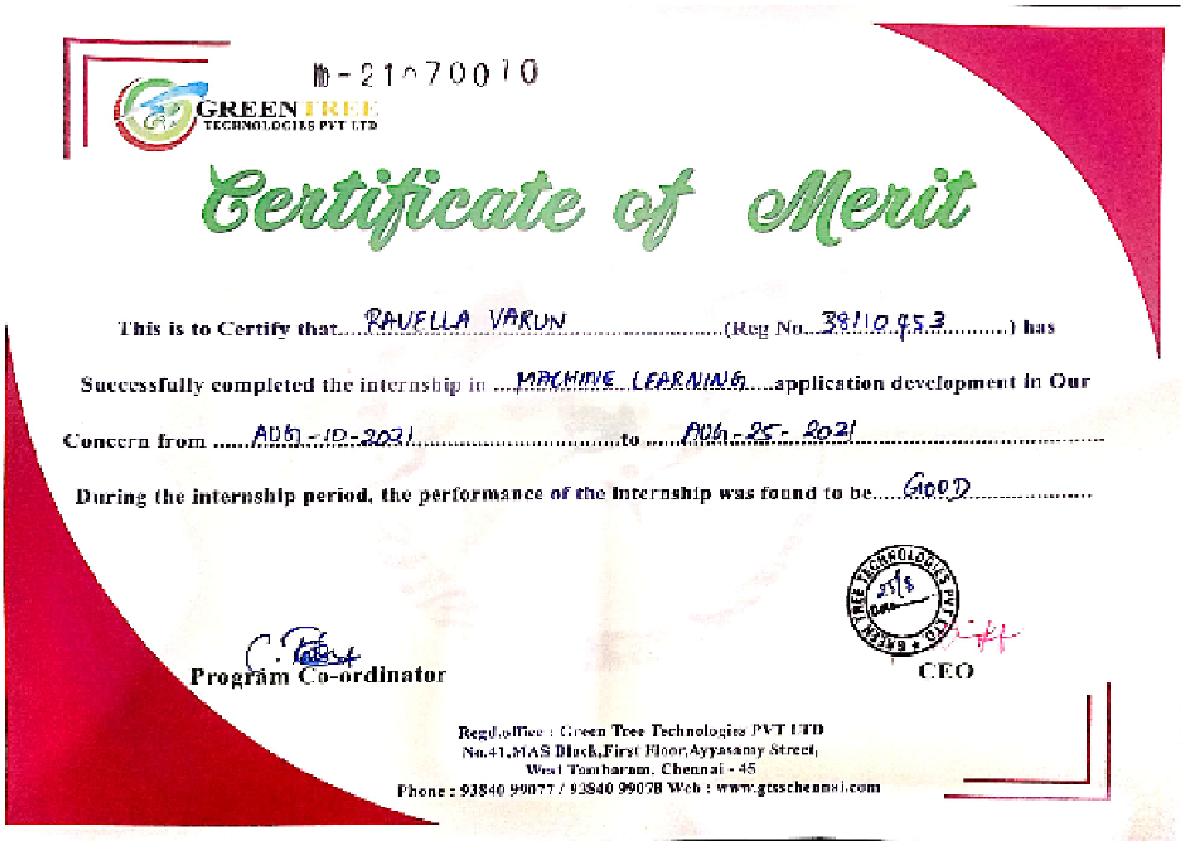
# TRAINING CERTIFICATE



M-21070010

GREEN TREE
TECHNOLOGIES PVT LTD

## Certificate of Merit

This is to Certify that......RAVELLA VARUN..............................(Reg No...38110453..........) has

Successfully completed the internship in ....MACHINE LEARNING.....application development in Our

Concern from ......AUG-10-2021....................................to .......AUG-25-2021..........................................

During the internship period, the performance of the internship was found to be.....GOOD......................

GREEN TREE TECHNOLOGIES PVT LTD

Program Co-ordinator

CEO

Regd.office : Green Tree Technologies PVT LTD
No.41,MAS Block,First Floor,Ayyasamy Street,
West Tambaram, Chennai - 45
Phone : 93840 99077 / 93840 99078 Web : www.gtschennai.com

# ABSTRACT

This paper predicted a model that indicates whether to buy a car insurance based on primary health insurance customer data. Currently, automobiles are being used to land transportation and living, and the scope of use and equipment is expanding. This rapid increase in automobiles has caused automobile insurance to emerge as an essential business target for insurance companies. Therefore, if the car insurance sales are predicted and sold using the information of existing health insurance customers, it can generate continuous profits in the insurance company's operating performance. Therefore, this paper aims to analyze existing customer characteristics and implement a predictive model to activate advertisements for customers interested in such auto insurance. The goal of this study is to maximize the profits of insurance companies by devising communication strategies that can optimize business models and profits for customers. This study was conducted through an automobile insurance purchase prediction model was implemented using Health Insurance Cross-sell Prediction data. The proposed system uses Random Forest Classifier. According to the results of this study, Train Accuracy is 0.992 and test Accuracy is 0.936, which has high accuracy. Therefore, the result was that customers with health insurance could induce a positive reaction to auto insurance purchases.

# TABLE OF CONTENTS

# CHAPTER 1

# INTRODUCTION

- Today, automobiles have become a key tool for land transportation, and their use range is expanding as a means of living.
- Cars are a running weapon and have a problem in that accident caused by banks also increase, leading to a significant increase in life and property damage.
- However, as relationship marketing becomes more critical, the repurchase of insurance and maintenance of insurance contracts by existing customers is an important success factor for the insurance industry.
- In the long term, minimizing the conversion of existing customers to other insurance companies is a positive factor that increases corporate profits.
- Thus, it became a major problem for insurers to increase business efficiency by maintaining existing contractors by controlling the factors of car purchase conversion (Kong et al., 2019).
- However, consumers are concerned that the quality of service of the insurance company they are currently subscribing to is lower than the quality inherently reserved by consumers, and the expected utility from searching for a new insurance company is If it is less than the expected utility, it is highly likely to switch to a sub-optimal insurance company; otherwise, it is highly likely to search for or switch to a new insurance company.
- In particular, the more information about premiums presented by insurance companies, the more likely a subscriber will switch to another insurance company.
- Therefore, this study was conducted by analyzing customer characteristics and implementing a prediction model to activate advertisements for customers interested in such auto insurance in this paper.
- The research results can also be used to increase purchase desire by effectively exposing automobile insurance to existing insurance customers.

## 1.1 PROPOSED ALGORITHMS

### RANDOM FOREST ALGORITHM

**Random forest algorithm can use both for classification and the regression kind of problems.** In this you are going to learn, how the **random forest algorithm** works in machine learning for the classification task.

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of **ensemble learning,** which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.

A random forest algorithm consists of many decision trees. The 'forest' generated by the random forest algorithm is trained through bagging or bootstrap aggregating. Bagging is an ensemble meta-algorithm that improves the accuracy of machine learning algorithms.

As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.

The below diagram explains the working of the Random Forest algorithm:

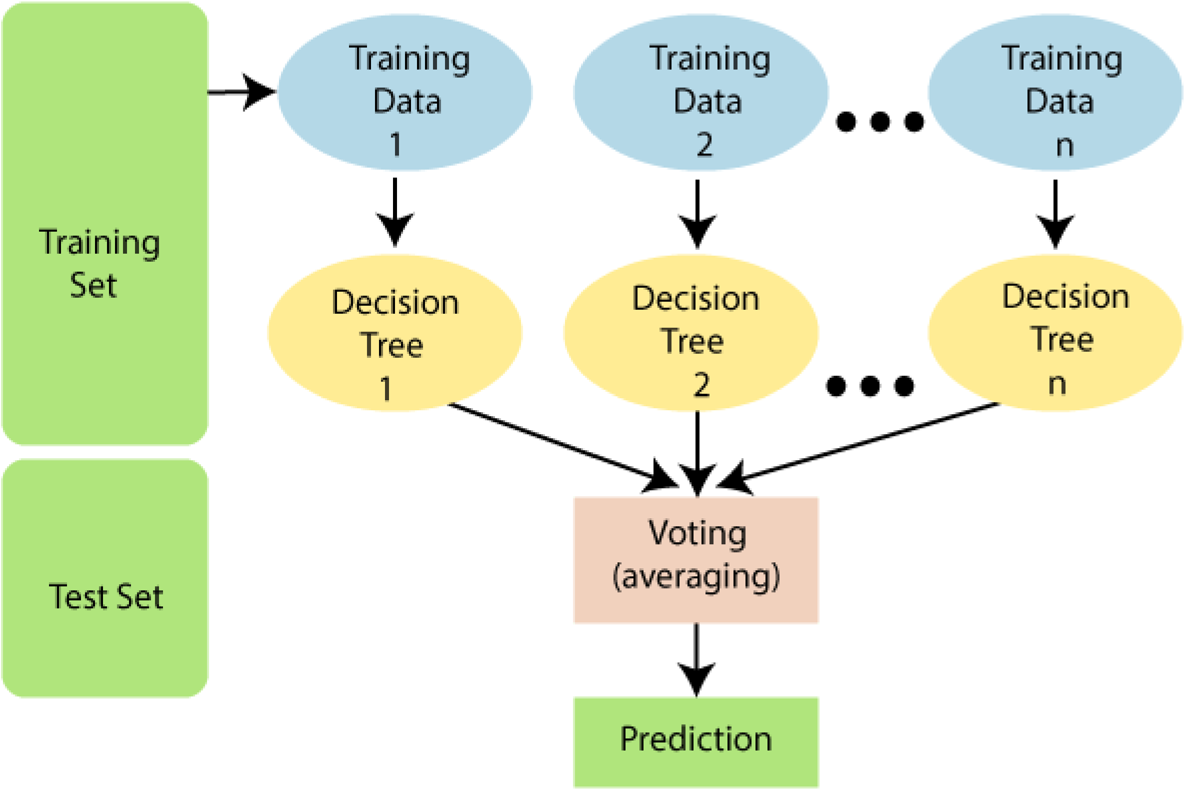


**Fig 2:** Explaining the working algorithm of the Random Forest algorithm

Below are some points that explain why we should use the Random Forest algorithm:

- It takes less training time as compared to other algorithms.
- It predicts output with high accuracy, even for the large dataset it runs efficiently.
- It can also maintain accuracy when a large proportion of data is missing.

## 1.2 FEATURES OF A RANDOM FOREST ALGORITHM

- It's more accurate than the decision tree algorithm.
- It provides an effective way of handling missing data.
- It can produce a reasonable prediction without hyper-parameter tuning.
- It solves the issue of overfitting in decision trees.
- In every random forest tree, a subset of features is selected randomly at the node's splitting point.

## 1.3 CLASSIFICATION IN RANDOM FORESTS

Classification in random forests employs an ensemble methodology to attain the outcome. The training data is fed to train various decision trees. This dataset consists of observations and features that will be selected randomly during the splitting of nodes.

A rain forest system relies on various decision trees. Every decision tree consists of decision nodes, leaf nodes, and a root node. The leaf node of each tree is the final output produced by that specific decision tree. The selection of the final output follows the majority-voting system. In this case, the output chosen by the majority of the decision trees becomes the final output of the rain forest system. The diagram below shows a simple random forest classifier.

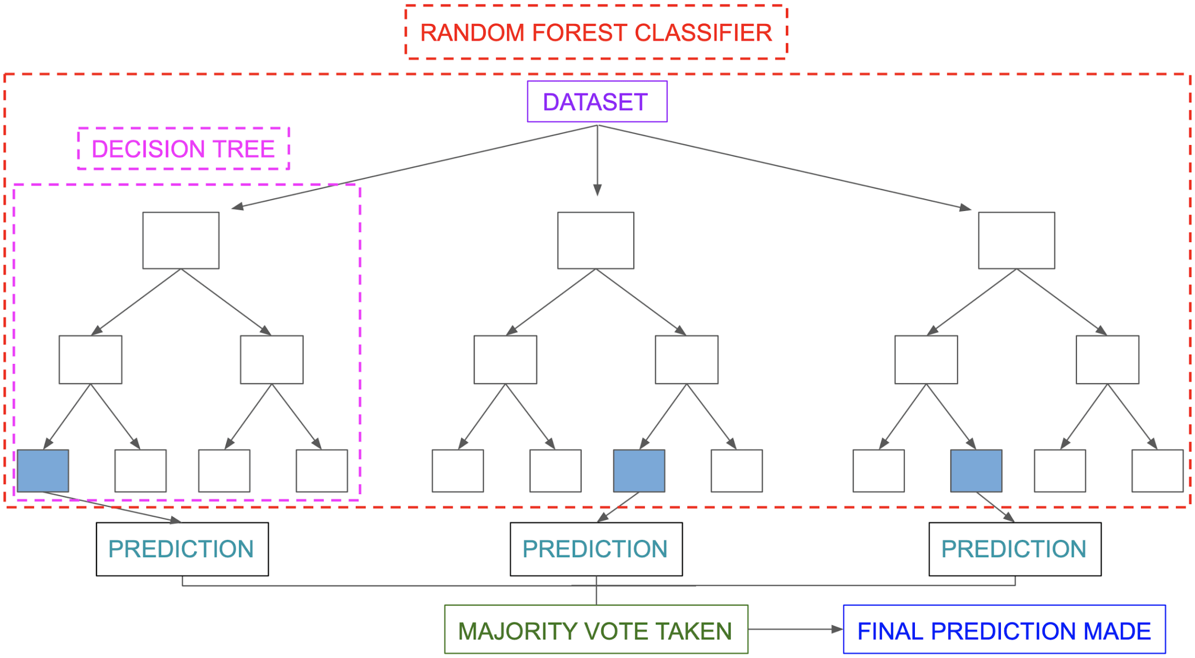


**Fig 3:** Explaining the Random Forest Classifier

## 1.4 RANDOM FOREST STEPS

1. Randomly select “k” features from total “m” features.
    1. Where k << m
2. Among the “k” features, calculate the node “d” using the best split point.
3. Split the node into daughter nodes using the best split.
4. Repeat 1 to 3 steps until “l” number of nodes has been reached.
5. Build forest by repeating steps 1 to 4 for “n” number times to create “n” number of trees.

The beginning of random forest algorithm starts with randomly selecting “k” features out of total “m” features. In the image, you can observe that we are randomly taking features and observations.

The working of the algorithm can be better understood by the below example:
**Example:** Suppose there is a dataset that contains multiple fruit images. So, this dataset is given to the Random forest classifier. The dataset is divided into subsets and given to each decision tree. During the training phase, each decision tree produces a prediction result, and when a new data point occurs, then based on the majority of results, the Random Forest classifier predicts the final decision. Consider the below image:

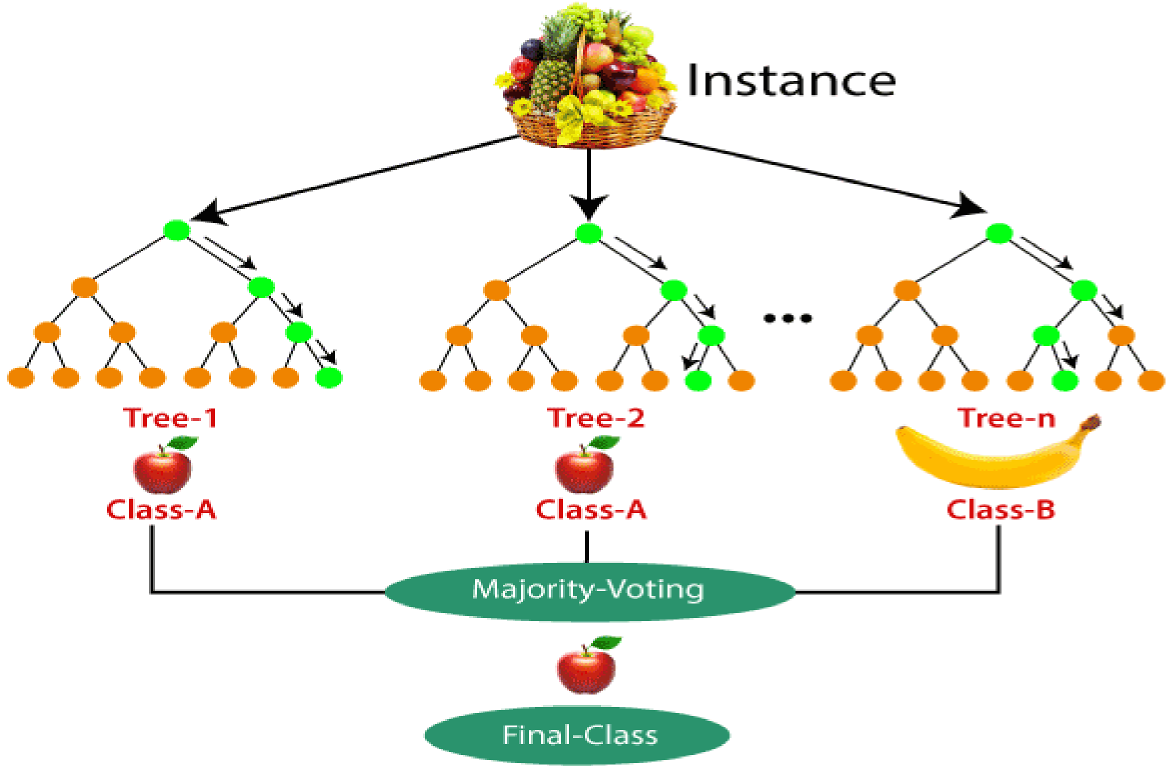


**Fig 4: Explaining the Random Forest Classifier algorithm with example**

## 1.5 APPLICATIONS OF RANDOM FOREST

There are mainly four sectors where Random forest mostly used:

1. **Banking:** Banking sector mostly uses this algorithm for the identification of loan risk.
2. **Medicine:** With the help of this algorithm, disease trends and risks of the disease can be identified.
3. **Land Use:** We can identify the areas of similar land use by this algorithm.
4. **Marketing:** Marketing trends can be identified using this algorithm.

## 1.6 ADVANTAGES OF RANDOM FOREST

Random Forest is capable of performing both Classification and Regression tasks.

- It is capable of handling large datasets with high dimensionality.
- It enhances the accuracy of the model and prevents the overfitting issue.

## 1.7 DISADVANTAGES OF RANDOM FOREST

- Although random forest can be used for both classification and regression tasks, it is not more suitable for Regression tasks.

# CHAPTER 2
# LITERATURE SURVEY

Literature survey is the most important step in software development process. Before developing the tool it is necessary to determine the time factor, economy and company strength. Once these things are satisfied, then the next step is to determine which operating system and language can be used for developing the tool. Once the programmers start building the tool the programmers need lot of external support. This support can be obtained from senior programmers, from book or from websites. Before building the system the above consideration are taken into account for developing the proposed system. The major part of the project development sector considers and fully survey all the required needs for developing the project. For every project Literature survey is the most important sector in software development process. Before developing the tools and the associated designing it is necessary to determine and survey the time factor, resource requirement, man power, economy, and company strength. Once these things are satisfied and fully surveyed, then the next step is to determine about the software specifications in the respective system such as what type of operating system the project would require, and what are all the necessary software are needed to proceed with the next step such as developing the tools, and the associated operations.

## 2.1 AN ENSEMBLE RANDOM FOREST ALGORITHM FOR INSURANCE BIG DATA ANALYSIS

Due to the imbalanced distribution of business data, missing user features, and many other reasons, directly using big data techniques on realistic business data tends to deviate from the business goals. It is difficult to model the insurance business data by classification algorithms, such as logistic regression and support vector machine (SVM). In this paper, we exploit a heuristic bootstrap sampling approach combined with the ensemble learning algorithm on the large-scale insurance business data mining, and propose an ensemble random forest algorithm that uses the parallel computing capability and memorycache mechanism optimized by Spark. We collected the insurance business data from China Life Insurance Company to analyze the potential customers using the proposed algorithm. We use F-Measure and G-mean to evaluate the performance of the algorithm. Experiment result shows that the ensemble random forest algorithm outperformed SVM and other classification algorithms in both