THE CAR INSURANCE CLAIM PREDICTION SYSTEM BY USING MACHINE LEARNING ALGORITHMS ON APACHE SPARK PLATFORM

THEIN THAN KO

M.I.Sc. **May 2023**

THE CAR INSURANCE CLAIM PREDICTION SYSTEM BY USING MACHINE LEARNING ALGORITHMS ON APACHE SPARK PLATFORM

By

Thein Than Ko

D.C.Sc.

A Dissertation Submitted in Partial Fulfillment of the Requirements for the Degree of

Master of Information Science

(M.I.Sc.)

University of Computer Studies, Yangon

May 2023

STATEMENT OF ORIGINALITY

I hereby certify that the work embodied in this thesis is the result of original research and has not been submitted for a higher degree to any other University or Institution.

…… ……

Date Thein Than Ko

ACKNOWLEDGEMENTS

First and Foremost, I would like to express my sincere gratitude to **Dr. Mie Mie Khin**, Rector, the University of Computer Studies, Yangon, for her kind permission to develop this thesis.

My sincere thanks and regards go to **Dr. Si Si Mar Win**, Professor, Faculty of Computer Science, University of Computer Studies, Yangon, for their kind management throughout the completion of this thesis.

I would like to express my heartfelt thanks to my supervisor, **Dr. Tin Zar Thaw,** Professor, Faculty of Computer Science, University of Computer Studies, Yangon for her encouragement, patient and invaluable supervision in compiling the materials for my thesis.

I also deeply thank **Daw Aye Aye Khine**, Associate Professor & Head, Department of English, University of Computer Studies, Yangon for editing my thesis from the language point of view.

Finally, I especially thank my parents, all of my friends for their suggestions, support and generous help rendered me during the development of thesis. Their love and concerned encouragement have strengthened me through the studies.

Moreover, I would like to thank all the staff and teachers from the University of Computer Studies, Yangon for their support.

ABSTRACT

Car insurance companies face a major challenge in dealing with insurance claims, which are prone to fraud and increasing in volume. This makes it difficult for insurers to classify claims during the review process. To address this issue, the aim of this study is to develop four Car Insurance Claim Prediction Classifiers with Random Forest and Logistic regression based on the car insurance claim dataset respectively and supports for comparison which method and attributes are more suitable for car insurance companies. Firstly, this proposed system creates a feature selection model using Variance Threshold Selector method to select the important attributes impact on the accuracy of car insurance claim prediction classifiers. The data set is split into training with 80% and testing sets with 20% randomly and the two classifiers with all attributes, the training dataset is used to create the LR classifier and RF classifier. For two classifiers with the feature selection method, the system creates the new training dataset and new testing dataset by removing low variance value of attributes using Variance Threshold Selector method. After that, two LR classifier and RF classifier are been created by using new datasets. The system has analyzed the different attributes: 30, 32, 34, 36, 38, 40 and 42 to choose the number of attributes and important attributes and tested 10 times for each attribute number because of splitting training and testing datasets randomly. Finally, the system compares the evaluation results with metrics: accuracy and f score. RF classifiers with and without the feature selection method are suitable for the proposed system than LR classifiers. Among different attribute numbers, the classifiers based on 38 attributes and 40 attributes are the best classifiers and classifier based on 42 attributes are the second best classifier.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

Page

LIST OF EQUATIONS

CHAPTER 1
INTRODUCTION

The insurance industry is a fast-growing sector [10] [13] that plays a crucial role in ensuring the economic well-being of a country. However, car insurance claims in insurance companies can be costly problems, and insurance providers must always make a great effort to combat the growing cost of insurance claims and claim loss due to insurance claim fraud [15]. Insurance companies face business problems, such as risk assessment, classification of policy holders, resource allocation, insurance claim classification, and prediction in the insurance claim handling process [2]. With the advancements in computing technology, machine learning approaches have emerged as a viable solution to these problems, particularly for handling and processing large amounts of data such as that found in insurance databases [5].

The use of machine learning classifiers in big data analysis helps the insurance industry to predict future trends in the competitive market. Big data, which includes structured, unstructured, and semi-structured data, has fundamentally changed data management across the insurance industry [14] as traditional relational database management systems and software tools are unable to cope with the sheer volume and variety of data [3] [19]. In this system, Apache Spark, open source processing engine, uses to control big data problem. Apache spark uses directed acyclic graph and its own data structure i.e., Resilient Distributed Dataset (RDD) to provide speed and analytics [16]. Spark helps in some challenging and computationally exhaustive tasks like processing high volumes of real-time and archived data, thereby integrating the complex capabilities such as ML and graph algorithms. It brings big data processing to the market and Spark has a library for ML labelled as MLib. Spark MLib library has algorithms for the functions of classification, regression, clustering, collaborative filtering, and dimensionality reduction. Machine learning approaches are essential to process the data and extract vital insurance claim information for decision-making processes [5][12]. In this system, Logistic Regression Classifier, Random Forest Classifier and Variance Threshold Selector method from MLib are used to apply the car insurance claim prediction system.

1.1 Objectives of the Thesis

The main objective of this study is to build a machine learning classifier that classifies and predicts car insurance claim status in the next six months. The other objectives of this study include the following:

- To predict whether the car claim care insurance within the next six months for car insurance companies and car owners.
- To apply Logistic Regression and Random Forest classifiers to the car insurance claim prediction system.
- To apply the feature selection method: Variance Threshold Selector to the car insurance claim prediction system with Logistic Regression and Random Forest classifiers.
- To support for selecting important attributes for the proposed system.
- To apply the proposed car insurance claim prediction system on the apache spare platform.

1.2 Motivation of the Thesis

Today, the car insurance companies have many difficulties and challenges to make decision correctly and to analyze large amount of structured data for car insurance claim. To improve their decision-making process and reduce their costs, a car insurance claim prediction system is proposed. This can help insurance companies to proactively identify high-risk claims and take measures to prevent them, leading to improved customer satisfaction and reduced costs. Large amounts of structured data, such as past insurance claims, the system can predict the likelihood of future claims are analyzed and fraudulent claims are detected by using machine learning techniques more accurately. Overall, the motivation behind developing a car insurance claim prediction system is to provide insurance companies with a powerful tool to manage their risks and improve their operational efficiency.

1.3 Organization of the Thesis

This thesis is organized into five chapters. In chapter one, the car insurance claim prediction system is introduced. This chapter also described the objectives of the thesis, the motivation of the thesis, and the organization of the thesis. In chapter two, background theory and related works are presented. In chapter three, the proposed system design, explanation of how works the car insurance claim prediction system, car insurance claim dataset, a feature selection method and two classification methods of Apache spark machine learning library (ML lib) are explained in detail. In chapter four, the implementation of the system and experimental results are expressed in detail. In chapter five, the conclusion of the thesis work is presented. In addition, further extensions of the system are depicted.

CHAPTER 2
BACKGROUND THEORY AND RELATED WORKS

This chapter provides the technical context of the presented system and explains the work related to the car insurance claim prediction system.

2.1 Apache Spark

Spark, an open-source processing engine, utilizes a directed acyclic graph and its proprietary data structure known as Resilient Distributed Dataset (RDD) to deliver high speed and analytics. It excels in handling demanding and computationally intensive tasks, such as processing large volumes of real-time and archived data, while seamlessly integrating complex capabilities like machine learning and graph algorithms. Spark revolutionizes big data processing and offers the MLlib library, also referred to as MLib [3]. The Spark MLib library encompasses a wide range of algorithms for classification, regression, clustering, collaborative filtering, dimensionality reduction, and more [8]. The key components in the Apache Spark pipeline are as follows [9] [23]:

MLlib simplifies the usage of machine learning algorithms by standardizing APIs, enabling the seamless combination of multiple algorithms into a unified pipeline or workflow. This section provides an overview of the essential concepts introduced by the Pipelines API, heavily influenced by the scikit-learn project.

In MLlib, the Data Frame from Spark SQL serves as the ML dataset, offering versatility in accommodating various data types. For instance, a Data Frame can encompass distinct columns storing text, feature vectors, true labels, and predictions.

A Transformer, an integral component, refers to an algorithm capable of converting one Data Frame into another. Notably, the creation of models such as the Logistic Regression Classifier and Random Forest Classifier involves utilizing Transformers.

An Estimator, on the other hand, represents an algorithm that can be trained on a Data Frame to generate a Transformer. For instance, a learning algorithm functions as an Estimator that undergoes training on a Data Frame, ultimately producing a model [20].

To specify an ML workflow, a Pipeline orchestrates the sequential chaining of multiple Transformers and Estimators. By integrating these elements, the Pipeline streamlines the execution of the entire ML process.

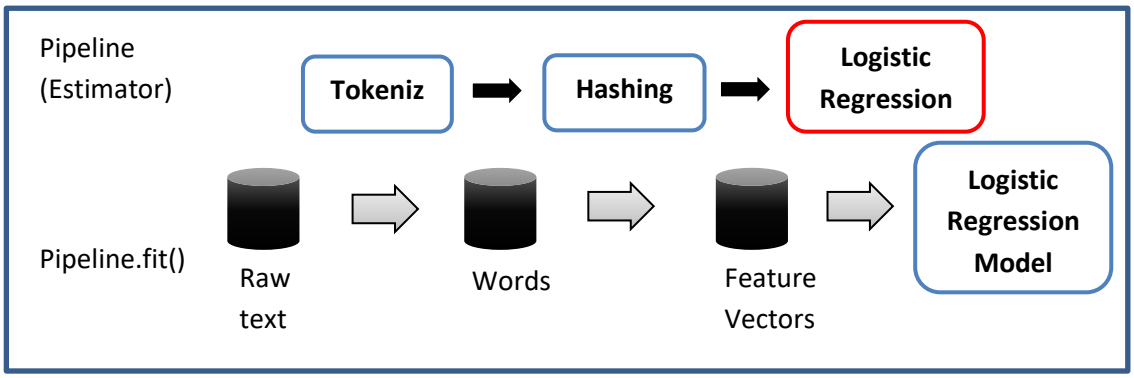

Figure 2.1 Transformer Pipeline of Logistic Regression

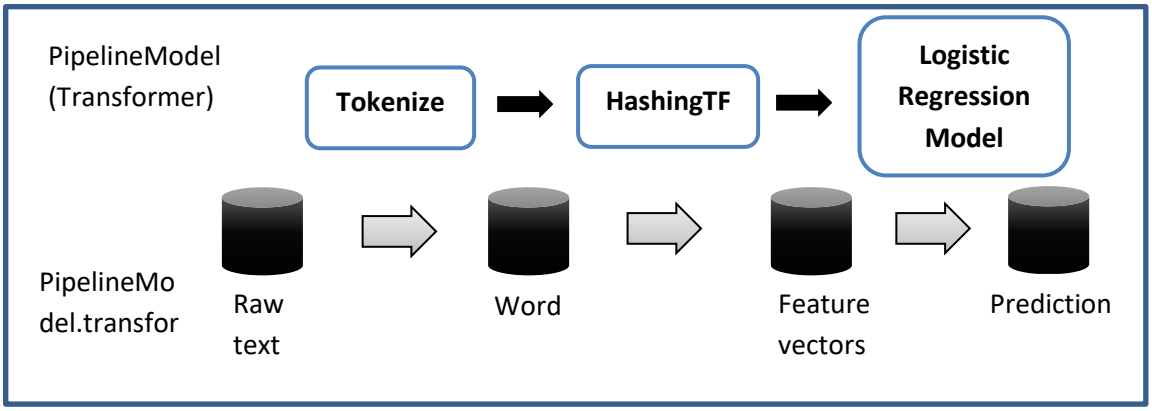

Figure 2.2 Estimator Pipeline of Logistic Regression

2.2 Apache Spark ML Lib

Spark is an open source processing engine, which uses directed acyclic graph and its own data structure i.e., Resilient Distributed Dataset (RDD) to provide speed and analytics [17] [26]. Learn how to use Apache Spark MLlib to create a machine learning application. The application will do predictive analysis on an open dataset. From Spark's built-in machine learning libraries, this example uses classification through logistic regression [11] [21].

MLlib is a core Spark library that provides many utilities useful for machine learning tasks, such as:

- Classification
- Clustering
- Modeling

- Singular value decomposition (SVD) and principal component analysis (PCA)
- Hypothesis testing and calculating sample statistics

2.2.1 Logistic Regression

Classification, a widely employed machine learning task, involves organizing input data into distinct categories or classes. The responsibility of a classification algorithm is to determine the appropriate "labels" to assign to the given input data [25]. To illustrate, consider a scenario where a machine learning algorithm receives stock information as input and categorizes the stocks into two groups: stocks that should be sold and stocks that should be retained. Logistic regression serves as the classification algorithm of choice. Spark's logistic regression API is particularly valuable for binary classification, where input data is classified into one of two groups. In essence, the logistic regression process generates a logistic function that can be employed to predict the probability of an input vector belonging to either group.

2.2.2 Random Forest

As the random forest algorithm combines multiple trees to make predictions on a dataset, it is possible for some decision trees to accurately predict the output while others may not. However, when all the trees are considered together, they collectively yield the correct output [18][22]. Consequently, there are two key assumptions for enhancing the performance of the Random Forest classifier:

- The dataset should contain genuine values in the feature variable to ensure accurate predictions by the classifier, as opposed to relying on guessed or approximate results.
- The predictions from each tree should exhibit minimal correlations with each other.

The following points highlight the advantages and reasons for utilizing the Random Forest Algorithm:

- It requires less training time in comparison to other algorithms.
- It achieves high prediction accuracy, even when dealing with large datasets, and demonstrates efficient performance.

- It can effectively handle situations where a significant proportion of the data is missing, without sacrificing accuracy.

2.2.3 Feature Selection

The biggest challenge of Machine Learning is to create models that have robust predictive power by using as few features as possible [26]. But given the massive sizes of today's datasets, it is easy to lose the oversight of which features are important and which ones are not.

That is why there is an entire skill to be learned in the ML field — feature selection. Feature selection is the process of choosing a subset of the most important features while trying to retain as much information as possible.

Variance Threshold Selector is a selector that removes low-variance features. Features with a variance not greater than the variance Threshold will be removed. If not set, variance Threshold defaults to 0, which means only features with variance 0 (i.e. features that have the same value in all samples) will be removed.

2.3 Related Works

The system utilizing Random Forest (RF) and Multi Class - Support Vector Machine (SVM) was developed for Motor Insurance Claim Status Prediction [7][10]. The system encompasses various stages including data understanding, explanatory data analysis, data preprocessing, model training, model testing, classification, prediction, and a comparison of the two models created. The dataset used consists of eleven attributes related to motor insurance claim data from AIC Company, with five target classes: close, notification, pending, re-open, and settled. To evaluate the model's performance, four metrics were employed: Accuracy, Precision, Recall, and F measure. In the domain of insurance, particularly motor insurance, the RF model exhibited slightly superior prediction accuracy compared to the SVM model.

The proposed system [13] introduced a hybrid predictive modeling approach for motor insurance claims by combining grey relational analysis with backpropagation neural network (BPNN). The performance of the predictive models, namely the hybrid model GRABPNN and the simple BPNN, was evaluated using four error measurements: mean squared error, root mean square error, mean absolute error, and mean absolute percentage error. The study provided evidence that, considering different numbers of features and hidden nodes to rank informative features, GRABPNN

outperformed other models in modeling claim frequency and claim severity for each claim type.

The prediction of motor insurance claims occurrence was approached as an imbalanced machine learning problem in the proposed system. Various algorithms, including Logistic Regression, Decision Tree, Random Forest, XGBoost, and Feed-forward Network, were employed [6]. The dataset used, called Fremotor1, described the car insurance claims and insurance policy parameters from an unknown French insurer. The primary objective of this work was to explore and implement different techniques to address the challenges posed by imbalanced datasets when predicting claim occurrences in car insurance. It is important to note that even a high-performing machine learning algorithm may not yield satisfactory results when dealing with imbalanced data. To mitigate this issue, the SMOTE oversampling technique was employed. The performance of the machine learning algorithms in the context of claim occurrence prediction in car insurance was evaluated using Accuracy and F1 score as metrics. Among the algorithms tested, XGBoost and Random Forest methods demonstrated superior accuracies compared to the other algorithms [1].

In the proposed system [15], a prediction model for auto insurance claims was developed using various machine learning techniques, including Artificial Neural Network (ANN), Decision Tree (DT), Naïve Bayes classifiers, and XGBoost. The experimental findings demonstrated that the model achieved satisfactory results. Notably, the XGBoost model and Resolution Tree exhibited the highest accuracy among the four models, achieving an accuracy of 92.53% and 92.22%, respectively.

CHAPTER 3
ARCHITECTURE OF THE PROPOSED SYSTEM

The main intention of this chapter is to present the main methodology that is the core part of this thesis book. To achieve this goal, this chapter firstly describes methodologies of three operational algorithms along with the overview design of the system. The proposed system develops two Car Insurance Claim Prediction Classifiers with Random Forest and Logistic regression based on the car insurance claim dataset respectively and compare which method is the most suitable for car insurance companies. Firstly, this proposed system creates a feature selection model using Variance Threshold Selector method to select the important attributes impact on the accuracy of car insurance claim prediction classifiers. The final selected data set is split into training with 80% and testing sets with 20% randomly and the prediction model was built using Logistic Regression (LR) and Random Forest (RF) classifiers.

There are a number of researches works related with attributes selection methods of Apache Spark Machine Learning Work. In the current work, Variance Threshold Selector method is used to select the car insurance attributes to be the more accurate classifier for Car Insurance Claim Prediction System. Logistic Regression Classifier and Random Forest Classifier are used to create two car insurance claim prediction systems respectively to select more suitable classifier for Car Insurance Companies. In accordance with this proposed approach, the following sections will describe the orientation of three algorithms in details.

3.1 Overview Design of the Proposed Car Insurance Claim Prediction

This system is revealed for tackling the car insurance claim prediction problem by offering computerized software that enables to find the most suitable prediction system for Car Insurance Companies. The Figure 3.1 depicts an overview of the proposed system.

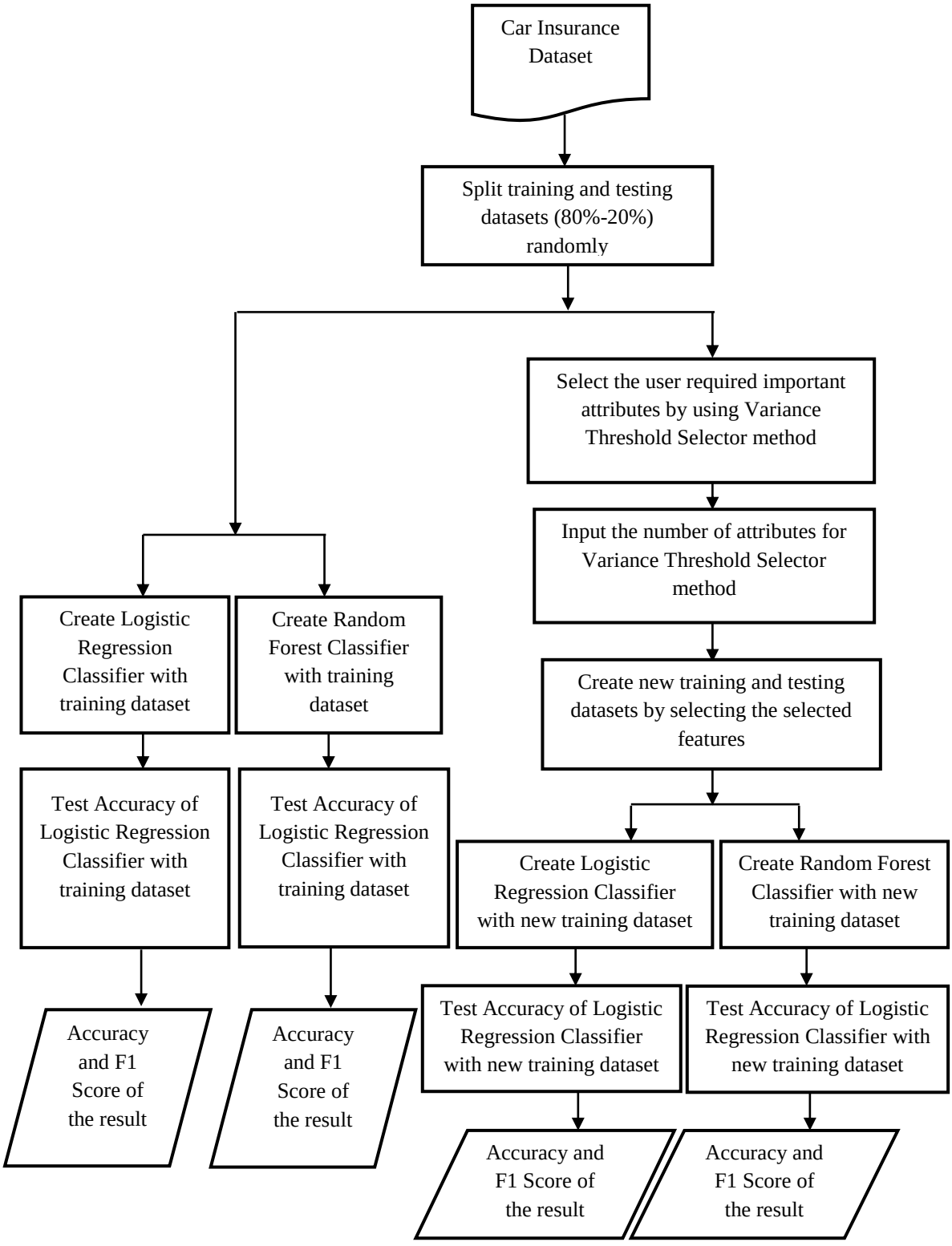

Figure 3.1 Proposed System Design of the Car Insurance Claim Prediction System based on Accuracy Measuring

The system predicts car insurance claim of customers by using LR and RF classifiers with or without using a feature selection method. To predict car insurance

claim with the feature selection method, the first task is that the system takes the car insurance claim dataset and selects the important features by using Variance Threshold Selector method based on user specified feature number. Secondly, the system splits the selected features data set into two datasets: training dataset with 80% and testing dataset with 20% and then creates two classifiers: Logistic Regression Classifier and Random Forest Classifier based on training dataset respectively. Finally, the system provides the particular prediction accuracy of two classifiers based on testing dataset. To predict car insurance claim without the feature selection method, the system takes the car insurance claim dataset and splits two data sets: training dataset with 80% and testing dataset with 20%. After preparing dataset, the system creates Logistic Regression Classifier and Random Forest Classifiers and predict the accuracy of two classifiers based on testing dataset. Proposed System Design of the Car Insurance Claim Prediction System based on User Input Car Insurance Data is shown in The Figure 3.2.

To predict car insurance claim or not in the next six months, the user inputs the car insurance claim data and predict yes or no results by using Logistic Regression Classifier with attribute selection, Random Forest Classifier without attribute selection, Logistic Regression Classifier without attribute selection and Random Forest Classifier with attribute selection.

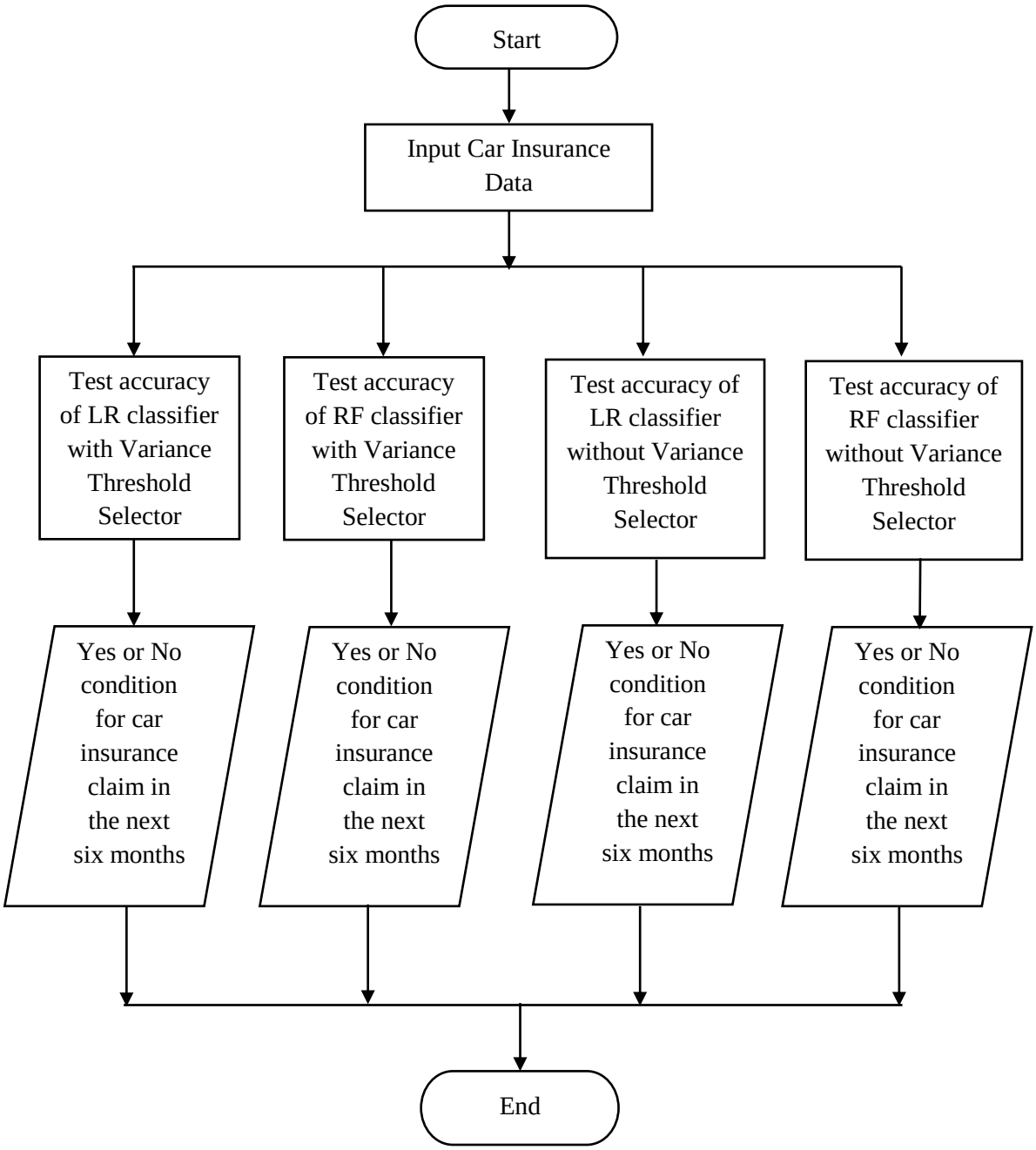

Figure 3.2 Proposed System Design of the Car Insurance Claim Prediction System based on User Input Car Insurance Data

3.2 Data Collection and Preprocessing of Car Insurance Claim Dataset

The Car Insurance Claim Prediction dataset is collected from the Kaggle website that specializes in running statistical analysis and predictive modeling competitions. The Dataset contains information on policyholders having the 44 attributes and 97656 cases.

3.2.1 Data Collection

The proposed system uses 44 attributes and 6000 cases that contain yes cases and no cases [24]. Firstly, the unimportant attribute, policy-id is removed from the total attributes that are shown in the Table 3.1.

Table 3.1 Car Insurance Claim Dataset's Attributes

Sr. No.	Car Insurance Claim Prediction Data List	Description	Data Type
1.	policy tenure	Time period of the policy	Integer
2.	age of the car	Normalized age of the car in years	Integer
3.	age of policyholder or owner	Normalized age of the policyholder in years	Integer
4.	area cluster	Area cluster of the policyholder(C1-C22)	String
5.	population density	Population density of the city (Policyholder City)	Integer
6.	Make	Encoded Manufacturer/company of the car	Integer
7.	Segment	Segment of the car(A=0/B1=1/B2=2/C1=3/C2=4/Utility=5)	String
8.	Model	Encoded name of the car (E.g, M1…M10)	String
9.	Fuel_type	Type of fuel used by the car (CNG=0, Diesel=1, Petrol=2)	String
10.	Max_torque	Maximum Torque generated by the car (Nm@rpm) e.g., (60Nm@3500rpm We divided into two Nm and rpm for Max_torque)	String
11.	Max_power	Maximum Power generated by the car (bhp@rpm) We divided into two bhp and rpm for Max_power.	String
12.	Engine_type	Type of engine used in the car(1.0 Sce=0, 1.2 L K Series Engine=1, 1.2 L K12N Dualjet=2, 1.5 L U2 CRDi=3, 1.5 Turbocharged Revotorq=4, 1.5 Turbocharged Revotron=5, F8D Petrol Engine=6, G12B=7, i-DTEC=8, K Series Dual jet=9, K10C=10)	String
13.	Airbags	Number of airbags installed in the car	Integer
14.	Is_esc	Boolean flag indicating whether Electronic Stability Control (ESC) is present in the car or not.	Boolean
15.	Is_adjustable_steering	Boolean flag indicating whether the steering wheel of the car is adjustable or not.	Boolean
16.	Is_tpms	Boolean flag indicating whether Tyre Pressure Monitoring System (TPMS) is present in the car or not.	Boolean
17.	Is_parking_sensors	Boolean flag indicating whether parking sensors are present in the car or not.	Boolean

18.	Is_parking_camera	Boolean flag indicating whether the parking camera is present in the car or not.	Boolean
19.	Rear_brakes_type	Type of brakes used in the rear of the car (Drum=0, Disc=1)	String
20.	Displacement	Engine displacement of the car (cc)	Integer
21.	Cylinder	Number of cylinders present in the engine of the car	Integer
22.	Transmission_type	Transmission type of the car (Manual=0, Electric=1, Automatic=2,)	String
23.	Gear_box	Number of gears in the car	Integer
24.	Steering_type	Type of the power steering present in the car (Electric=0 and Power=1, Manual=2)	String
25.	Turining_radiucs	The space a vehicle needs to make a certain turn (Meters)	Integer
26.	Length	Length of the car (Millimeter)	Integer
27.	Width	Width of the car (Millimeter)	Integer
28.	Height	Height of the car (Millimeter)	Integer
29.	Gross_weight	The maximum allowable weight of the fully-loaded car, including passengers, cargo and equipment (Kg)	Integer
30.	Is_front_fog_lights	Boolean flags indicating whether these are available in the car or not.	Boolean
31.	Is_rear_window_wiper	Boolean flags indicating whether these are available in the car or not.	Boolean
32.	Is_rear_window_washer,	Boolean flags indicating whether these are available in the car or not.	Boolean
33.	Is_rear_window_defogger	Boolean flags indicating whether these are available in the car or not.	Boolean
34.	Is_brake_assist,	Boolean flags indicating whether these are available in the car or not.	Boolean
35.	Is_power_door_lock	Boolean flags indicating whether these are available in the car or not.	Boolean
36.	Is_central_locking	Boolean flag indicating whether the central locking feature is available in the car or not.	Boolean
37.	Is_power_steering	Boolean flag indicating whether power steering is available in the car or not.	Boolean
38.	Is_driver_seat_height_adjustable	Boolean flag indicating whether the height of the driver seat is adjustable or not.	Boolean
39.	Is_day_night_rar_view_mirror	Boolean flag indicating whether day & night rearview mirror is present in the car or not.	Boolean
40.	Is_ecw	Boolean flag indicating whether Engine Check Warning (ECW) is available in the car or not.	Boolean
41.	Is_speed_alert	Boolean flag indicating whether the speed alert system is available in the car or not.	Boolean
42.	Ncap_rating	Safety rating given by NCAP (out of 5)	Integer
43.	Is_claim	Outcome: Boolean Flag indicating whether the policyholder file a claim in the 6 months or not.	Boolean

3.2.2 Preprocessing of Car Insurance Claim Dataset

The proposed system makes the data pre-processing step manually and convert their value types of attributes: String and Boolean into number type. Maximum Torque generated by the car (Max_torque) is divided into according to the Nm unit and RPM unit and Maximum Power generated by the car (Max_power) attribute is divided into two-unit attributes according to the BHP unit and RPM unit.

In preprocessing step, string type attributes are converted into number data according to equations (3.1) to (3.8).

$$f(area\_cluster) = \begin{cases} 1,\ if\ (area\_cluster = "C1") \\ 2, else\ if\ (area\_cluster = "C2") \\ 3, else\ if\ (area\_cluster = "C3") \\ . \\ . \\ 21, else\ if\ (area\_cluster = "C21") \\ 22,\ (otherwise) \end{cases} \tag{3.1}$$

Where f($area\_cluster$) function is converted the Area cluster attribute into number data based on x's string value.

$$f(segment\_attribute) = \begin{cases} 0, if\ (segment\_attribute = "A") \\ 1, else\ if\ (segment\_attribute = "B1") \\ 2, else\ if\ (segment\_attribute = "B2") \\ 3, else\ if\ (segment\_attribute = "C1") \\ 4, else\ if\ (segment\_attribute = "C2") \\ 5, (otherwise) \end{cases} \tag{3.2}$$

Where f($segment\_attribute$) function is converted the $segment\_attribute$ of the car into number data based on y's string value.

$$f(model) = \begin{cases} 1, if\ (model = "M1") \\ 2,\ else\ if\ (model = "M2") \\ 3,\ else\ if\ (model = "M3") \\ . \\ . \\ 9,\ else\ if\ (model = "M10") \\ 10, (otherwise) \end{cases} \tag{3.3}$$

Where f(model) function is converted the value's data type of the car model attribute into number data type based on model's string value.

$$f(fuel\_type) = \begin{cases} 0, if\ (fuel\_type = "CNG") \\ 1, else\ if\ (fuel\_type = "Diesel") \\ 2,\ (otherwise) \end{cases} \quad (3.4)$$

Where f(fuel_type) function is converted the value's data type of the fuel type attribute into number data type based on fuel type attribute's string value.

$$f(engine\_type) = \begin{cases} 0,\ if\ (engine\_type = "1.0\ Sce") \\ 1,\ else\ if\ (engine\_type = "1.2\ L\ K\ Series\ Engine") \\ 2,\ else\ if\ (engine\_type = "1.2\ L\ K12N\ Dualjet") \\ 3,\ else\ if\ (engine\_type = "1.5\ L\ U2\ CRDi") \\ 4, else\ if\ (engine\_type = "1.5\ Turbocharged\ Revotorq") \\ 5, else\ if\ (engine\_type = "1.5\ Turbocharged\ Revotron") \\ 6, else\ if\ (engine\_type = "F8D\ Petrol\ Engine") \\ 7, else\ if\ (engine\_type = "G12B") \\ 8, else\ if\ (engine\_type = "\ i-DTEC") \\ 9, else\ if\ (engine\_type = "K\ Series\ Dual\ jet") \\ 10, (otherwise) \end{cases} \quad (3.5)$$

Where f($engine\_type$) function is converted the value's data string type of engine type attribute into number data type based on engine type of the car.

$$f(rear\_brakes\_type) = \begin{cases} 0,\ if\ (rear\_brakes\_type = "Drum") \\ 1,\ (otherwise) \end{cases} \quad (3.6)$$

Where f($rear\_brakes\_type$) function is converted the value's data type of the rear brakes type attribute into number data type based on their string value.

$$f(transmission\_type) = \begin{cases} 0, if\ (transmission\_type = "Manual") \\ 1, else\ if\ (transmission\_type = "Electric") \\ 2, (otherwise) \end{cases} \quad (3.7)$$

Where f(transmission_type) function is converted the value's data type of the transmission type attribute into number data type based on fuel type attribute's string value.

$$f(steering\_type) = \begin{cases} 0, if\ (steering\_type = "Electric") \\ 1, else\ if\ (steering\_type = "Power") \\ 2, (otherwise) \end{cases} \quad (3.8)$$

Where f(steering_type) function is converted the value's data type of the power steering present in the car into number data type based on steering_type attribute's string value.

3.3 Apache Spark ML Lib

Spark is an open source processing engine, which uses directed acyclic graph and its own data structure i.e., Resilient Distributed Dataset (RDD) to provide speed and analytics. Spark helps in some challenging and computationally exhaustive tasks like processing high volumes of real-time and archived data, thereby integrating the complex capabilities such as ML and graph algorithms. It brings big data processing to the market and Spark has a library for ML labelled as MLib. Spark MLib library has algorithms for the functions of classification, regression, clustering, collaborative filtering, dimensionality reduction, etc. The proposed system applies the car insurance claim prediction system using Spark MLib library with a feature selection method and two classification methods.

3.3.1 Feature Selection with Variance Threshold Selector Method

The system uses Feature selection method, Variance Threshold Selector to choosea subset of the most important car insurance claim features while trying to retain as much information as possible for car insurance companies. This method is a selector that removes low-variance car insurance claim features that have the same value in all records. Features with a variance not greater than the variance Threshold will be removed. This technique is a quick and lightweight way of eliminating features with very low variance, i. e. features with not much useful information [27].

Variance Method:

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{3.9}$$

Where, n is the number of records, x_i is the value at position i and x is the mean of particular attribute. The proposed system calculated car insurance claim attributes' variances and removes attributes with low variance according to the user specified needed attributes number.

3.3.2 Logistic Regression Classifier

Logistic regression is the algorithm that is used for classification. Spark's logistic regression API is useful for binary classification, or classifying input data into one of two groups. The proposed system accepts car insurance claim information as

input. Then divides the information into two categories: customers that is claim car insurance in the next six month and customers that car insurance company should keep [25]. To predict car insurance claim, the process of logistic regression produces a logistic function. Use the function to predict the probability that an input vector belongs in one group or the other. The Logistic regression equation can be obtained from the Linear Regression equation:

$$\log[\frac{y}{1-y}] = b_0+b_1x_1+b_2x_2+\ldots\ldots\ldots\ldots+b_nx_n \quad (3.10)$$

The above equation is the final equation for Logistic Regression where b are the regression coefficients, x_i is the real data and y is the predict class data of the particular record and $\log[\frac{y}{1-y}]$ can be between -[infinity] for y=0 and +[infinity] for y=1. Finally, this proposed system uses Logistic Regression Classifier to predict the car insurance claim condition.

3.3.3 Random Forest Classifier

Random Forest works in two-phase first is to create the random forest by combining N decision tree, and second is to make predictions for each tree created in the first phase. The classifier can predict accurate results rather than a guessed result when the predictions from each tree must have very low correlations [22]. Random Forest algorithm has many advantages:

- 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.

The Random Forest working process can be explained in the below steps:

- Step-1: Select random K data points from the training set.
- Step-2: Build the decision trees associated with the selected data points (Subsets).
- Step-3: Choose the number N for decision trees that you want to build.
- Step-4: Repeat Step 1 & 2.
- Step-5: For new data points, find the predictions of each decision tree, and assign the new data points to the category that wins the majority votes.

In this system, the classifier creates decision trees according to the car insurance claim dataset and then creates the random forest by combining car insurance claim decision trees to predict accurate results.

3.4 Chapter Summary

This chapter has firstly presented system overview design with figure and detailed explanation of how works the car insurance claim prediction system. Secondly, the system demonstrates how to collect and preprocess car insurance claim dataset with equations. Finally, a feature selection method and two classification methods of Apache spark machine learning library (ML lib) are explained according to the car insurance claim prediction system. Therefore, car insurance claim prediction system uses Feature Selection with Variance Threshold Selector method and two Logistic Regression and Random Forest methods to predict accurate results.

CHAPTER 4

IMPLEMENTATION AND EXPERIMENTAL RESULTS

This chapter presents the system implementation of the car insurance claim prediction system and its performance evaluation with accuracy and F-score metrics.

4.1 System Implementation

The car insurance claim prediction system is implemented with java programming language and apache spark 2.4.0 on Windows operating system. In this section, the processes are grouped according to their characteristics namely: Variance Threshold Selector, Training and Testing (80%-20%), Testing with Variance Threshold Selector and Evaluation.

4.1.1 Feature Selection System Menu of the Car Insurance Claim Prediction System

When the system is started, the main menu is appeared and it contains four tabs: 'Facture Selection', 'Training and Testing (80%-20%)', 'Testing with Variance Threshold Selector' and 'Evaluation' that are shown in Figure 4.1.

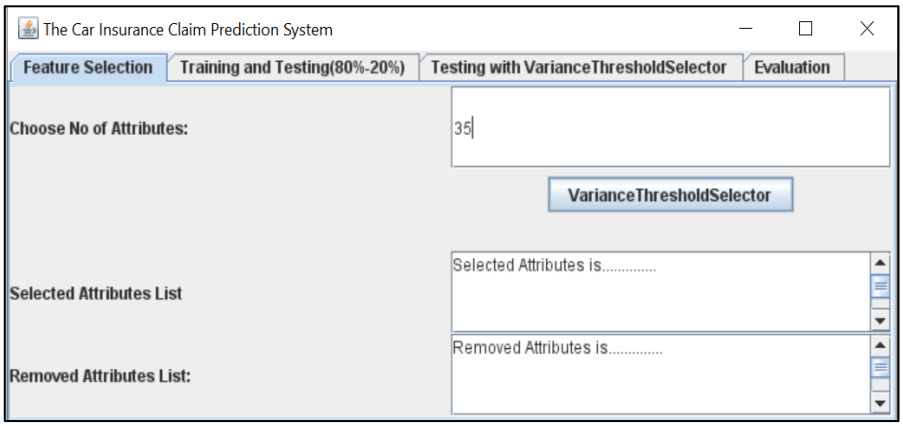

Figure 4.1 Main Menu of the car insurance claim prediction system

In feature selection pane, the user types the number of attributes and press 'Variance Threshold Selector Button'. The system selects height variance care insurance claim attributes and displays the selected 35 attributes in the textbox of 'Selected Attributes List'. The nine low variance attributes are also show in the textbox of 'Removed Attributes List' that are displayed in Figure 4.2.

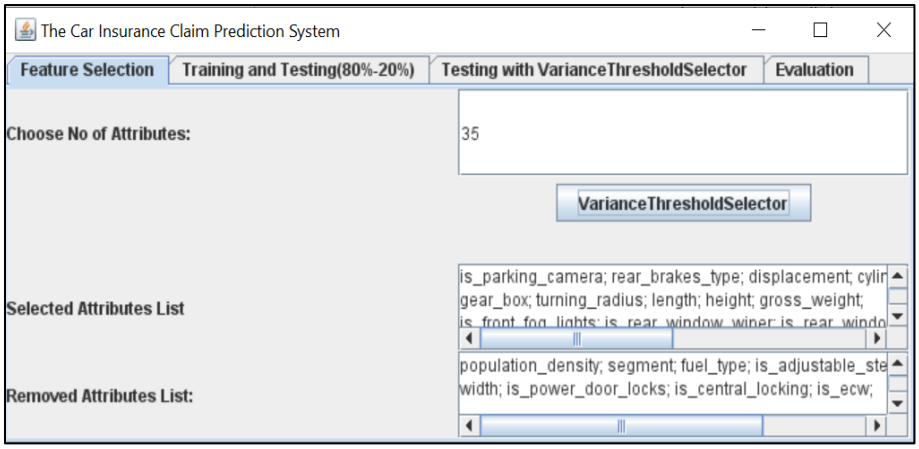

Figure 4.2 Selected Attributes of the car insurance claim prediction system

The Table 4.1 shows the selected attributes and removed attributes based on user specified number of attribute number.

Table 4.1 The Selected Attributes of the Car Insurance Claim Prediction System based on User Specified Attributes Number

Sr. No.	User Specified Number of Attributes	Selected Attributes of the Car Insurance Claim Prediction System	Removed Attributes of the Car Insurance Claim Prediction System
1.	30	policy_tenure; age_of_car; age_of_policyholder; area_cluster; make; model; max_torque (Nm); max_torque (RPM); max_power (RPM); engine_type; airbags; is_esc; is_parking_sensors;is_parking_camera; displacement; transmission_type; gear_box; turning_radius; length; height; is_front_fog_lights; is_rear_window_wiper; is_rear_window_washer; is_rear_window_defogger; is_brake_assist; is_power_steering; is_driver_seat_height_adjustable; is_day_night_rear_view_mirror; is_speed_alert; ncap_rating;	population_density; segment; fuel_type; max_power (BHP); is_adjustable_steering; is_tpms; rear_brakes_type; cylinder; steering_type; width; gross_weight; is_power_door_locks; is_central_locking; is_ecw;
2.	32	policy_tenure; age_of_car; age_of_policyholder; area_cluster; make; model; max_torque (Nm); max_torque (RPM); max_power (RPM); engine_type; airbags; is_esc; is_tpms; is_parking_sensors; is_parking_camera; rear_brakes_type; displacement; transmission_type; gear_box; turning_radius; length; height;	population_density; segment; fuel_type; max_power (BHP); is_adjustable_steering; cylinder; steering_type; width; gross_weight;

		is_front_fog_lights; is_rear_window_wiper; is_rear_window_washer; is_rear_window_defogger; is_brake_assist; is_power_steering; is_driver_seat_height_adjustable; is_day_night_rear_view_mirror; is_speed_alert; ncap_rating;	is_power_door_locks; is_central_locking; is_ecw;
3.	34	policy_tenure; age_of_car; age_of_policyholder; area_cluster; make; model; max_torque (Nm); max_torque (RPM); max_power (BHP); max_power (RPM); engine_type; airbags; is_esc; is_tpms; is_parking_sensors; is_parking_camera; rear_brakes_type; displacement; cylinder; transmission_type; gear_box; turning_radius; length; height; is_front_fog_lights; is_rear_window_wiper; is_rear_window_washer; is_rear_window_defogger; is_brake_assist; is_power_steering; is_driver_seat_height_adjustable; is_day_night_rear_view_mirror; is_speed_alert; ncap_rating;	population_density; segment; fuel_type; is_adjustable_steering; steering_type; width; gross_weight; is_power_door_locks; is_central_locking; is_ecw;
4.	36	policy_tenure; age_of_car; age_of_policyholder; area_cluster; population_density; make; model; max_torque (Nm); max_torque (RPM); max_power (BHP); max_power (RPM); engine_type; airbags; is_esc; is_tpms; is_parking_sensors; is_parking_camera; rear_brakes_type; displacement; cylinder; transmission_type; gear_box; turning_radius; length; height; gross_weight; is_front_fog_lights; is_rear_window_wiper; is_rear_window_washer; is_rear_window_defogger; is_brake_assist; is_power_steering; is_driver_seat_height_adjustable; is_day_night_rear_view_mirror; is_speed_alert; ncap_rating;	segment; fuel_type; is_adjustable_steering; steering_type; width; is_power_door_locks; is_central_locking; is_ecw;

5.	38	policy_tenure; age_of_car; age_of_policyholder; area_cluster; population_density; make; model; fuel_type; max_torque (Nm); max_torque (RPM); max_power (BHP); max_power (RPM); engine_type; airbags; is_esc; is_adjustable_steering; is_tpms; is_parking_sensors; is_parking_camera; rear_brakes_type; displacement; cylinder; transmission_type; gear_box; turning_radius; length; height; gross_weight; is_front_fog_lights; is_rear_window_wiper; is_rear_window_washer; is_rear_window_defogger; is_brake_assist; is_power_steering; is_driver_seat_height_adjustable; is_day_night_rear_view_mirror; is_speed_alert; ncap_rating;	segment; steering_type; width; is_power_door_locks; is_central_locking; is_ecw;
6.	40	policy_tenure; age_of_car; age_of_policyholder; area_cluster; population_density; make; model; fuel_type; max_torque (Nm); max_torque (RPM); max_power (BHP); max_power (RPM); engine_type; airbags; is_esc; is_adjustable_steering; is_tpms; is_parking_sensors; is_parking_camera; rear_brakes_type; displacement; cylinder; transmission_type; gear_box; steering_type; turning_radius; length; width; height; gross_weight; is_front_fog_lights; is_rear_window_wiper; is_rear_window_washer; is_rear_window_defogger; is_brake_assist; is_power_steering; is_driver_seat_height_adjustable; is_day_night_rear_view_mirror; is_speed_alert; ncap_rating;	segment; is_power_door_locks; is_central_locking; is_ecw;
7.	42	policy_tenure; age_of_car; age_of_policyholder; area_cluster; population_density;	is_central_locking; is_ecw;

		make; segment; model; fuel_type; max_torque (Nm); max_torque (RPM); max_power (BHP); max_power (RPM); engine_type; airbags; is_esc; is_adjustable_steering; is_tpms; is_parking_sensors; is_parking_camera; rear_brakes_type; displacement; cylinder; transmission_type; gear_box; steering_type; turning_radius; length; width; height; gross_weight; is_front_fog_lights; is_rear_window_wiper; is_rear_window_washer; is_rear_window_defogger; is_brake_assist; is_power_door_locks; is_power_steering; is_driver_seat_height_adjustable; is_day_night_rear_view_mirror; is_speed_alert; ncap_rating;	

4.1.2 Training (80%) and Testing (20%) Result of the Car Insurance Claim Prediction System

After the feature selection process, the system creates the new dataset and splits that into two parts: training data set with 80 % and testing dataset with 20% randomly. The training dataset has used two classifiers: Linear Regression Classifier and Random Forest Classifier. Similarly, the old datasets are divided into two datasets: training dataset with 80 % and testing dataset with 20% and used to create two Linear Regression and Random Forest Classifies. Finally, the testing datasets are used to produce the accuracy results of four classifiers with accuracy and F1 Score values that are shown in Figure 4.3.

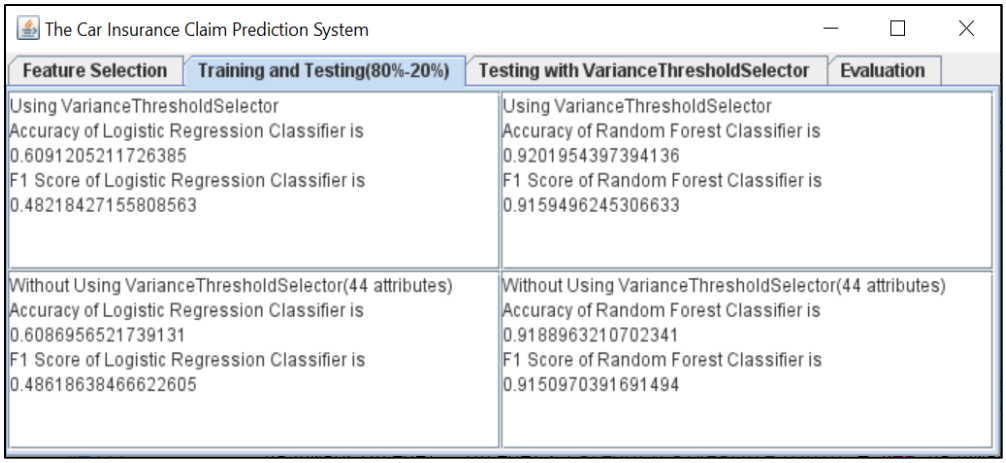

Figure 4.3 Accuracy and F1 score of the car insurance claim prediction system

According to the accuracy result, lthough Logistic Regression Classifier is not suitable for the care insurance claim prediction data nature, Random Forest Classifier is more suitable classifier for this data nature with above 90% of accuracy and F1 score values respectively.

4.1.3 Classification of the Car Insurance Claim Prediction System for User Input Data

After the process of creation of four classifiers, the users can input the car insurance claim data with the specified attributes number. Moreover, the removed attributes, low variance attributes, are disable that shown with label with “Removed Attribute”. According to Figure 4.4, the two classifiers with the feature selection method can predict the customer is not claim this car insurance in the next six months correctly.

The system can show classification results of two classification with all attributes by clicking ‘Add Removed Attributes” button that are shown in Figure 4.5. According to the results, these two classifications can specify the result of this care insurance claim record correctly.

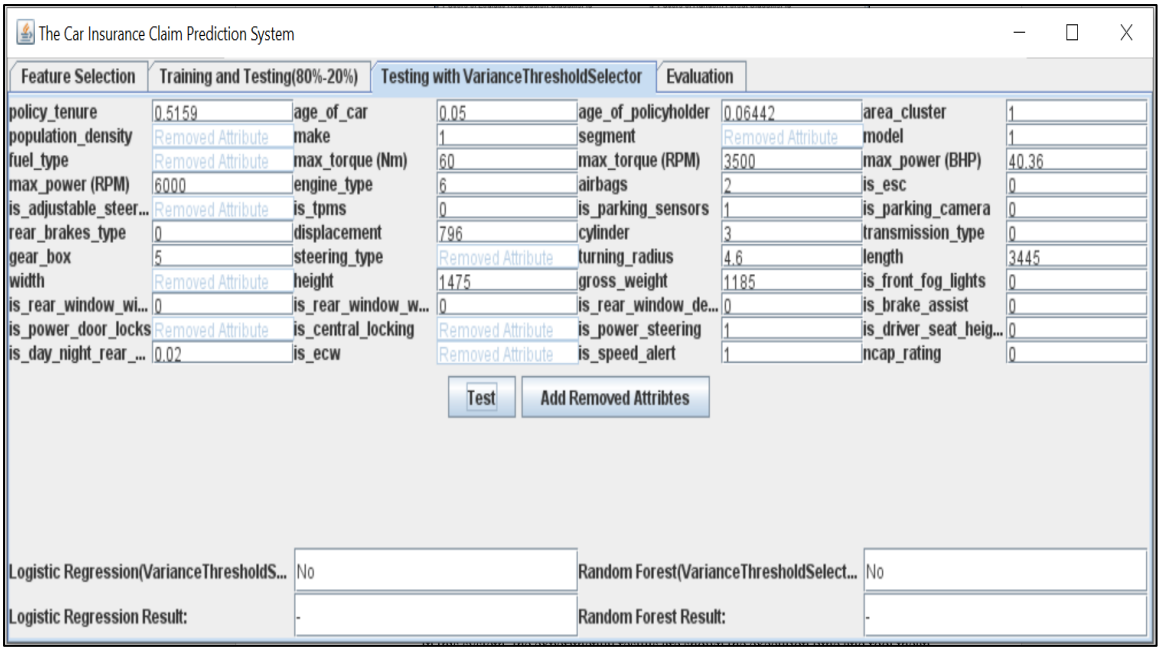

Figure 4.4 Classification result of the car insurance claim prediction system with the selected feature

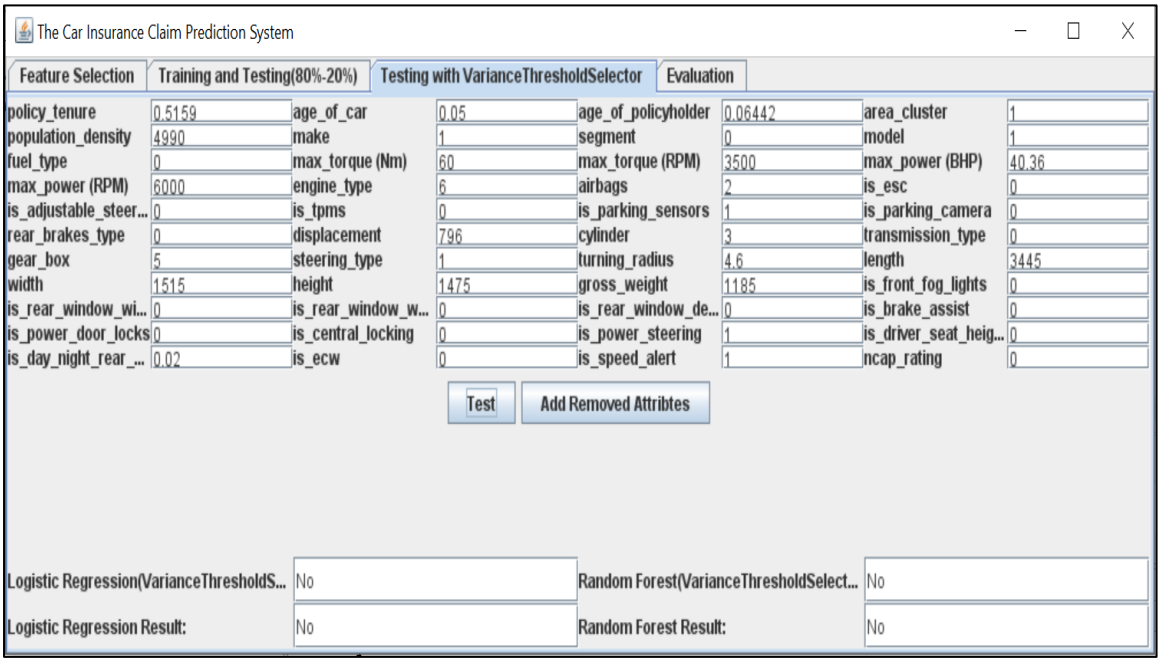

Figure 4.5 Classification result of the car insurance claim prediction system with all attributes

4.1.4 Evaluation Result Illustration According to the Selected Attribute Number

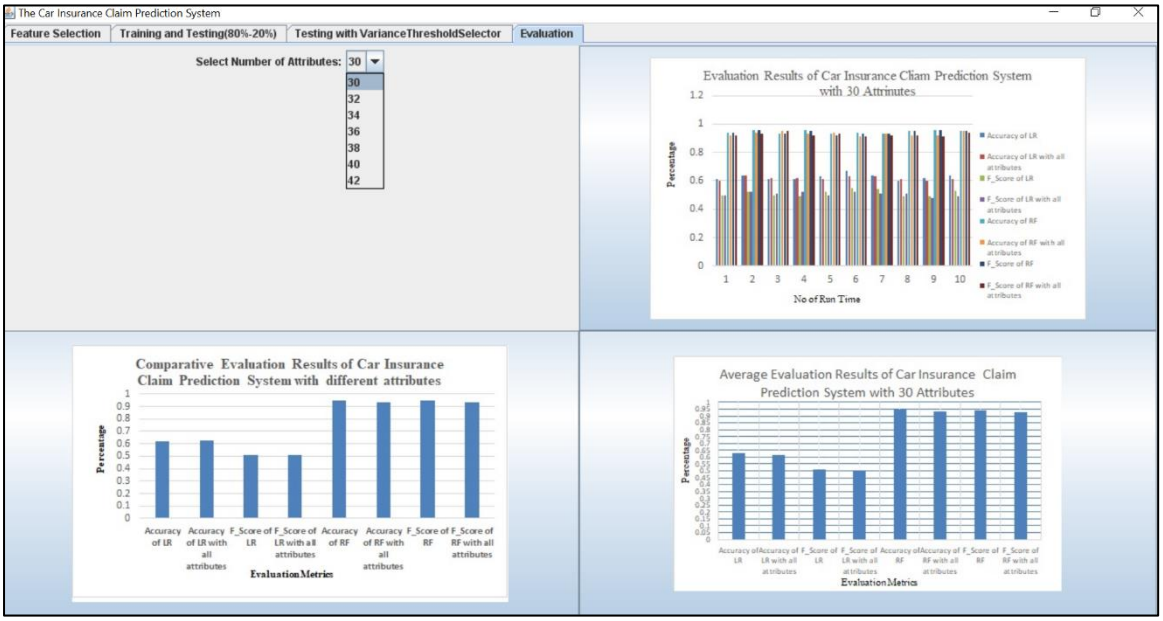

Figure 4.6 Evaluation Result Illustration of Attribute Number: 30

Figure 4.6 show the evaluation result of Attribute Number 30 and the evaluation result of 10 time testing for attribute number 30.

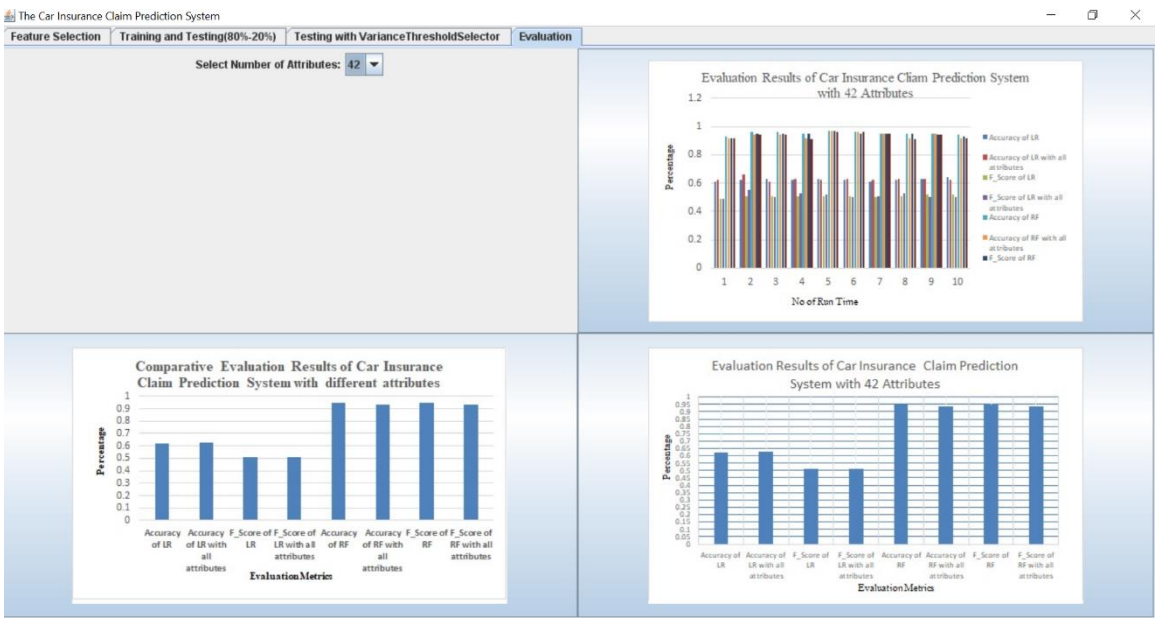

Figure 4.7 Evaluation Result Illustration of Attribute Number: 40

Figure 4.7 show the evaluation result of Attribute Number 40 and the evaluation result of 10 time testing for attribute number 40.

4.2 Experimental Results

In this system, the experimental results are shown with accuracy and F1 score measuring metrics by supporting to compare which classifiers is the most suitable for the car insurance claim dataset. To analyze the experiment result, the number of attributes: 30, 32, 34, 36, 38, 40, 42 are taken to measure the accuracy and f-score and each attribute selection run ten times because of splitting training and testing randomly.

4.2.1 Evaluation Result of the Car Insurance Claim Prediction System

For the number of 30 attributes based on 10 tests, accuracy and f-score results of LR classifier are less than 65 percent and accuracy and f-score results of RF classifier are greater than 90 percent. So, LR classifier is not suitable for this car insurance claim dataset nature. RF classifier is suitable for this proposed system. Moreover, accuracy and f-score values of RF classifier with feature selection method are greater than RF classifier with origin attributes that are shown in Figure 4.8.

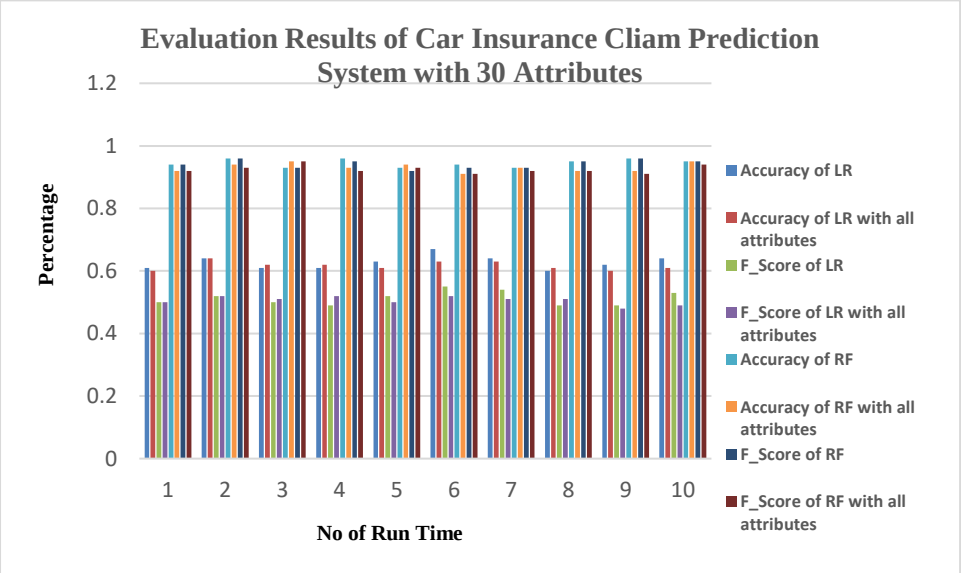

Figure 4.8 Comparison of 30 Attributes Evaluation Results Based on 10 Tests

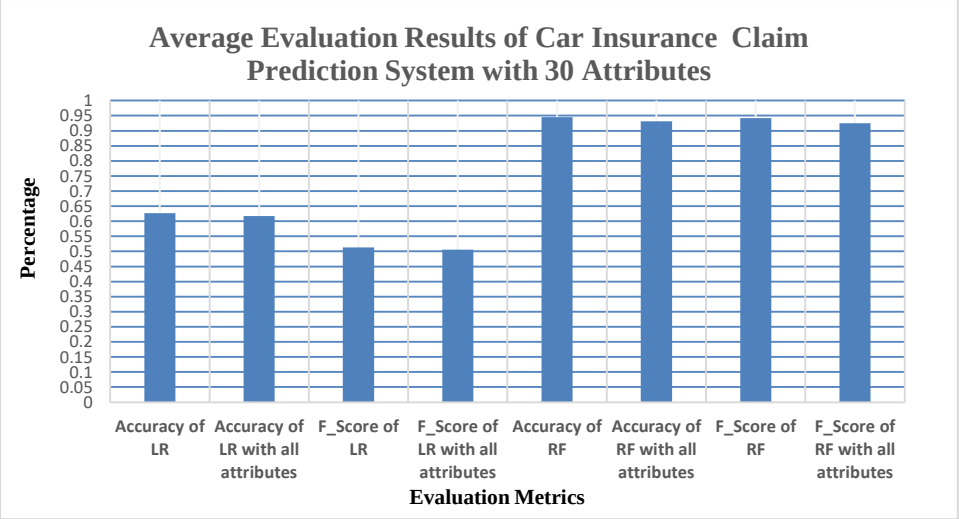

Figure 4.9 Comparison of Accuracy and F-Score Values for LR and RF Based on 30 Attributes

Figure 4.9 demonstrates evaluation metric values of four classifiers: LR with and without attribute selection method and RF with and without attributes selection method. Accuracy and f-score results of LR classifiers are less than 65 percent. Accuracies of RF with the attribute selection method and without the attribute selection method are 0.945 percent and 0.931 percent respectively. F scores of RF with the attribute selection method and without the attribute selection method are 0.942 percent and 0.925 percent respectively. Therefore, RF classifier with the attribute selection method is more suitable for the proposed car insurance claim prediction system based on 30 attributes.

For the number of 32 attributes based on 10 tests, accuracy and f-score results of LR classifier are less than 63 percent and accuracy and f-score results of RF classifier are greater than 87 percent. So, LR classifier is not suitable for this car insurance claim dataset nature. RF classifier is suitable for this proposed system. Moreover, accuracy and f-score values of RF classifier with feature selection method are greater than RF classifier with origin attributes that are shown in Figure 4.10.

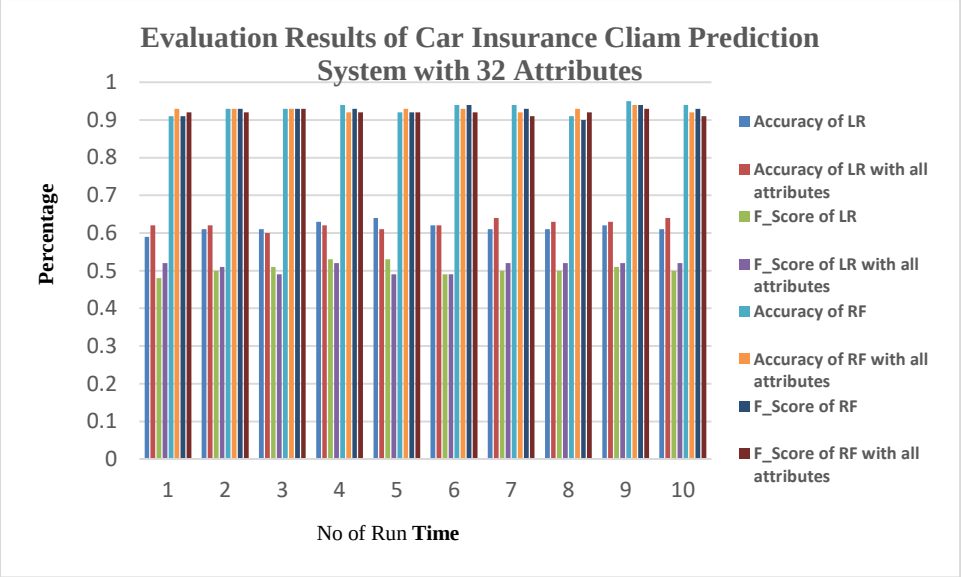

Figure 4.10 Comparison of 32 Attributes Evaluation Results Based on 10 Tests

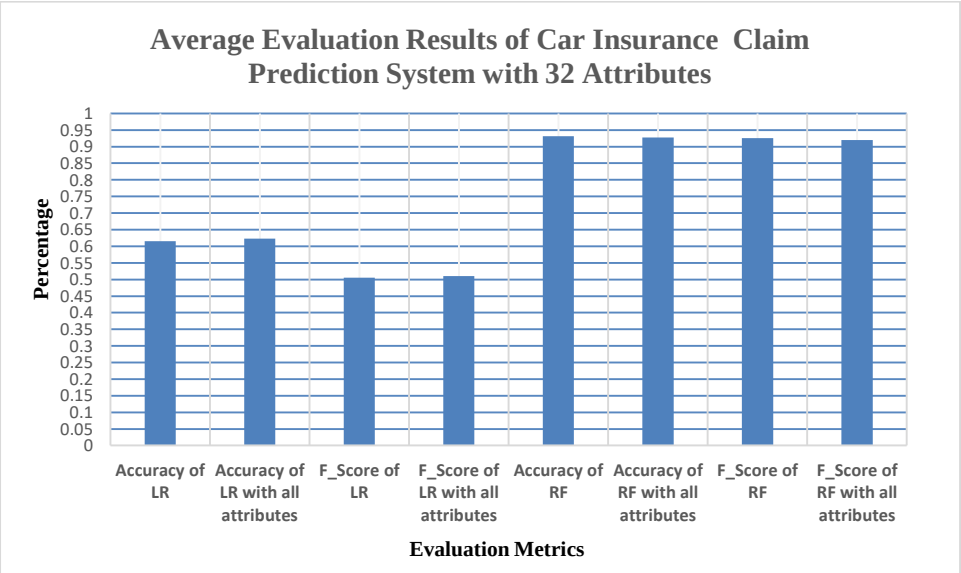

Figure 4.11 Comparison of Accuracy and F-Score Values for LR and RF Based on 32 Attributes

Figure 4.11 demonstrates evaluation metric values of four classifiers: LR with and without attribute selection method and RF with and without attributes selection method. Accuracy and f-score results of LR classifiers are less than 63 percent. Accuracies of RF with the attribute selection method and without the attribute selection method are 0.931 percent and 0.928 percent respectively. F scores of RF with the

attribute selection method and without the attribute selection method are 0.926 percent and 0.92 percent, respectively. Therefore, RF classifier with the attribute selection method is more suitable for the proposed car insurance claim prediction system based on 32 attributes.

For the number of 34 attributes based on 10 tests, accuracy and f-score result of LR classifier are less than 64 percent and accuracy and f-score results of RF classifier are greater than 86 percent. So, LR classifier is not suitable for this car insurance claim dataset nature. RF classifier is suitable for this proposed system. Moreover, accuracy and f-score values of RF classifier with feature selection method are greater than RF classifier with origin attributes that are shown in Figure 4.12.

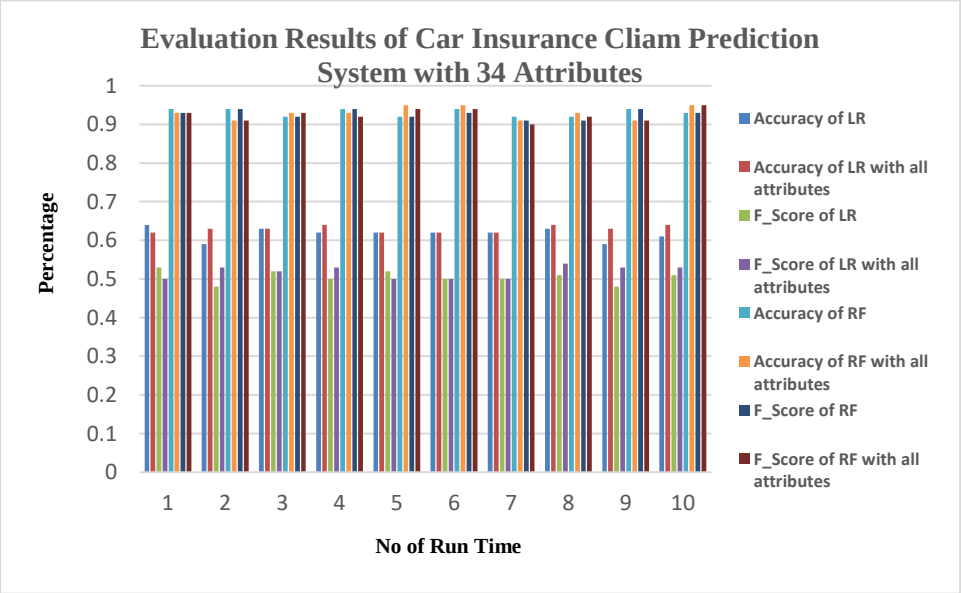

Figure 4.12 Comparison of 34 Attributes Evaluation Results Based on 10 Tests

Figure 4.13 demonstrates evaluation metric values of four classifiers: LR with and without attribute selection method and RF with and without attributes selection method. Accuracy and f-score results of LR classifiers are less than 63 percent. Accuracies of RF with the attribute selection method and without the attribute selection method are 0.931 percent and 0.93 percent respectively. F scores of RF with the attribute selection method and without the attribute selection method are 0.927 percent and 0.925 percent respectively. For 34 attributes, evaluation metric values of two classifiers: RF with and without attributes selection methods produce similar results.

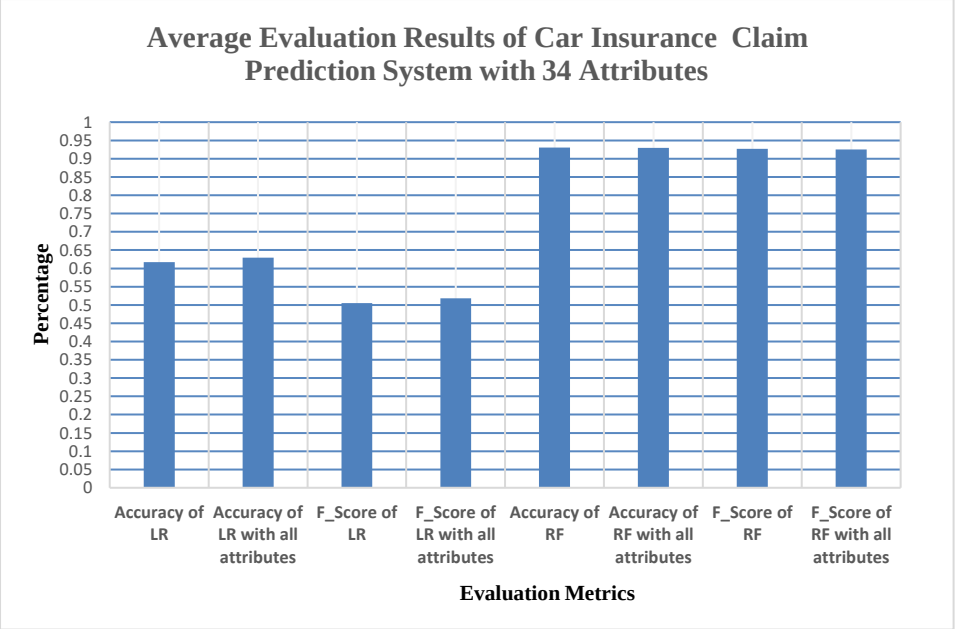

Figure 4.13 Comparison of Accuracy and F-Score Values for LR and RF Based on 34 Attributes

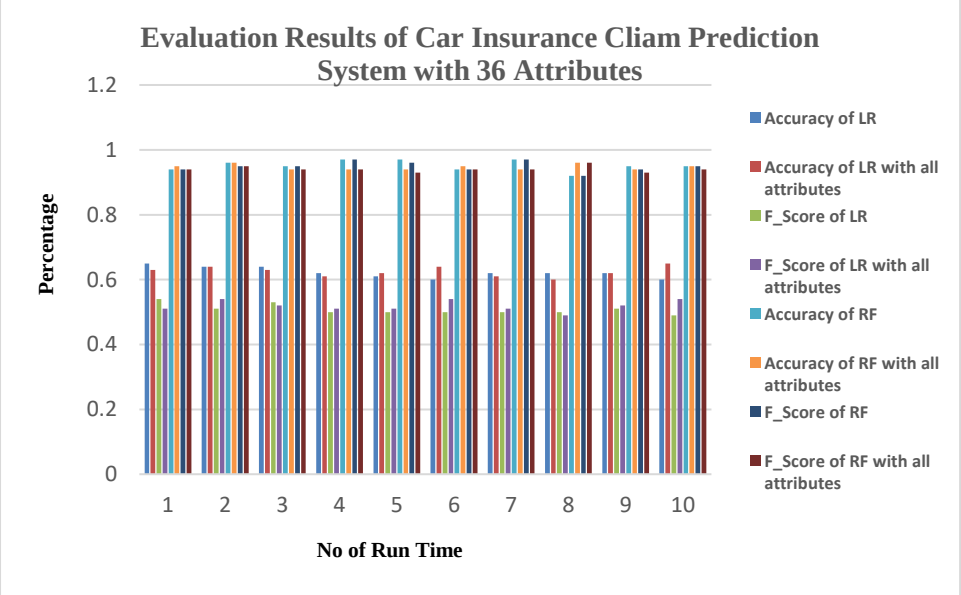

Figure 4.14 Comparison of 36 Attributes Evaluation Results Based on 10 Tests

For the number of 36 attributes based on 10 tests, accuracy and f-score results of LR classifier are less than 65 percent and accuracy and f-score results of RF classifier are greater than 91 percent. So, LR classifier is not suitable for this car insurance claim dataset nature.

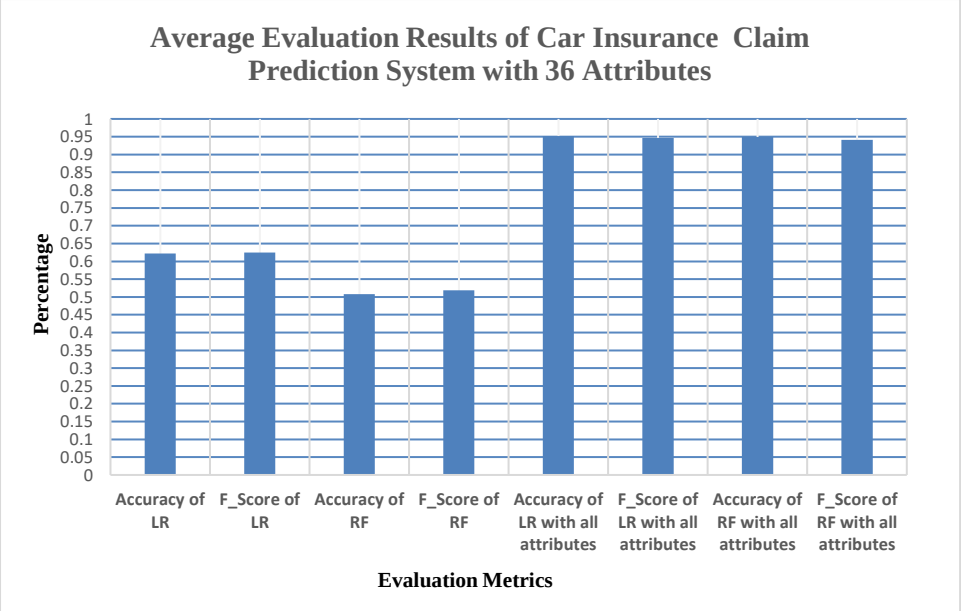

Figure 4.15 Comparison of Accuracy and F-Score Values for LR and RF Based on 36 Attributes

RF classifier is suitable for this proposed system. Moreover, accuracy and f-score values of RF classifier with feature selection method are greater than RF classifier with origin attributes that are shown in Figure 4.14.

Figure 4.15 demonstrates evaluation metric values of four classifiers: LR with and without attribute selection methods and RF with and without attributes selection methods. Accuracy and f-score results of LR classifiers are less than 63 percent. Accuracies of RF with the attribute selection method and without the attribute selection method are 0.952 percent and 0.947 percent respectively. F scores of RF with the attribute selection method and without the attribute selection method are 0.949 percent and 0.941 percent respectively. Therefore, RF classifier with the attribute selection method is more suitable for the proposed car insurance claim prediction system based on 36 attributes.

For the number of 38 attributes based on 10 tests, accuracy and f-score result of LR classifier are less than 66 percent and accuracy and f-score result of RF classifier are greater than 92 percent. So, LR classifier is not suitable for this car insurance claim dataset nature. RF classifier is suitable for this proposed system. Moreover, accuracy and f-score values of RF classifier with feature selection method are greater than RF classifier with origin attributes that are shown in Figure 4.16.

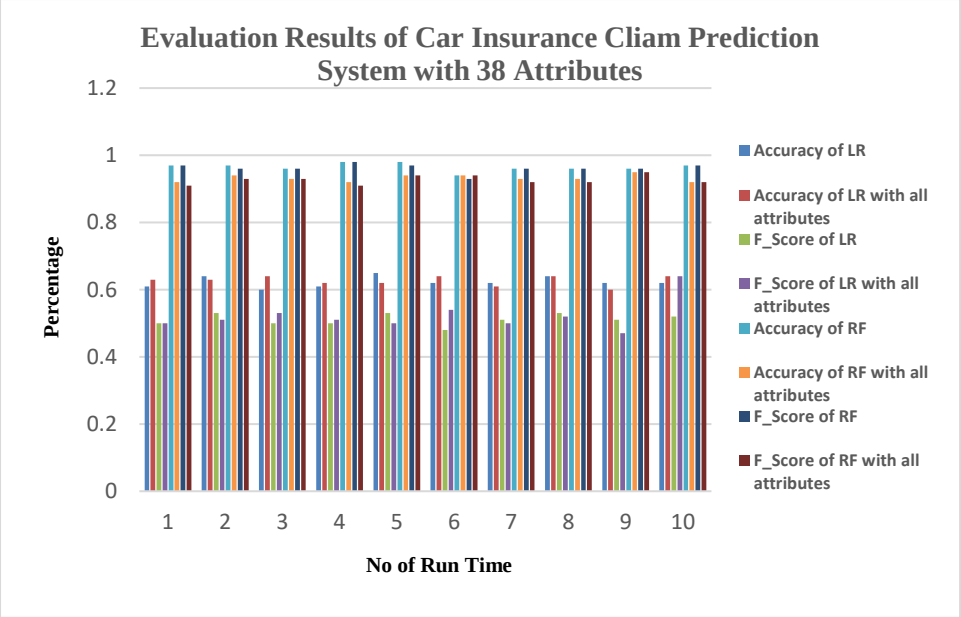

Figure 4.16 Comparison of 38 Attributes Evaluation Results Based on 10 Tests

Figure 4.17 demonstrates evaluation metric values of four classifiers: LR with and without attribute selection methods and RF with and without attributes selection methods. Accuracy and f-score results of LR classifiers are less than 63 percent. Accuracies of RF with the attribute selection method and without the attribute selection method are 0.965 percent and 0.932 percent respectively. F scores of RF with the attribute selection method and without the attribute selection method are 0.965 percent and 0.927 percent, respectively. Therefore, RF classifier with the attribute selection method is more suitable for the proposed car insurance claim prediction system based on 38 attributes.

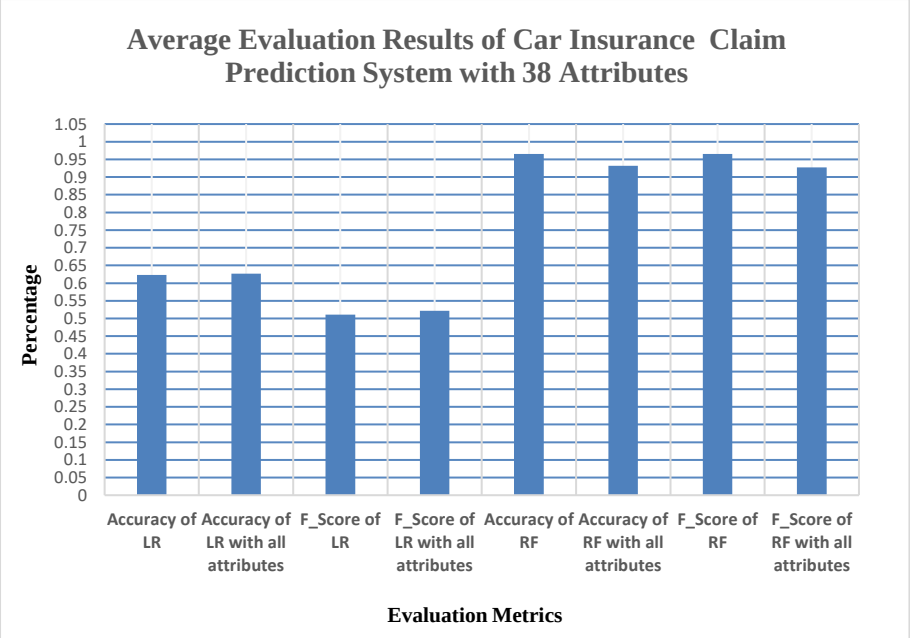

Figure 4.17 Comparison of Accuracy and F-Score Values for LR and RF Based on 38 Attributes

For the number of 40 attributes based on 10 tests, accuracy and f-score result of LR classifier are less than 65 percent and accuracy and f-score results of RF classifier are greater than 92 percent. So, LR classifier is not suitable for this car insurance claim dataset nature. RF classifier is suitable for this proposed system. Moreover, accuracy and f-score values of RF classifier with feature selection method are greater than RF classifier with origin attributes that are shown in Figure 4.18.

Figure 4.19 demonstrates evaluation metric values of four classifiers: LR with and without attribute selection method and RF with and without attributes selection method. Accuracy and f-score results of LR classifiers are less than 62 percent. Accuracies of RF with the attribute selection method and without the attribute selection method are 0.966 percent and 0.948 percent respectively.

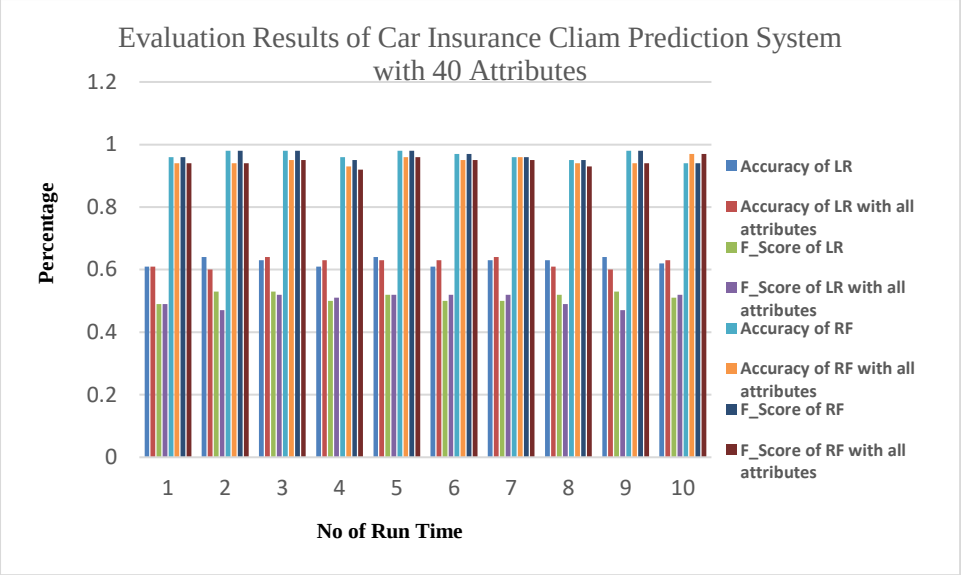

Figure 4.18 Comparison of 40 Attributes Evaluation Results Based on 10 Tests

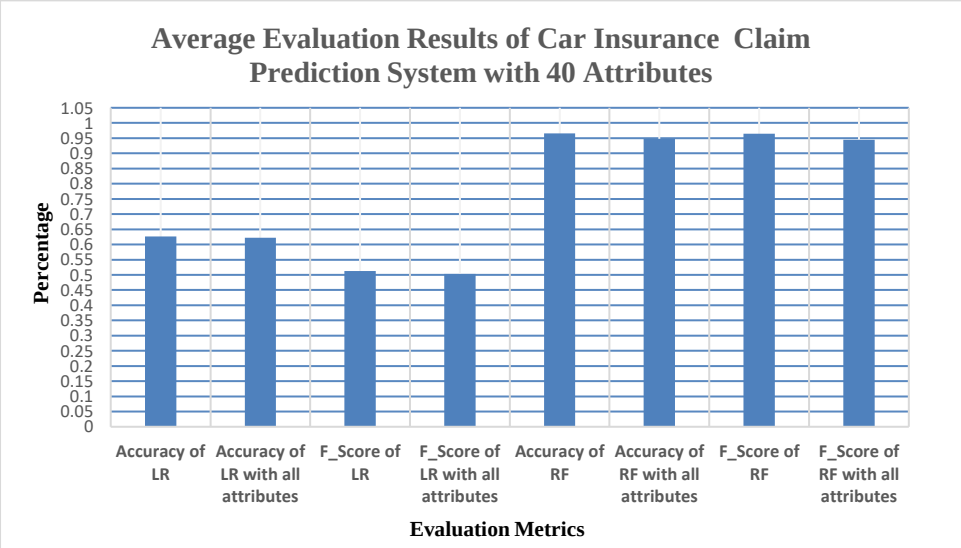

Figure 4.19 Comparison of Accuracy and F-Score Values for LR and RF Based on 40 Attributes

F scores of RF with the attribute selection method and without the attribute selection method are 0.965 percent and 0.945 percent, respectively. Therefore, RF classifier with the attribute selection method is more suitable for the proposed car insurance claim prediction system based on 40 attributes.

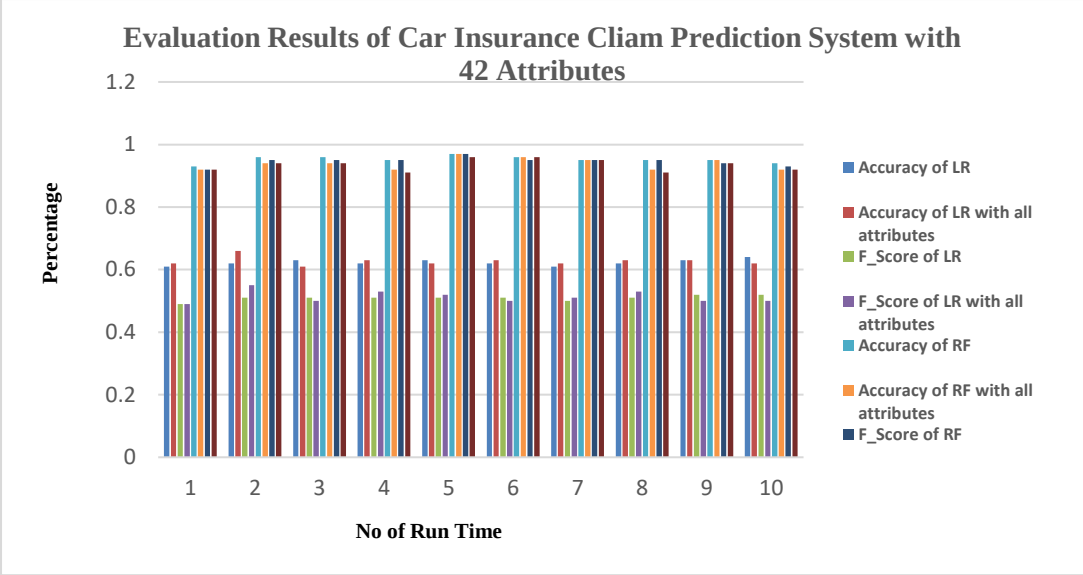

Figure 4.20 Comparison of 42 Attributes Evaluation Results Based on 10 Tests

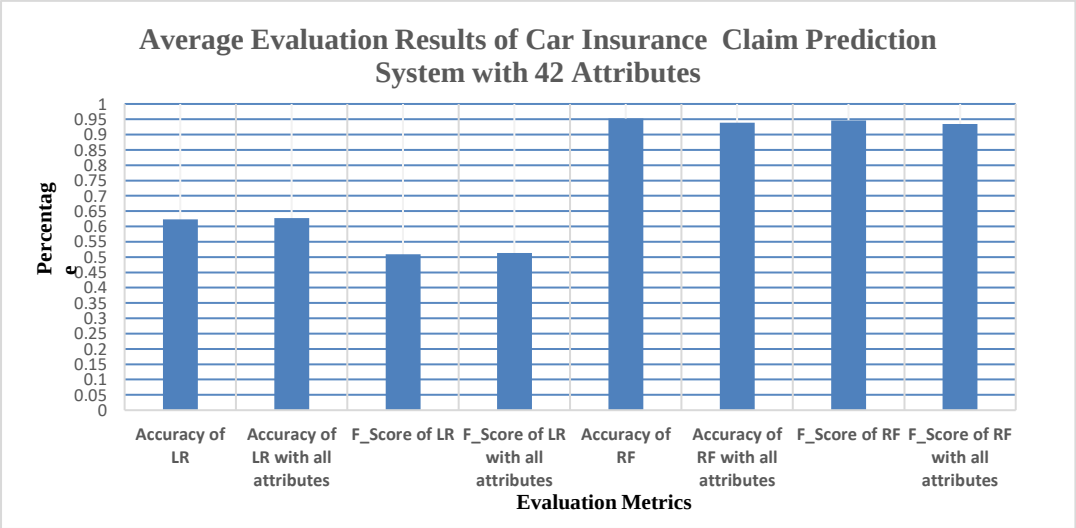

Figure 4.21 Comparison of Accuracy and F-Score Values for LR and RF Based on 42 Attributes

For the number of 42 attributes based on 10 tests, accuracy and f-score results of LR classifier are less than 62 percent and accuracy and f-score results of RF classifier are greater than 91 percent. So, LR classifier is not suitable for this car insurance claim dataset nature. RF classifier is suitable for this proposed system. Moreover, accuracy and f-score values of RF classifier with feature selection method are greater than RF classifier with origin attributes that are shown in Figure 4.20.

Figure 4.21 demonstrates evaluation metric values of four classifiers: LR with and without attribute selection method and RF with and without attributes selection

method. Accuracy and f-score results of LR classifiers are less than 63 percent. Accuracies of RF with the attribute selection method and without the attribute selection method are 0.952 percent and 0.939 percent respectively. F scores of RF with the attribute selection method and without the attribute selection method are 0.946 percent and 0.935 percent, respectively. Therefore, RF classifier with the attribute selection method is more suitable for the proposed car insurance claim prediction system based on 42 attributes.

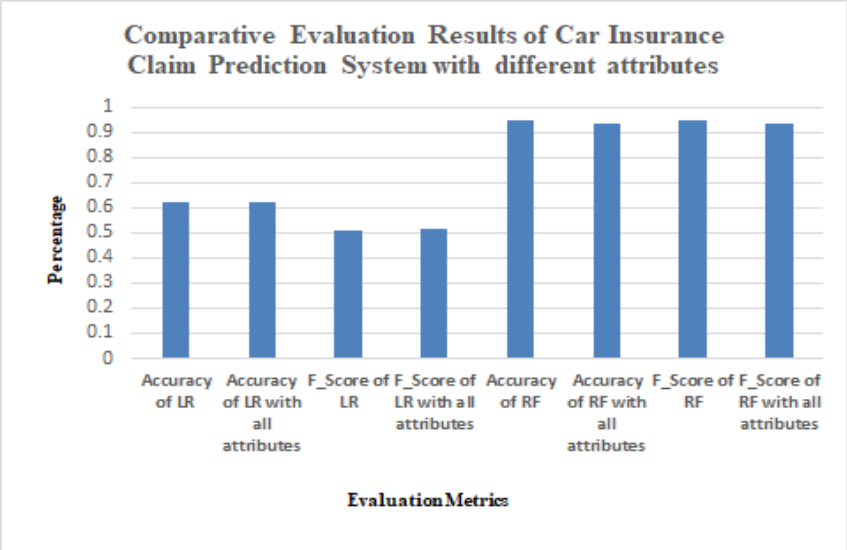

Figure 4.22 Comparison of Accuracy and F-Score Values for LR and RF Classifiers

According to the average experimental results from comparing accuracy and f score values of car insurance claim prediction system with different attributes, the classifiers based on 38 attributes and 40 attributes are the best classifiers with the same accuracy (0.97 percent) and the same f score (0.97 percent) that are shown in Figure 4.15. The second best classifier based on 42 attributes are accuracy with 0.95 percent and 0.95 percent that are shown in Figure 4.22.

4. 3 Chapter Summary

This chapter explains the proposed system with GUI and its performance evaluations with accuracy and f score. The evaluated results of the system can be analyzed to consider which classifier is more suitable for the proposed system.

CHAPTER 5
CONCLUSION AND FURTHER EXTENSION

This system has proposed four classifiers: Linear Regression (LR) based on Variance Threshold Selector with selected attributes, LR classifier with all attributes, Random Forest (RF) classifier based on Variance Threshold Selector with selected attributes and RF classifier with all attributes. To create four classifiers, the system has divided the dataset into training dataset with 80% and testing dataset with 20% randomly. For the two classifiers with all attributes, the training dataset is used to create the LR classifier and RF classifier. For two classifiers with the feature selection method, the new training dataset and testing dataset by removing low variance value of attributes using Variance Threshold Selector method. After that, two LR classifier and RF classifier are been created by using new datasets. The system has analyzed the difference attributes: 30, 32, 34, 36, 38, 40 and 42 to choose the number of attributes and important attributes. The system has tested 10 times for each attribute number because of splitting training and testing datasets randomly. Finally, the system compares the evaluation results with metrics: accuracy and f score. LR classifiers with and without the feature selection method are not suitable for the car insurance claim prediction system because the dataset is not linearly separable the two classes of data from each other and their accuracies are not greater than 0.65 percent. For RF classifiers with and without the feature selection method are suitable for the proposed system with accuracy 93 percent and f score 92 percent. Moreover, the classifiers based on 38 attributes and 40 attributes are the best classifiers with the same accuracy (0.97 percent) and the same f score (0.97 percent) while the second best classifier based on 42 attributes are accuracy with 0.95 percent and 0.95 percent.

5.1 Advantages and disadvantages of the System and Further Extension

The proposed system can predict whether the car claim care insurance within the next six months for car insurance companies and car owners by combining Random Forest Classifier and Variance Threshold Selector with above 90% of accuracy and F score. The proposed system can prove more accuracy of Random Forest Classifier using

Variance Threshold Selector than Random Forest Classifier without using Variance Threshold Selector. Applying the proposed car insurance claim prediction system on the apache spare platform, the size of training and testing dataset can increase in the future. This proposed system can be applied on the apache spark platform. Moreover, the selecting important attributes can support for the proposed system. The current work of this study implemented Logistic Regression and Random Forest with the feature selection method.

Although the evaluation results of two RF classifiers are more than 92 percent, the evaluation results of LR classifier are less than 63% because of the data set nature. The choice of LR classifier is one disadvantages of the proposed system. It is needed to consider some more approach that is suitable for the dataset natures by considering accuracy. Furthermore, the choice of input attributes makes the model more challengeable topic. Moreover, the performance of this system can be evaluated by using various evaluation methods. In the future, this system can be used to test a large amount of training data and testing data.

REFERENCES

[1] A. C. Tan and D. Gilbert, “An empirical comparison of supervised machine learning techniques in bioinformatics,” First Asia Pacific Bioinforma. Conf. (APBC 2003), vol. 19, no. Apbc, 2003.

[2] A. C. Yeo, K. A. Smith, R. J. Willis, and M. Brooks, “Clustering Technique for Risk Classification and Prediction of Claim Costs in the Automobile Insurance Industry,” Int. J. Intell. Syst. Accounting, Financ. Manag., no. November 1999, pp. 39–50, 2001.

[3] A. L. Heureux and M. Grolinger, Katarina and Caprtz, “Machine Learning With Big Data : Challenges and Approaches,” IEEE Access, vol. 5, pp. 7776–7797, 2017.

[4] A. S. Alshamsi and A. Ain, “Predicting Car Insurance Policies Using Random Forest,” IEE, pp. 128–132, 2014.

[5] A. S. Alshamsi and A. Ain, “Predicting Car Insurance Policies Using Random Forest,” pp. 128–132, 2014.

[6] Baran, Sebastian, and Przemysław Rola. "Prediction of motor insurance claims occurrence as an imbalanced machine learning problem.", Cornell University, arXiv preprint arXiv:2204. 06109 (2022).

[7] C. Using, S. Vector, F. K. C-means, Z. Rustam, and F. Yaurita, “Support Vector Machines for Classifying Policyholders Satisfactorily in Automobile Insurance .,” J. Phys. Conf. Ser. Pap., 2018.

[8] Endalew Alamir, Teklu Urgessa, T. GopiKrishna and Ellappan V, “Application of Machine Learning with Big Data Analytics in the Insurance,” vol. 11, no. 12, pp. 1064–1073, 2020.

[9] E. Shaikh, I. Mohiuddin, Y. Alufaisan and I. Nahvi, "Apache Spark: A Big Data Processing Engine," 2019 2nd IEEE Middle East and North Africa COMMunications Conference (MENACOMM), Manama, Bahrain, 2019, pp. 1-6, doi: 10.1109/MENACOMM46666.2019.8988541.

[10] Hailu Zeleke ., “Insurance in Ethiopia: Historical Development, Present Status and Future Challenges .,” vol. 1, no. 1, p. 308, 2009.

[11] J. Brownlee, Machine Learning Mastery with python, V1.4. 2016.

[12] K. A. Smith, R. J. Willis, M. Brooks, K. A. Smith, R. J. Willis, and M. Brooks, “An analysis of customer retention and insurance claim patterns using data mining : a case study,” J. Oper. Res. Soc. ISSN, no. 5682, pp. 1476–9360, 2017.

[13] K. P. M. L. P. W. and M. C. W. Depa, “A Comparative Study of Data Mining Algorithms in the Prediction of Auto Insurance Claims,” Eur. Int. J. Sci. Technol., vol. 5, no. 1, pp. 47–54, 2016.

[14] L. Wang and C. A. Alexander, “Big Data : Infrastructure , technology progress and challenges,” J. Data Manaagement Comput. Sci. Vol., vol. 2, no. 1, pp. 1–6, 2015.

[15] M. C. Wijegunasekara and Weerasingheand M.C. Wijegunasekara , “A Comparative Study of Data Mining Algorithms in the Prediction of Auto Insurance Claims,” vol. 5, no. 1, pp. 47–54, 2016.

[16] M. Zaharia, M. Chowdhury, T. Das, A. Dave, J. Ma, M. McCauley, M. J. Franklin, S. Shenker, I. Stoica, “Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing”, in: 9th USENIX Conference on Networked Systems Design and Implementation, NSDI’12, USENIX Association, Berkeley, USA, 2012, pp. 2–2.

[17] M. Zaharia, R. S. Xin, P. Wendell, T. Das, M. Armbrust, A. Dave,X. Meng, J. Rosen, S. Venkataraman, M. J. Franklin, A. Ghodsi,J. Gonzalez, S. Shenker, I. Stoica, Apache spark: A unified enginefor big data processing, Commun. ACM 59 (11) (2016) 56–65.

[18] N. K. Frempong, N. Nicholas, and M. A. Boateng, “Decision Tree as a Predictive Modeling Tool for Auto Insurance Claims,” Int. J. Stat. Appl., vol. 7, pp. 111–120, 2017.

[19] P. Bharal and A. Halfon, “Making Sense of Big Data in Insurance,” ACORD and MarkLogic, 2013.

[20] R. J. Kate and A. M. Swartz, “Assessment of various supervised learning algorithms using different performance metrics Assessment of various supervised learning algorithms using different performance metrics,” IOP Conf. Ser. Mater. Sci. Eng., 2017.

[21] T. Kavipriya and N. Kumar, “A Study on Machine Learning Algorithms for Big Data Analytics,” IOSR J. Eng., no. Iccids, pp. 40–46, 2018.

[22] W. Lin, Z. Wu, L. Lin, A. Wen, and J. I. N. Li, “An Ensemble Random Forest Algorithm for Insurance Big Data Analysis,” IEEE Acess, vol. 5, 2017.

[23] X. Meng, J. Bradley, B. Yavuz, E. Sparks, S. Venkataraman, D. Liu,J. Freeman, D. Tsai, M. Amde, S. Owen, et al., Mllib: machine learningin apache spark, The Journal of Machine Learning Research 17 (1)(2016) 1235–1241.

[24] Car Insurance Claim Prediction Dataset From Kaggle:https://www.kaggle .com /datasets/ifteshanajnin/ carinsuranceclaimprediction-classification.

[25] Understanding Logistic Regression – GeeksforGeeks:https://www.geeksforgeeks.org/under standing- logistic-regression.

[26] “Documentation | Apache Spark”, https://spark.apache.org/documentation.html

[27] “How to Use Variance Thresholding For Robust Feature Selection”, https://towardsdatascience.com/how-to-use-variance-thresholding-for-robust-feature-selection-a4503f2b5c3f

AUTHOR'S PUBLICATION

[1] Thein Than Ko and Tin Zar Thaw, "The Car Insurance Claim Prediction System by Using Machine Learning Algorithms on Apache Spark Platform", the Proceedings of the Conference on Parallel and Soft Computing (PSC 2023), University of Computer Studies, Yangon, Myanmar, 2023.