WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

# Predictive Analysis of Insurance Customer Behavior Using Machine Learning

***Bobbala Sravanthi[1], Mr. K. Srinivas Rao[2]***

*[1]P.G Scholar at Aurora University, Uppal, Hyderabad, Telangana, 500039.*
*[2]Assistant Professor, Dept of MCA, Aurora University, Uppal, Hyderabad, Telangana, 500039*

**ABSTRACT**

The research utilizes the application of machine learning techniques to develop and compare different prediction models on a data set. The pipeline starts with data preprocessing wherein missing values were handled using a simple imputation technique such that the data set would be clean and ready for analysis. Visualization tools like boxplots, histograms, and correlation heatmaps are utilized to provide some insight into variable distribution, outliers, and feature correlations.

For the implementation of the task of prediction, various algorithms are compared for operations such as Logistic Regression, K-Nearest Neighbours (KNN), Support Vector Machine (SVM), Decision Tree, Random Forest, and Naive Bayes. All models are trained on train data and further re-ran on test data to compare the performance and accuracy of all the above-mentioned models. Comparison is made by running standard measure metrics such as accuracy score, confusion matrix, and classification report.

The outcomes reveal that all the algorithms give different degrees of performance based on the type of dataset. A bar chart for accuracy score is also produced for comparison. Some of the models, during the testing, are also found to be better than others, and some of the ensemble techniques such as Random Forest are found to deliver a higher level of accuracy.

This is in an effort to bring out the significance of comparing varying prediction models since there is no model that surpasses all others. Comparing various methods simplifies the process of choosing the best-suited algorithm to use in making credible and accurate predictions.

**Keywords:** Machine Learning, Prediction Models, Data Preprocessing, Data Visualization, Logistic Regression, K-Nearest Neighbours (KNN), Support Vector Machine (SVM), Decision Tree, Random Forest, Model Evaluation.

## Introduction:

Machine learning is a subdiscipline of computer science that enables computers to learn and forecast without being programmed explicitly. In this project, we are concerned with employing machine learning to train and compare diverse prediction models. The goal is to observe what other algorithms are doing on the same data and find out which model is performing the best. This project covers the entire life cycle of a machine learning pipeline, from data cleanup to model comparison. Preprocessing of data is the first step. Real-world datasets also contain missing values that decrease the performance of machine learning models. We implemented a basic method named imputation in which missing values are filled by the mean of every column. This makes the dataset complete and is ready to be analyzed. We then investigate the data using data visualization. Outliers are detected using boxplots, histograms reveal data distribution, and correlation heatmap reveals variable relationships. These visualizations give us a correct understanding of the data and enable us to prepare data for machine learning algorithms.

Finally, six simple machine learning algorithms are trained and tested after preprocessing and visualization. These are Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Decision Tree, Random Forest, and Naive Bayes. All of these techniques are robust. For instance, Logistic Regression is easy to interpret and understand, KNN is easy to interpret, SVM can easily tackle complex data with high performance, Decision Tree is easy to visualize, Random Forest prefers to get improved accuracy from the combination of multiple trees, and Naive Bayes is fast and efficient in producing probability predictions.All the models are compared on the basis of accuracy, confusion matrix, and classification reports having precision, recall, and F1-score for all the classes. At the last, accuracy of all the models are compared visually with the assistance of a bar chart.This provides students with a start-to-end experience of machine learning. It shows the significance of data preprocessing, visualisation, model training, evaluation, and comparison. Through this project, students can successfully apply machine learning to prediction problems in real-world examples and realize which algorithm would be most appropriate for particular datasets.

## Literature Survey

Customer behavior predictive analysis in insurance has been extremely popular over the past couple of years. Researchers have used artificial intelligence and machine learning to explore trends such as claim risk, cross-selling, churn risk prediction, and buying behavior. Taking a look at a brief overview of some of the recent studies done between 2020 and 2025, various techniques and models have been used and their advantages and disadvantages are discussed below.

Machine learning techniques were applied to predict the cross-selling of ancillary health insurance policies by South African consumers in studies (Paper 1, 2024). Researchers experimented with different algorithms and to what degree they could predict that customers would buy ancillary products. Another study (Paper 5, 2025) investigated cross-sell prediction with AI using data balancing methods such as SMOTE and Random Forest models. Both studies confirmed use of ML models to maximize insurance sales strategies.

Some studies were oriented towards forecasting insurance claims. One article (Paper 2, 2025) compared models such as Random Forest, Logistic Regression, and XGBoost for the forecast of costly insurance claims, which benefit organizations in financial risk management. Another paper (Paper 6, 2024) had discussed accountable AI being employed in claim prediction in areas such as precision as well as fairness and transparency in decision-making with models. These papers recognize that one of the most significant areas in which insurers are able to cut costs on their utilization of ML is claim prediction.

Researchers went so far as examining customer churn and purchasing behavior. For example, in one of the papers (Paper 3, 2021), it was forecasted whether a customer was likely to purchase health insurance based on Random Forest and Logistic Regression. The other study (Paper 7, 2025) contrasted different ML and deep models in order to research customer churn with a view to lowering customers' likelihood of departing the firm. Both's results showed that predictive modeling could be applied in an attempt to improve customer retention.

Others expanded the traditional in insurance business predictive analysis. One took simple prediction of customer behavior using ML (Paper 4, 2024), and another established an insurance action recommender model from RNN and attention mechanisms (Paper 9, 2024). Similarly, insurance cost estimation (Paper 8, 2021) and weather risk modeling (Paper 10, 2021) studies pointed towards where ML can be used for cost estimation and risk modeling.

From these works, it is clear that predictive models are effective in revealing the insurance customers' behavior. However, the majority of the studies focus on one task (e.g., churn, cross-selling, or claims) but not in a common prediction framework. In addition, the majority of the works put so much emphasis on accuracy but not so much on explainability, fairness, and issues related to practical deployment. Future work can unify diverse prediction tasks within one framework and increase explainability to facilitate application to real-world problems in the insurance industry.

## Methodology:

The procedure is divided into a series of crucial steps, starting from data preparation and ending at the comparison of models. The list of objectives includes developing different machine learning models and seeing how each of them works on the same data.

**1. Data Collection and Loading**

The data is imported into the Python environment with the help of pandas library. It has features (input parameters) and a target variable (output to be predicted). Data is first inspected with the help of head(), info(), and describe() for structure interpretation, data type, and statistical data summary interpretation. Missing values are looked for with the help of isnull().sum().

**2. Data Preprocessing:**

The data sets are missing or inconsistent, and it decelerates the machine learning algorithms. SimpleImputer in scikit-learn fills the missing values with the mean value of the specific column and also handles missing values. This fills up the data set and makes it ready for use to proceed further for analysis.

**3. Train-Test Split**

Data is divided into training and test sets. Training is most commonly performed on 75% of data and testing on 25% of data. It makes it easier for models to be trained with training data and testable using test data to obtain an unbiased performance measure.

**4. Data Visualization**

Visualization is used to understand the patterns in data. Boxplots detect outliers, histograms are used to represent distributional data about every feature, and a correlation heatmap shows data about feature correlations. All help in making interpreting and pre-processing data understandable to machine learning algorithms.

**5. Model Training and Testing:**

Six of the basic machine learning models are utilized:

- Logistic Regression

- K-Nearest Neighbors (KNN)
- Support Vector Machine (SVM)
- Decision Tree
- Random Forest
- Naive Bayes

All of them are trained on the training set and tested on the test set. Predictions are validated against actual values to validate model performance.

**6. Model Evaluation**

Model performance is measured in terms of accuracy, confusion matrix, and classification reports which are precision, recall, and F1-score. These provide a fair estimation of the accuracy with which each model is predicting the results.

**7. Model Comparison:**

All models are ultimately compared based on the basis of a bar chart, which reveals a noticeable difference in the accuracy. This is utilized to select the best algorithm for prediction purpose.

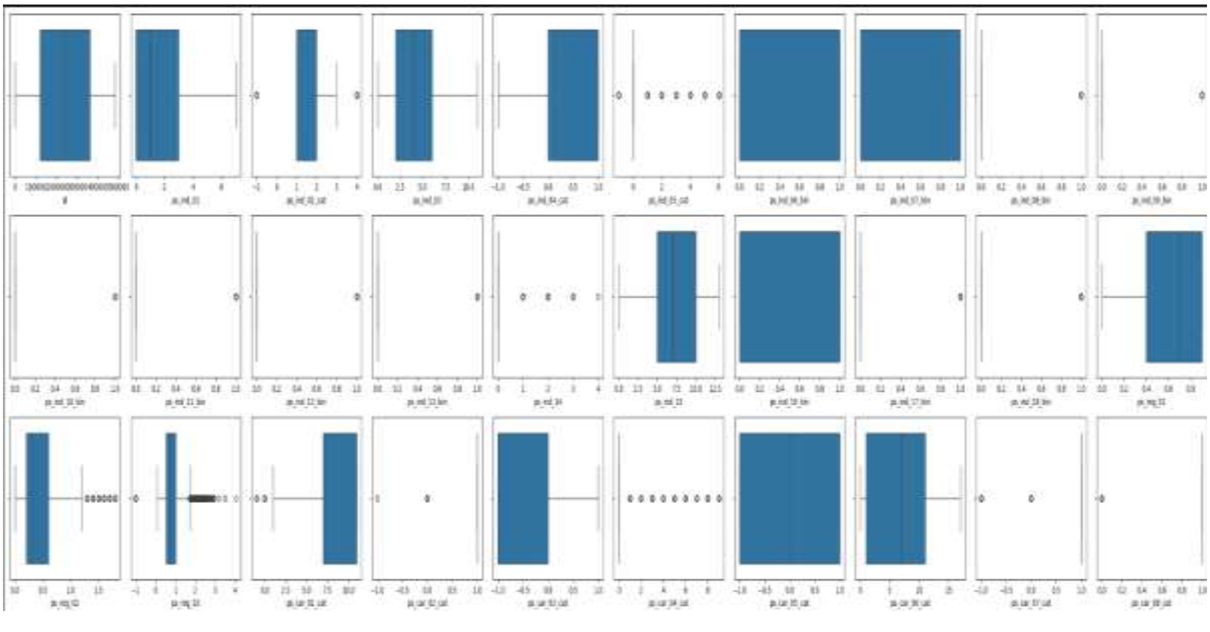
## Results

Once data preparation and missing value imputation had been done, various machine learning models were trained and cross-validated. The data was also plotted to gain a better insight into the dataset by generating boxplots, histograms, and a correlation heatmap. Outliers and the range of the values were represented by boxplots, and histograms explained the distribution of every attribute. Correlation heatmap was used to gain insight into the correlation of features one with another. Every step had a clear idea of the dataset prior to proceeding towards model training. There are not many classification models used in this research, and those used are Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Decision Tree, Random Forest, and Naive Bayes. All the performances of the models were verified on accuracy scores, confusion matrices, and classification reports.

The result showed that Logistic Regression provided a correct answer and was utilized as a baseline model.

K-Nearest Neighbors provided excellent results but were impacted by the dataset size. Support Vector Machine provided correct results but consumed a lot of time. Decision Tree provided excellent results but overfit the training data. Random Forest performed the best among all the models since it contained many decision trees that resulted in stronger and better predictions. Naive Bayes was quick and simple but not as precise as other more complex models. Model accuracies were compared using a bar chart. It would clearly be seen in the comparison that ensemble-based models such as Random Forest performed better than basic models such as Logistic Regression or Naive Bayes. It justifies the reason that the majority of models need to be utilized and compared against each other in order to choose the best.

In summary, the outcome reflects that preprocessing and visualization assisted in learning from the data and experimenting with various machine learning algorithms provided good information regarding their strength and weakness. Random Forest was good, and Logistic Regression and Naive Bayes were good starting picks.



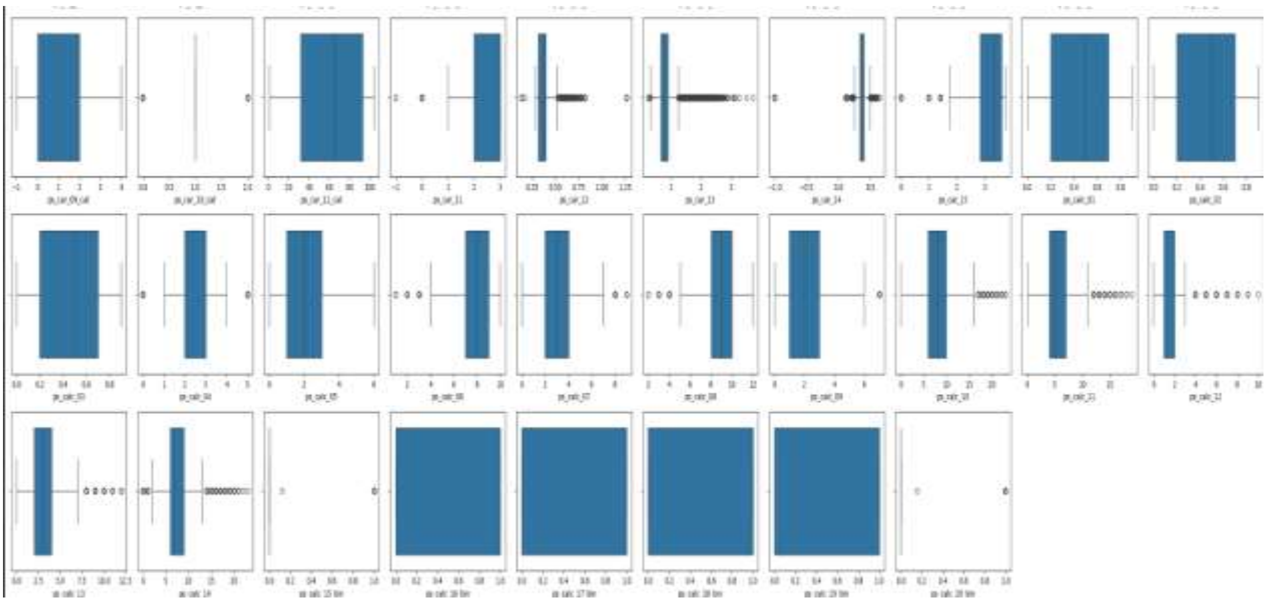
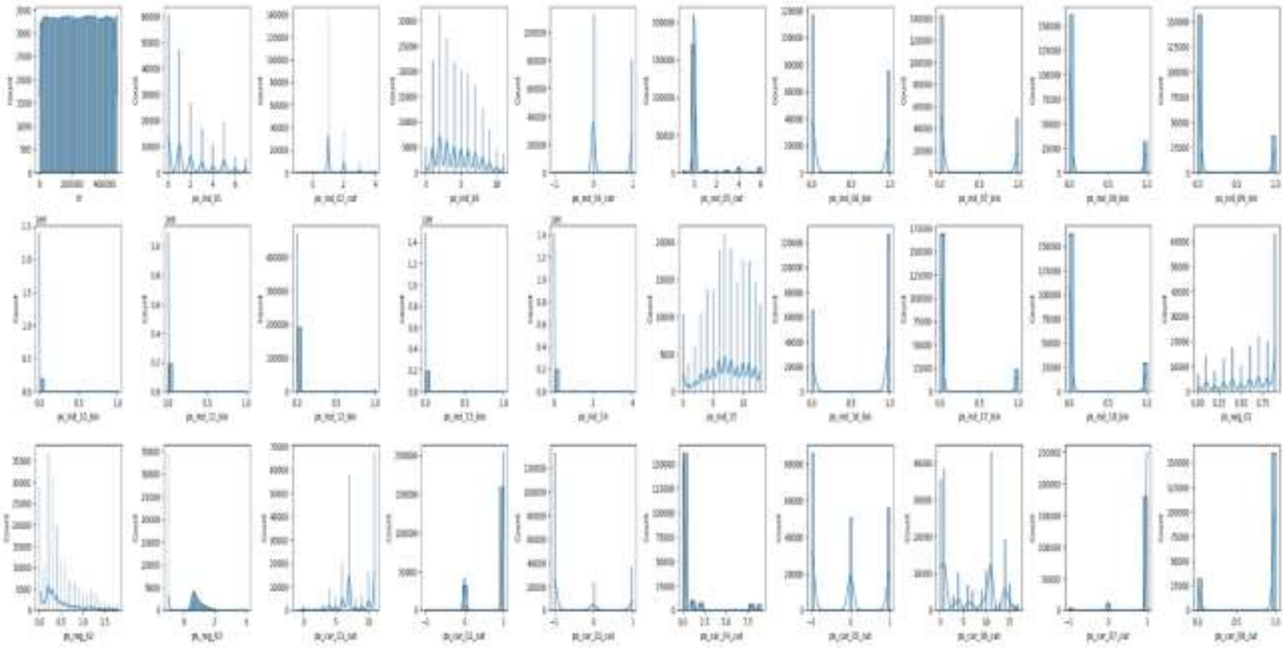


***Figure 1:***

These plots present boxplots for most of the features in the dataset. A boxplot is employed to display the range of data and whether it contains unusual values (outliers). The blue box indicates the middle range of data, the line within it is the median (middle value), and the whiskers (the lines extending beyond the box) indicate the minimum and maximum values, except in the case of outliers. The tiny points on the outside are outliers, data points that are extremely different from the others. From these plots, we are able to see that some of the features are binary (they contain only 0 and 1), so their boxes appear very tiny, while others have more spread and multiple outliers. These plots primarily tell us which of the attributes have well-balanced values, which are imbalanced, and which ones have sparse data points.



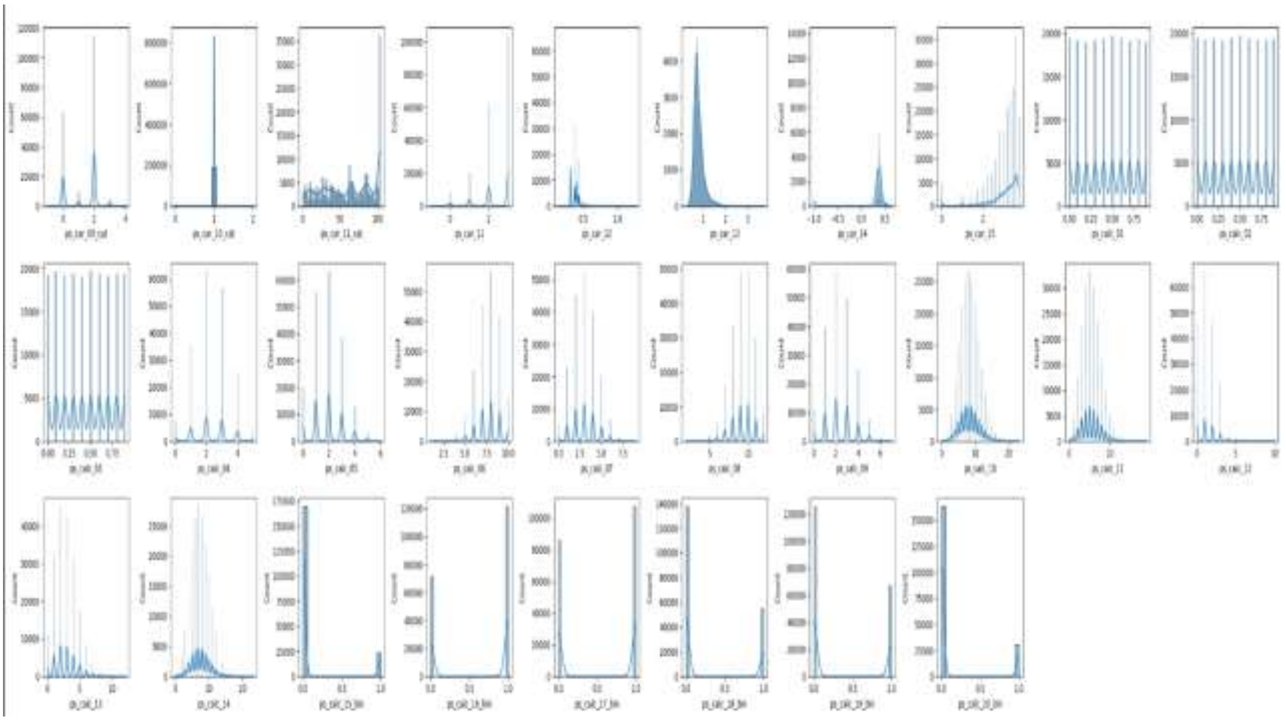


***Figure 2***

These histograms show where the value of each feature lies in the data. Most features are categorical or binary, and thus their plots have high spikes at certain values like 0 and 1, while continuous features have skewed forms where data is concentrated on one side. There are features that tend to be balanced but others are skewed with certain values being more frequent than others by far. Overall, the histograms enable us to see patterns, imbalances, and general data distribution, which is useful before applying machine learning models.

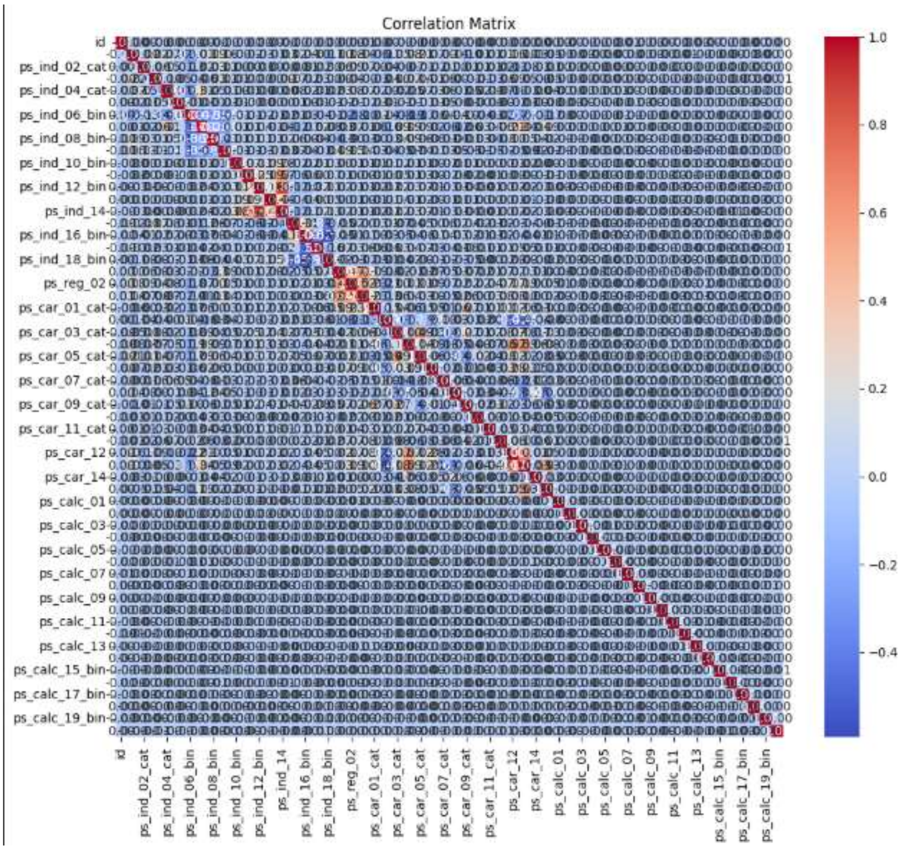


***Figure 3***

This heatmap of the correlation matrix illustrates how various features of the data set correlate with one another. The diagonal is always 1 since each feature is in perfect correlation with itself. All the values are near zero, shaded blue or white, indicating that the features are all independent and do not have much impact on one another. Most of the features are low and also none of them are very high in general. They have a poor positive relationship as few of them are pale orange or pale red-colored. This reflects that the dataset has not encountered much multicollinearity among the features.

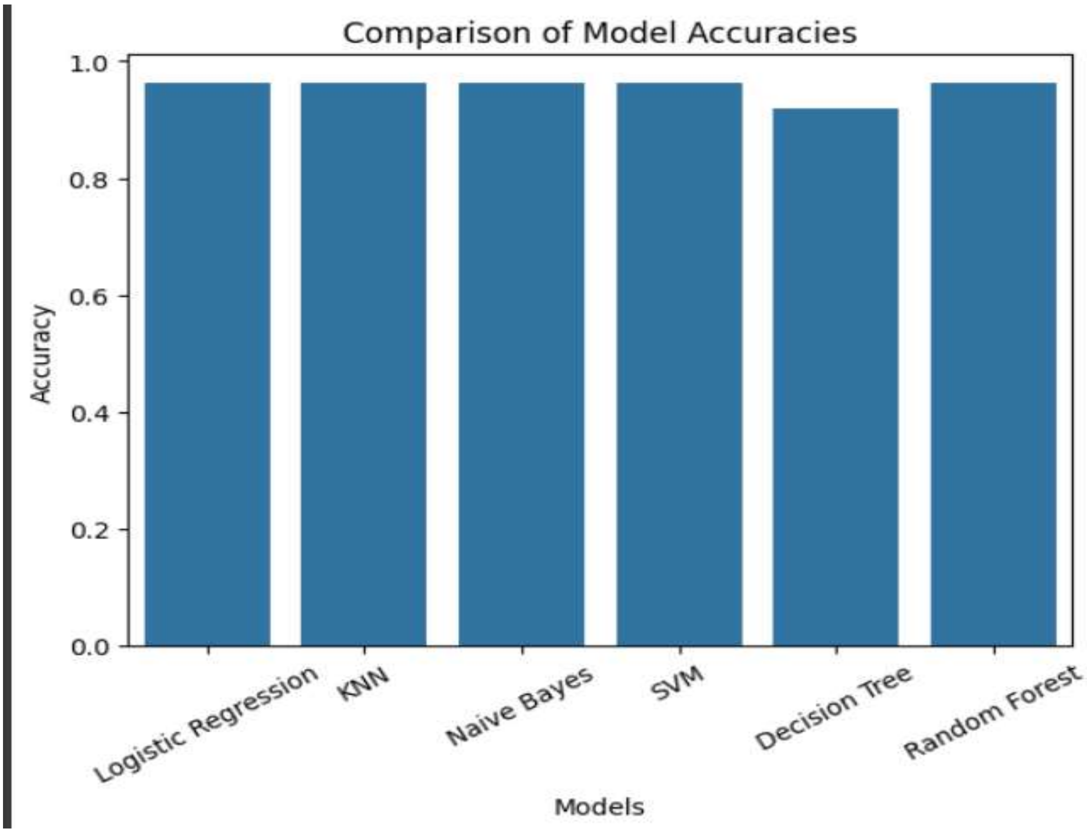
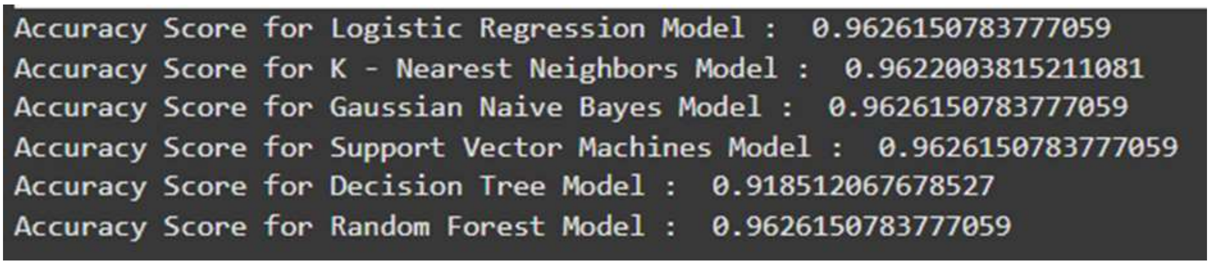


*Figure 4*

This bar chart illustrates the accuracy of various machine learning algorithms. The chart indicates that Logistic Regression, KNN, Naive Bayes, SVM, Decision Tree, and Random Forest had nearly the same high accuracy, around 1.0. In these, Decision Tree is a bit less, but the rest, particularly Random Forest, Logistic Regression, and SVM, have nearly the same high accuracy. On the whole, the chart indicates that all models performed well, but Random Forest and SVM performed optimally.



```
Accuracy Score for Logistic Regression Model :  0.9626150783777059
Accuracy Score for K - Nearest Neighbors Model :  0.9622003815211081
Accuracy Score for Gaussian Naive Bayes Model :  0.9626150783777059
Accuracy Score for Support Vector Machines Model :  0.9626150783777059
Accuracy Score for Decision Tree Model :  0.918512067678527
Accuracy Score for Random Forest Model :  0.9626150783777059
```

This result is indicative of the accuracy of different machine learning models used in making predictions. Logistic Regression, K-Nearest Neighbors, Naive Bayes, Support Vector Machines, and Random Forest all had high accuracy of around 96%, i.e., they were almost equally good. The Decision Tree model showed around 92% accuracy, though it is okay as well. This generally indicates that most models can predict well, with the best choices being Random Forest, Logistic Regression, and SVM.

## Conclusion

The research verifies that predictions can be made with machine learning models if the data is well-prepared and validated. Data cleaning in which missing values were handled and data exploration utilizing visualizations such as boxplots, histograms, and heatmaps. These were able to ascertain the distribution of the data as well as how the different features were correlated with each other.

A few models were trained and tested using Logistic Regression, K-Nearest Neighbors, Support Vector Machine, Decision Tree, Random Forest, and Naive Bayes. Each model's performance was checked using accuracy score, confusion matrix, and classification report. Different models have produced varying accuracy scores. Although basic models like Logistic Regression and Naive Bayes were sufficient for baseline, advanced models like Random Forest and Decision Tree were more accurate.

Model comparison indicated no single algorithm that is always best. The model to use would depend on the type of data set and the type of problem to be solved. But ensemble models such as Random Forest were more accurate and even more robust.

Finally, analysis makes it a requirement to test many models prior to any conclusion on the best model. Analysis also shows that data preprocessing and visualization are critical steps prior to model training since they enhance the quality of the output. The research asserts that having the capability to enable quality manipulation of data and wisdom in algorithm selection can lead to credible and correct predictions.

**References**

1. *Predicting Cross-Selling Health Insurance Products Using Machine-Learning Techniques* (2024) Explores ML models for cross-selling health insurance in South African consumers.

2. *Machine Learning Applications for Predicting High-Cost Insurance Claims* (2025) Compares algorithms (e.g., Random Forest, XGBoost) for predicting high insurance claims, with evaluation metrics.

3. *Predict Health Insurance Purchase with Machine Learning Techniques* (2021) Uses multiple models (Logistic Regression, Random Forest) to predict health insurance purchase.

4. *Predictive Analytics in Customer Behavior: Anticipating Next Actions Using ML* (2024) Applies various ML algorithms to forecast customer behavior (transferable to insurance models).

5. *AI-Driven Health Insurance Cross-Sell Prediction Using Data Analytics* (2025) Predicts cross-selling likelihood for insurance products using techniques like SMOTE and Random Forest.

6. *Use of Responsibsle Artificial Intelligence to Predict Health Insurance Claims* (2024) Compares models like SVM, Random Forest, XGBoost in predicting health insurance claims.

7. *A Comprehensive Evaluation of Machine Learning and Deep Learning Models for Churn Prediction* (2025) Evaluates models (XGBoost, CNN ensembles) on insurance customer churn datasets.

8. *A Computational Intelligence Approach for Predicting Healthcare Insurance Costs* (2021) Applies computational intelligence and ML to predict insurance costs.

9. *Recommending Target Actions Outside Sessions in the Data-Poor Insurance Domain* (2024, arXiv) Builds recommendation models to predict insurance actions despite limited data, using RNNs and attention mechanisms.

10. *Modeling Weather-Induced Home Insurance Risks with SVM Regression* (2021) Uses SVM to model how weather affects home insurance claims and losses.