Innovative Applications of Machine Learning Classifiers for Medical Cost and Prediction

Arpit Sharma
Department of ECE G.B.Pant DSEU
Okhla Campus-1
New Delhi, India

Kartikey Sharma
Department of CSE G.B.Pant DSEU
Okhla Campus-1
New Delhi, India

Avanish Pratap Singh
Department of ECE G.B.Pant DSEU
Okhla Campus-1
New Delhi, India

Prerna Bhardwaj
Department of CSE Indira Gandhi
Delhi Technical
University for women
New Delhi, India

Akhtar Warsi
Department of CSE G.B.Pant DSEU
Okhla Campus-1
New Delhi, India

***Abstract*— The surging global healthcare costs have sparked a rise in interest for accurate and dependable methods to forecast medical expenses. Precise cost predictions can benefit healthcare providers, insurers, and policymakers in effectively allocating resources and enhancing the efficiency of healthcare services.**

To address this issue, this study delves into the use of machine learning classifiers for medical cost forecasting. Multiple models such as linear regression, polynomial classifier, random forest classifier, and decision tree classifier are examined. Moreover, feature importance analysis is conducted to identify the key variables that impact cost prediction.

With a diverse dataset encompassing patient demographics, clinical data, and healthcare utilization variables, various machine learning classifiers are employed, including linear, polynomial, decision trees, and random forests, to construct predictive models. In addition, feature engineering methods are applied to extract valuable insights from the data, and hyperparameter tuning is utilized to optimize model performance.

In conclusion, this research adds to the growing body of knowledge on healthcare and cost prediction by implementing machine learning classifiers. The developed models can assist healthcare stakeholders in making informed decisions regarding resource allocation, cost control, and patient care. The application of these predictive models has the potential to improve the overall effectiveness and sustainability of healthcare systems, ultimately resulting in better patient outcomes and cost-efficient delivery of healthcare services.

Keywords—*Medical cost prediction, Linear Regression, Random Forest classifier, polynomial classifier, Decision tree classifier, hyper parameters, optimization.*

I. INTRODUCTION

Medical costs are one of the most common and expensive expense in a person's life. Recent research shows a tremendous growth in medical expenses. It is due to the fact that many viruses have come in our world which are gradually causing great harm to the human body. Apart from this, there are some diseases which are found in every common person which gradually harm their body like diabetes, Blood Pressure and allergy. The medical cost for these diseases also becomes expensive if they remain for a long period of time. But now as people are becoming aware about their health and expenditures on their health.

In order to calculate the correct medical expense so that we can save money and time we have used machine learning algorithms. Machine learning is the scientific study of algorithms and statistical models that computer systems use to perform a specific task without being explicitly programmed. [1] Machine learning is used to teach machines how to handle the data more efficiently.[1] Basically, we have taken dataset from Kaggle which contains data of people like their age, sex, body mass index, children, smoker, region. Then we have applied machine learning algorithms to find the medical cost prediction. The algorithms used are Linear regression (It is the easiest machine learning algorithm. It is basically used for predictive analysis. This algorithm shows a linear relationship between a dependent and one or more independent variables.), Ridge regression (this regression is basically used to shrink the coefficients towards zero.), Lasso regression (this regression is basically used to shrink the coefficients exactly zero.), Decision Tree (It is basically a graphical representation which shows all the possible solutions to a problem based on some given conditions. There are two nodes in a decision tree that is Decision Node and Leaf Node. The decision nodes are responsible for making decisions whereas leaf nodes are the output of those decisions.), Random Forest (This algorithm says that we don't need to depend on a single decision tree instead of that the random forest takes the prediction from each tree and based on those predictions it predicts the final output.), K-Nearest Neighbor (It is the simplest machine learning algorithm in this we basically classify a new data point based on the similarity.) The findings of our study demonstrate the effectiveness of machine learning classifiers in predicting medical costs. We evaluate model performance using metrics such as mean absolute error, root mean squared error, and R-squared. Our results indicate that machine learning classifiers outperform traditional statistical methods, offering higher accuracy and improved predictive capabilities.

Furthermore, we investigate the impact of different features on cost prediction, highlighting the significance of variables such as age, chronic conditions, and healthcare utilization patterns. Insights from feature importance analysis can inform healthcare providers in identifying high-risk patients and implementing targeted interventions.

Clearly, we can say that the use of algorithms in biological research has altered dramatically as a result of the advent of computational approaches [2]. Machine learning and deep learning are examples of computational technologies that have been successful in uncovering biological mechanisms [2]. The uprising computational intelligence trend like machine learning with NGS data gives a new direction to disease identification and analysis [2].

II. LITERATURE REVIEW

This section introduces various predictive studies by researchers on medical cost prediction using machine learning.

Lorenzo Famiglini [3] proposed three models for the prediction of ICU admission for COVID-19 patients. One model, which is based on the ensembling of 3 models, has been selected for its high accuracy, despite its low clinical interpretability because of the black-box nature [3]. The other two models, i.e. Decision Tree is a nonparametric supervised learning method that is used for classification and regression [4]. It implements a simple set of rules to create partitions of the generated data and iterates the partitioning process to produce predictions [4]. Decision Tree can classify data without complicated calculations and can be used for both categorical and classification variables [4]. It is generally suitable for predicting categorical outcomes [4].), and a logistic regression is an efficient and straightforward method for binary or multiple classification problems [4]. It uses the logit or natural log of the odds so that the probability of the data belonging to a particular class is not excluded from the [0, 1] range [4]. Logistic Regression is a supervised learning algorithm that categorizes classes according to probability and provides accurate predictions [4]., have been selected because of their explain ability [13]. The results reported was AUC of .81 (the decision tree) and AUC of .83 for (logistic regression) and AUC of .88 for the black-box model (an ensemble). As, a result they reported a retrospective study to address the challenging task of predicting whether a COVID-19 patient will have to be transferred to the ICU within the next 5 days during their hospital stay [3]. The proposed approach, based on both interpretable and black-box models, reported good results [3].

MD Sadique Hasan [5] proposed model for Rapid Bacterial Detection and Identification of Bacterial Strains using Machine learning. The algorithms used were: The unsupervised algorithm PCA was used for the lower-dimensional representation of the whole data set and top feature extraction. SVM, KNN, DT, NB, Ensemble, LDA, Linear, RNN, and NN were used for the binary classification of bacterial and non-bacterial samples or negative control initially [24]. Their results show that around 97.9% accuracy can be achieved for bacterial contamination detection for as low as 1 CFU/mL while 92.1% accuracy can be achieved for differentiating the gram-positive and gram-negative strains [24].

Eunbi kim [4] proposed five predictive models for the analysis on benefits and costs of machine learning based on early hospitalization prediction. Most studies related to hospitalization prediction have used machine learning algorithms[4]. In this study, we also use machine learning algorithms to classify ED patient hospitalization[4]. So the algorithms used were Logistic Regression is an efficient and straightforward method for binary or multiple classification problems[4]. It uses the logit or natural log of the odds so that the probability of the data belonging to a particular class is not excluded from the [0, 1] range. LR is a supervised learning algorithm that categorizes classes according to probability and provides accurate predictions [4]., XGBoost is a highly scalable algorithm developed to improve performance and computational speed[4]. Boosting is an ensemble technique that adds new models to accommodate for errors made by existing models[4]. Gradient boosting is used to create new predictive models using the residuals of fitted models and minimize losses[4]. XGBoost can be used for both regression and classification [4]., Support Vector Machine(SVM) is a linear learning method and classification method in supervised learning that finds the optimal hyperplane that separates two classes. It maximizes the distance between the two closest classes to achieve a high classification performance [4]. The data points for the two classes closest to the determined decision boundary are called the support vectors. The distance between the support vector and decision boundary is called the margin, and the decision boundary that maximizes the margin is optimal [4]., Decision Tree is a nonparametric supervised learning method that is used for classification and regression [14]. It implements a simple set of rules to create partitions of the generated data and iterates the partitioning process to produce predictions [4]. Decision Tree can classify data without complicated calculations and can be used for both categorical and classification variables [4]. It is generally suitable for predicting categorical outcomes [4].), Natural Gradient Boosting is a supervised learning algorithm with stochastic prediction capabilities[4]. It estimates the parameters of the conditional probability distribution P(y|x) as a function of x by boosting[4]. N G Boost outputs the overall probability distribution for predictions using natural gradients [15]. As, a result they concluded that XG Boost is the best predictive model because X G Boost has the second-highest specificity of 0.9582 (95% CI 0.58–0.98). X G Boost has the highest AUC of 0.9332 (95% CI 0.92–0.94). A large portion of the economy is devoted to paying for health care. Spending on healthcare accounts for around 30% of the GDP. In terms of both absolute spending and as a percentage of the economy, health spending in developed countries is the greatest[6]. Through its Medicare programmed, the government foots a sizable percentage of the older population's medical costs. India's government spends 1.5 percent of its annual GDP on public healthcare, which is significantly less than that of other countries. Global public health spending, on the other hand, has almost doubled in line with inflation in the last two decades, reaching US $8.5 trillion in 2019, or 9.8% of global GDP.[16] Multinational multi-private sectors provide around 60% of comprehensive medical treatments and 70% of out-patient care, which charge patients astronomically high fees[7], Because of the rising expense of quality healthcare, increased life expectancy, and the epidemiological shift toward non-communicable diseases, health insurance is becoming an essential commodity for everyone. Accordingly, predicting such costs with accuracy is a significant first step in addressing this problem[8]. Since the 1980s, there has been research on the predictive modeling of medical costs based on (health insurance) claims data using heuristic rules and regression methods. These methods, however, have not been appropriately validated using populations that the methods have not seen. We utilize modern data-mining methods, specifically classification trees and clustering algorithms, along with claims data from over 800,000 insured individuals over three years, to provide rigorously validated predictions of health-care costs in the third year, based on medical and cost data from the first two years.[9] The analysis of costs in clinical and public health care has become a standard part of both experimental and epidemiological research. [17] This is motivated by the growing interest in controlling public expenditure, in view of adopting interventions or treatments on the basis of their cost-effectiveness.[10] The healthcare system is obviously very concerned about cost evaluation and control, and predictive models are needed to understand how costs behave as a function of given patients' or centers' characteristics. Insurance data has increased dramatically in the last decade, and carriers now have access to it. The health insurance system explores predictive modeling to boost its business operations and services[11]. Computer algorithms and Machine Learning (ML) is used to study and analyze the past insurance data and predict new output values based on trends in customer behavior, insurance policies, and data-driven business decisions, and support in formulating new schemes. Additionally, ML has found enormous and potential applications in the insurance industry[12].

.

III. PROPOSED WORK

The model begins with gathering the data and mapping out the crucial attributes. Then the data is cleaned to match the format of the model. Then the data is segregated into training and testing data. After that classifier uses the trainingdata to train the classifier and later evaluate the classifier. Then a program is used to calculate a score that depicts the correctness of the model.

1. Collection of Data, 2. Data Preprocessing, 3. Data Analysis or EDA, 4. Train and Test Split, 5. Machine Learning Models, 6. Evaluation

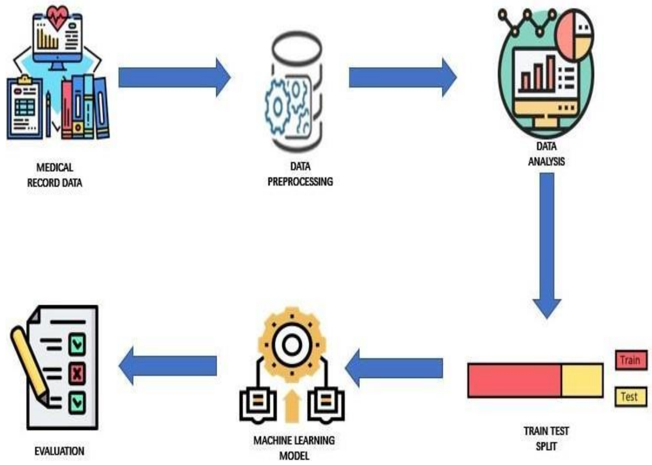


Figure 1. Architecture Diagram

As shown in Fig. 1, this model works in the following six steps: -

1. Collection of data

The first step of the implementation includes collecting the dataset for the model. After getting an adequate dataset, data analysis can be conducted on the dataset to explore the dataset further. This dataset contains medical records a generalized data regarding the patients. The dataset is containing the basic details and, in this data, we following with region wise exploration.

2. Data Preprocessing

Data preprocessing is a vital step in the development ofthe model. In the beginning, when we collect the data, it might not be suitable for the model we are building, which can result in altered outputs. As the dataset selected might follow a different format than the one needed, we need to standardize the structure of all the attributes in the data. Further, we need to deal with noisy data, duplicates, and missing data and perform attribute scaling. These steps are essential as the raw and unprocessed data can be put into the model as it might affect its working and thus the accuracy of the model. The primary procedurefollows:

- Filling in the missing data: A missing value, as the name suggests, means an empty cell left due to a mistake or by choice.
- Inconsistencies: the collisions occurring for any two data points must be removed.
- Outliers: Any abnormal values must be removed to maintain the integrity of the dataset.

The data in the database used for this model is clean.

3. Data Analysis or EDA

The first step of the implementation includes collecting the dataset for the model. After getting an adequate dataset, data analysis can be conducted on the dataset to explore the dataset further. This dataset contains the medical data set for a particular patient as the data contains the basic attributes, the significant thing is we are analyzing the region wise analysis, and also taking in consideration of one classification on basis of smoking, which provide the better classification with respect to analysis.

In Fig2, we provided the Age Distribution, here we referred to the given data set and shows the significance of the age distribution graph, In the healthcare, predicting medical costs is a critical task that plays a pivotal role in resource allocation, insurance premium determination, and healthcare planning. One of the significant factors influencing medical costs is the age of the individuals seeking healthcare services. Analyzing the age distribution of data within the context of medical cost prediction can provide valuable insights into healthcare expenditure patterns and assist in making informed decisions. The significance of the age distribution comes into role of predicting the analysis over any practical data. Age is a fundamental demographic variable that profoundly impacts an individual's healthcare needs and costs. As people age, they are more likely to experience chronic illnesses and age-related health conditions, which often require ongoing medical attention and costly treatments. Additionally, age can influence health behaviors, lifestyle choices, and the frequency of healthcare utilization. Here we taken a data set of the age group of 18 years to 64 years, as our analysis is based on smoking habits, that's why we didn't included the age group of below 18. In the given data set we also mentioned the density of the individual over an age of persons.

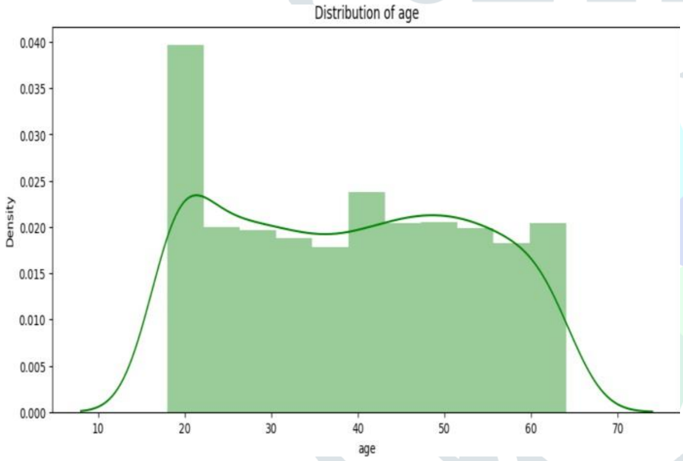


Figure 2. Distribution of age

In Fig.3, we plotted the Distribution of charges of smokers, Analyzing the distribution of charges among smokers within a medical cost prediction dataset is crucial for understanding the financial implications of smoking on the healthcare system. It can provide evidence-based insights for healthcare policy decisions, insurance premium calculations, and targeted smoking cessation programs. By exploring the dataset's descriptive statistics, visual representations, hypothesis testing, and regression analysis, we provided the impact of smoking on medical costs and ultimately contribute to improving public health outcomes. The graphs show the results which shows the trend for the smoking people over there charges of expenditure, and we plotted against the density of peoples for which it concerned,

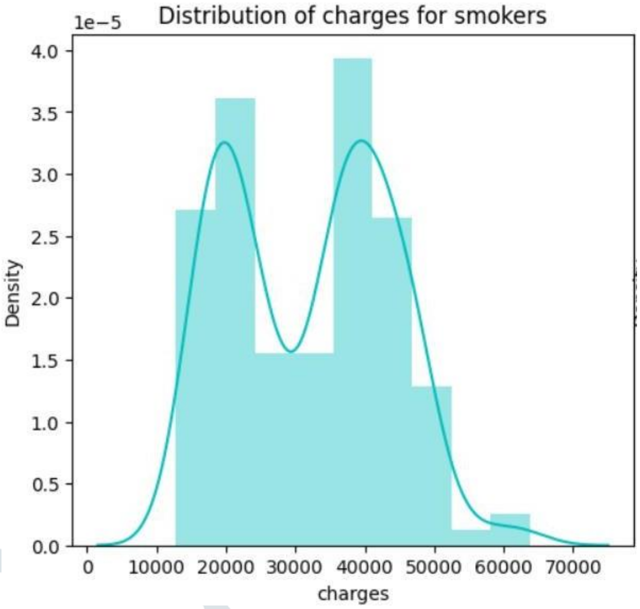


Figure 3. Distribution of charges for smokers

In Fig.4, We plotted the Distribution of charges for the non-smokers, the distribution of medical charges among non-smokers in a medical cost prediction dataset is a crucial component in understanding the financial implications of smoking behavior. By examining the mean, median, standard deviation, and visual representations of charges, we can gain valuable insights into the financial aspects of healthcare for non-smokers. In this graphical analysis, we found that the non-smokers characteristics is drastically varies as compare to the smoker distribution.

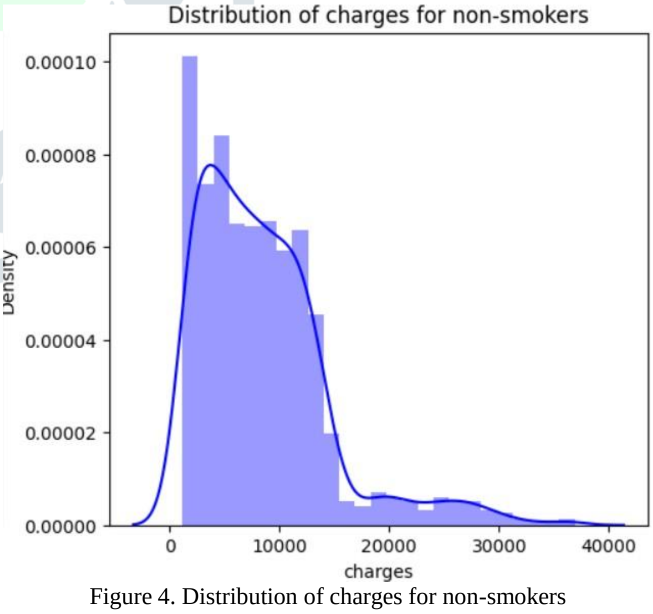


Figure 4. Distribution of charges for non-smokers

In Fig.5, we plotted the data over a smoker and non-smoker of age group 18, in which it provided the graphical representation, the distribution of smokers and non-smokers within a dataset is a critical factor in various health-related analyses, such as understanding the impact of smoking on health outcomes, designing public health interventions, and estimating healthcare costs. This data explores the significance of discerning the number of smokers and non-smokers within a dataset and discusses how this information can influence health research and policy decisions. Tracking the distribution of smokers and non-smokers over time can provide insights into the effectiveness of tobacco control policies and public health campaigns. Researchers can analyze changes in smoking prevalence and assess whether policy interventions have had a positive impact on reducing the number of smokers. This longitudinal analysis is vital for continuously improving public health strategies. The aspects of this exploration provide the distribution of smokers and non-smokers within a dataset is a vital aspect of health research and policymaking. It influences our understanding of health outcomes, guides the design of public health interventions, helps estimate healthcare costs, identifies health disparities, and allows for long-term trend analysis. Accurate and up-to-date data on smoking prevalence are essential for informed decision-making, ultimately contributing to improved public health outcomes and reduced healthcare costs. The graph includes the insights of the gender of the particular individual over a smoker and non-smoker habit, which provides the precisive data analysis over a data set. The representation of the gender is shown by 0 and 1 , here 0 represents the male and 1 represents female over gender prospects.

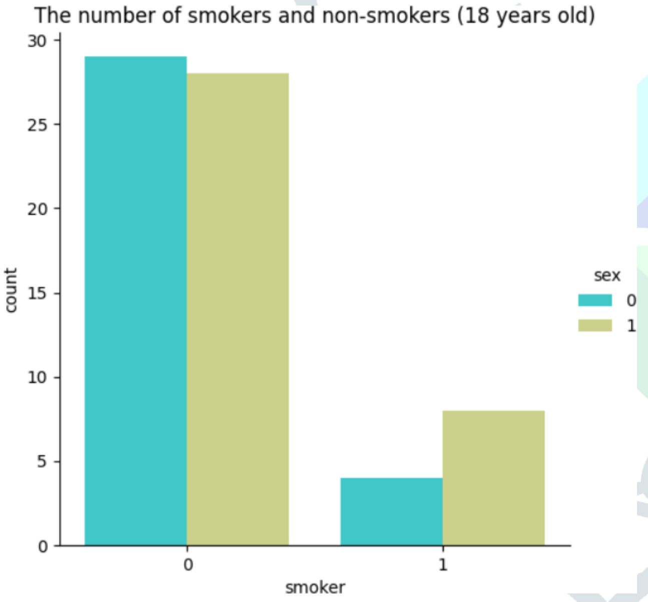


Figure 5. The number of smokers and non-smokers (18-year-old)

In Fig6, we provided the analysis over a correlation matrix, the context over a terms as, correlation matrix provides a quantitative measure of the linear relationship between pairs of variables. For example, it can reveal whether there is a correlation between age and medical charges. If there is a strong positive correlation, it suggests that as age increases, medical charges tend to increase as well. Correlation analysis can help in feature selection for building predictive models. Variables that have a strong correlation with the target variable (in this case, medical charges) are more likely to be important predictors. This information guides the selection of relevant features, improving the model's accuracy and interpretability. Multicollinearity occurs when two or more independent variables in a regression model are highly correlated. It can lead to unstable model coefficients and difficulties in interpreting their effects. By examining the correlation matrix, you can identify and address multicollinearity issues, potentially by selecting one of the correlated variables or by using advanced modeling techniques. As we include the "region" variable, the correlation matrix can help you understand how healthcare costs vary by region. For instance, it might indicate whether certain regions have higher or lower medical charges. This information can be valuable for resource allocation and policy planning. By examining correlations, you can assess risk factors associated with medical costs. For instance, if being male (sex=0) has a strong positive correlation with higher medical charges, it implies that being male is associated with increased healthcare costs. This insight is valuable for actuarial purposes and setting insurance premiums. Here keep all points in consideration we provided the correlation matrix of all parameters.

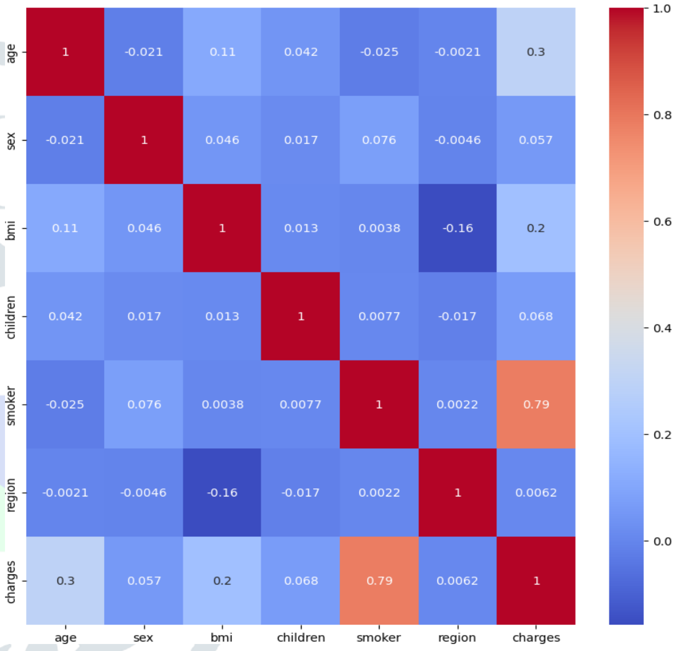


Figure 6. Correlation Matrix

In Fig 7-8, In this we provided the Box plot representation of particular data, in which it's shows the data ranges over a span of range, also the assumption is true, that the medical expense of males is greater than that of females. In addition to that medical expense of smokers is greater than that of non-smokers. We will focus on creating box plots for medical charges, comparing men and women, to better understand the distribution of charges within each gender group. This visualization technique allows us to identify differences in the spread and central tendency of charges, as well as the presence of potential outliers. Box plot analysis of medical charges for men and women within the dataset allows us to explore gender-based variations in healthcare expenses. By visualizing the central tendency, spread, and potential outliers in medical charges, we can identify trends and disparities that can inform decision-making in the fields of insurance, healthcare provision, and public health policy. In box plotting we already took all charges parameters which is making the relationship between the variation of charges in the context of both men and women.

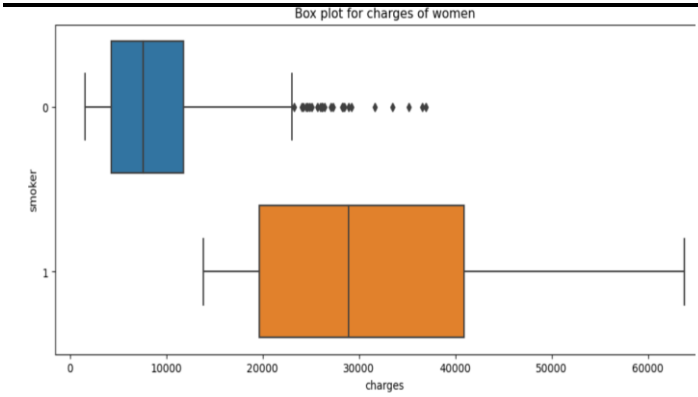


Figure 7. Box Plot for Charges of women

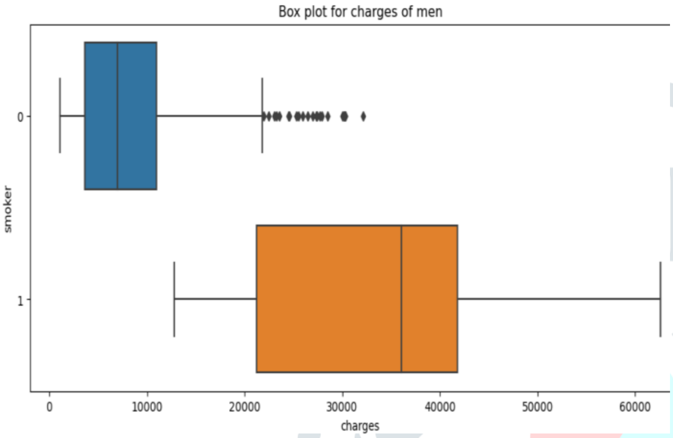


Figure 8. Box Plot for Charges of men

In Fig 9-10, here we provided the scatter plot of charges with non-smokers, smoker, and cumulative plot.

In Fig.9, Majority of the points shows that medical expense increases with age which may be due to the fact that older people are more prone to illness. But there are some outliners which shows that there are other illness or accidents which may increase the medical expense. In Fig10, here we see peculiarity in the graph. In the graph there are two segments, one with high medical expense which may be due to smoking related illness and the other with low medical expense which may be due age-related illness.

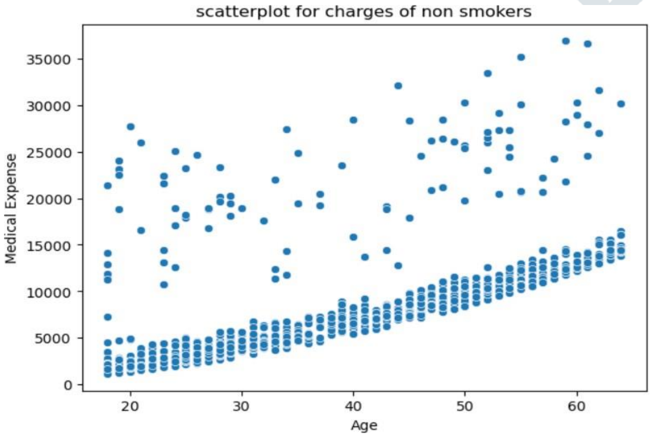


Figure 9. Scatterplot for charges of non-smokers

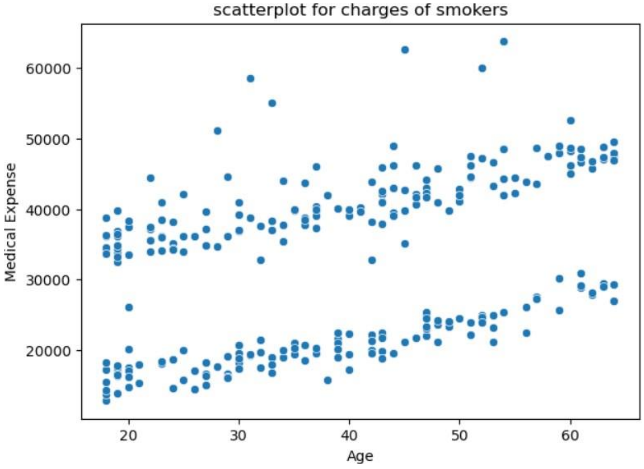


Figure 10. Scatterplot for charges of smokers

In Fig11, we provided the overall outliers and marginal flow of the medical expenses vs patient's age with smoking ability, Now, we clearly understand the variation in charges with respect to age and smoking habits. The medical expense of smokers is higher than that of non-smokers. In non-smokers, the cost of treatment increases with age which is obvious. But in smokers, the cost of treatment is high even for younger patients, which means the smoking patients are spending upon their smoking related illness as well as age related illness. The advantage of the scatter plot is that Scatter plots are valuable tools for visualizing the distribution of medical charges among men and women based on demographic parameters such as age and BMI. By analyzing these plots, we can gain insights into how these factors influence healthcare expenses and potentially inform insurance pricing, healthcare policy decisions, and personalized patient care strategies. Additionally, these visualizations can help identify any gender-based disparities in medical costs, contributing to a more equitable healthcare system.

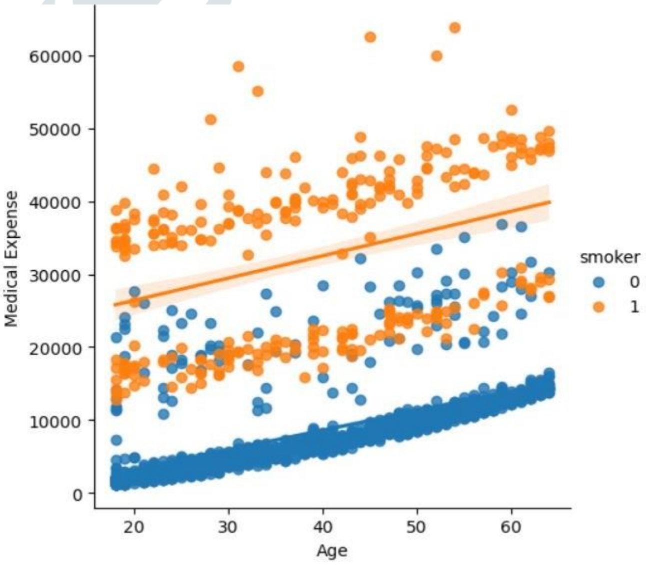


Figure 11. Scatterplot for charges of smokers

In Fig 12-13, here we provided the charges distribution of the obese vs non obese people, in which we provided the trends of charges, here the observaion is like, Charges distribution for patients with BMI less than 30 i.e. healthy patients , Therefore, patients with BMI less than 30 are spending less on medical treatment than those with BMI greater than 30. In examining the distribution of charges within a dataset that distinguishes between obese and non-obese individuals, we gain crucial insights into the financial implications of obesity in healthcare. Typically, obesity is associated with higher healthcare costs due to its correlation with various health conditions. Analyzing the distribution of charges allows us to quantify this relationship. By comparing the two groups, we can ascertain whether medical expenses significantly differ between obese and non-obese individuals. This analysis not only aids insurers in refining pricing models but also informs healthcare providers and policymakers about the economic consequences of obesity. It underscores the importance of preventive measures and interventions aimed at reducing obesity rates, as they can potentially alleviate the financial burden on both individuals and the healthcare system, while improving overall public health and well-being.

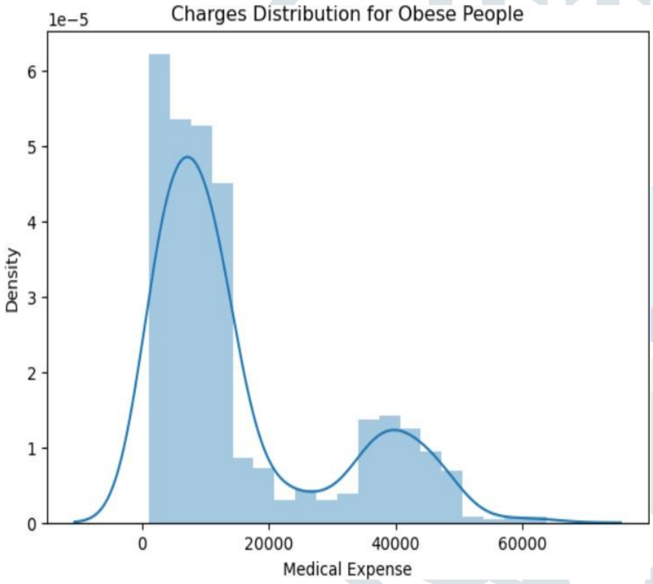


Figure 12. Charges Distribution for Obese People

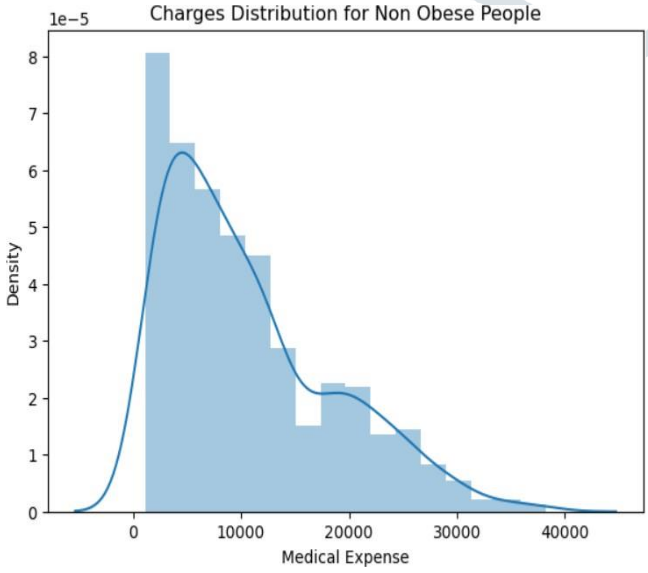


Figure 13. Charges Distribution for Non-Obese People

In Fig 14-15, here we provided the resultant graph regarding the charges distribution non-smokers and smokers which include the age vs charges distribution, here x-axis denotes the age, and y-axis denotes charges.

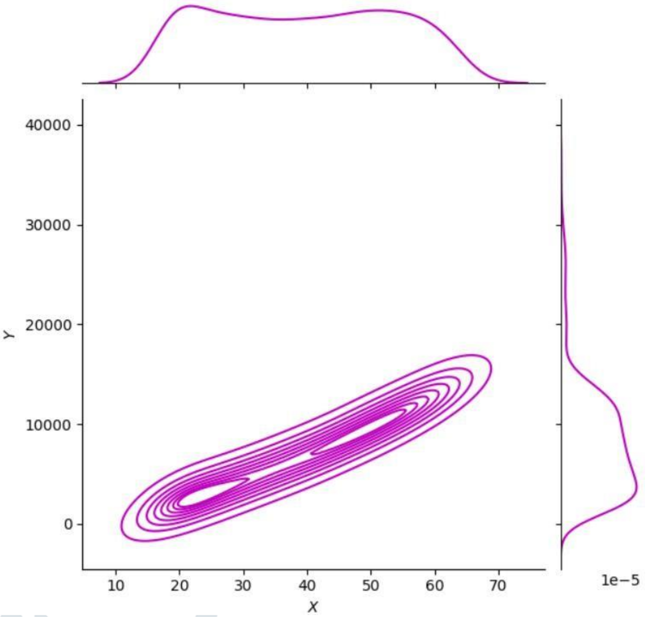


Figure 14. Distribution of charges and age for non-smokers

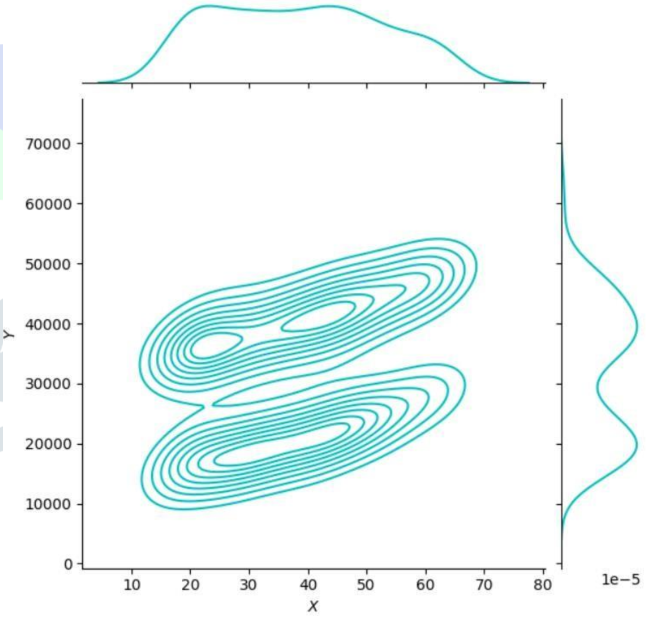


Figure 15. Distribution of charges and age for smokers

In Fig 16., The comparison of actual vs. predicted values is a fundamental step in evaluating the performance and accuracy of a linear regression model. In linear regression, the model's primary objective is to estimate a linear relationship between independent variables and a dependent variable. The actual values refer to the real-world data points of the dependent variable that we are trying to predict, while the predicted values are the values generated by the linear regression model based on the input data and model parameters. When we compare actual and predicted values, we essentially assess how well the model has captured the underlying relationships within the data. This comparison allows us to quantify the model's predictive power. By calculating metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), or R-squared (coefficient of determination), we can determine the extent to which the model's predictions align with the actual data. A lower MSE or RMSE and a higher R-squared value indicate a better fit and higher predictive accuracy, suggesting that the linear regression model provides a good representation of the data. Conversely, a large discrepancy between actual and predicted values may signal that the model needs further refinement or that other variables should be considered.

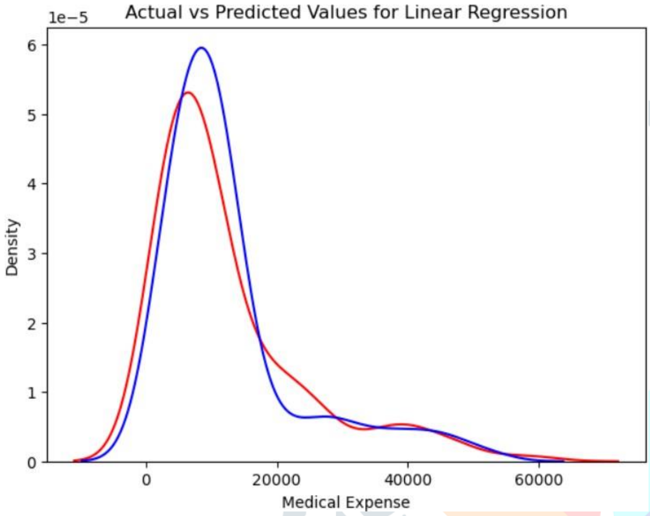


Figure 16. Actual vs Predicted Values for Linear Regression

In Fig 17., The comparison of actual vs. predicted values is a fundamental step in evaluating the performance and accuracy of a polynomial regression model. In polynomial regression, which is a type of regression analysis used for modeling non-linear relationships between variables, the comparison between actual and predicted values is crucial for assessing the model's performance. When we conduct polynomial regression, we fit a polynomial equation to the data to estimate a relationship between the independent and dependent variables. After building the model, we use it to predict values of the dependent variable based on the given independent variable(s). The comparison between the actual observed values and the values predicted by the polynomial regression model allows us to evaluate how well the model captures the underlying patterns in the data. By calculating the difference between the actual and predicted values, we can assess the accuracy and effectiveness of the model. A good polynomial regression model will minimize these differences, yielding predicted values that closely align with the actual data points. Conversely, a poor model will exhibit larger discrepancies between actual and predicted values, indicating a lack of fit or predictive power. This evaluation process is vital for understanding the model's reliability and its ability to make accurate predictions, which is essential in various fields such as economics, biology, and engineering, where non-linear relationships are common and understanding these relationships can have significant implications.

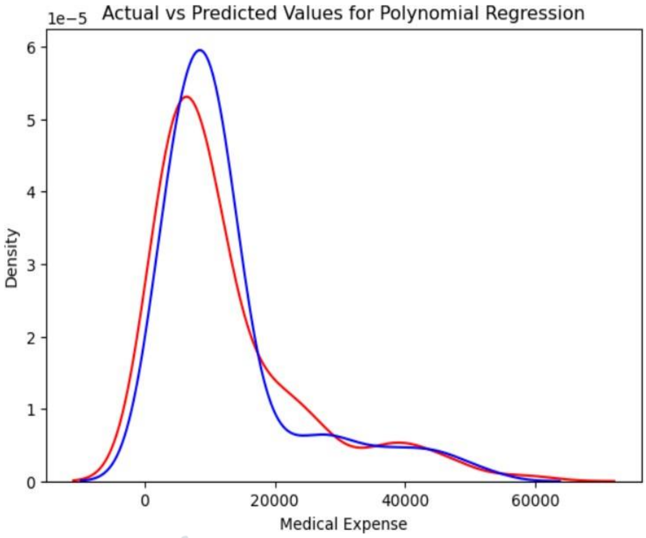


Figure 17. Actual vs Predicted Values for Polynomial Regression

In Fig 18., The comparison of actual vs. predicted values is a fundamental step in evaluating the performance and accuracy of a decision tree regression model, In the context of decision tree regression, the comparison between actual and predicted values is a fundamental step in assessing the model's performance and accuracy. When using a decision tree for regression tasks, the model learns to make predictions by recursively partitioning the dataset into subsets and assigning a predicted value (typically the mean or median of the target variable) to each subset. To evaluate how well the decision tree model is performing, we compare its predictions to the actual values from the dataset. This comparison allows us to quantify the extent to which the model's predictions align with the real-world data. By calculating metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), or R-squared (R2), we can objectively assess the model's ability to capture the underlying patterns and relationships within the data. The closer the predicted values are to the actual values, the lower the error metrics, indicating a more accurate regression model. This evaluation is crucial in determining whether the decision tree model is suitable for making reliable predictions and can inform any necessary adjustments or improvements in the modeling process.

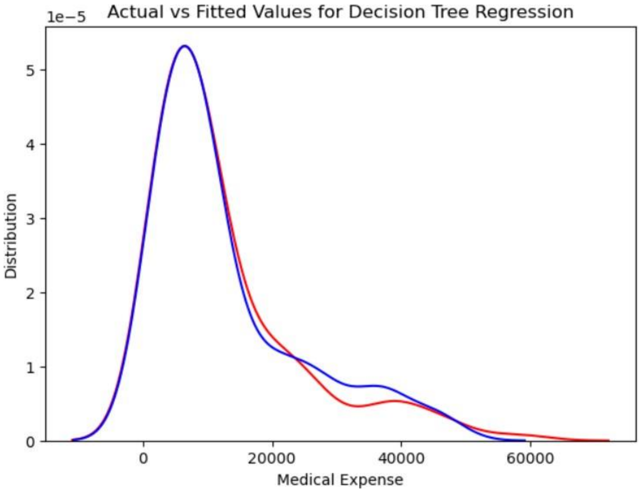


Figure 18. Actual vs Predicted Values for Decision Tree Regression

In Fig 19., The comparison of actual vs. predicted values is a fundamental step in evaluating the performance and accuracy of a random forest regression model ,In a Random Forest Regression model, the comparison between actual and predicted values is a crucial step to assess the model's performance and accuracy. After training the Random Forest on a dataset with known outcomes, it uses an ensemble of decision trees to make predictions on new, unseen data. The "actual values" represent the real target values from the dataset, which serve as the ground truth or the correct values we aim to predict. On the other hand, the "predicted values" are the outcomes generated by the Random Forest model for the same set of input features. By comparing these actual and predicted values, we can quantify how well the model performs in approximating the true relationship between the input features and the target variable. In essence, this comparison helps us understand the model's ability to capture the underlying patterns and variability in the data. Evaluating the closeness of actual and predicted values through metrics like Mean Absolute Error (MAE), Mean Squared Error (MSE), or R-squared allows us to gauge the predictive power of the Random Forest model and make informed decisions about its suitability for a particular regression task.

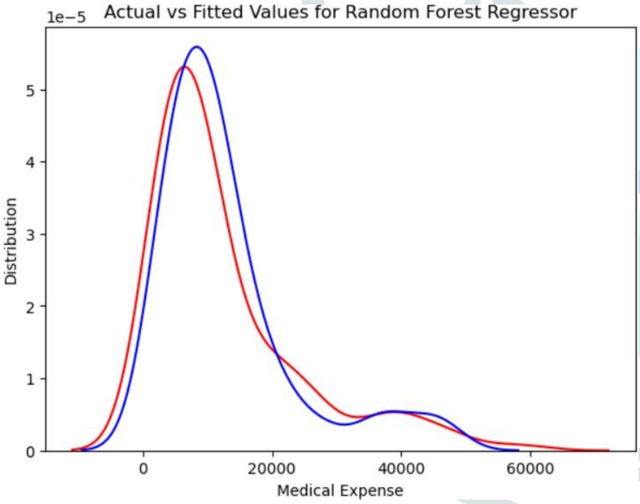


Figure 19. Actual vs Predicted Values for Random Forest Regression

4. Train and Test Split

Under this split, the dataset is divided into a training database to construct the model(classifier) and a testing database to evaluate the model.In this model, the split is conducted through cross-validation, which divides the database into k parts and iterates k times. In the initial iteration, the dataset is trained with (k-1) parts, and the remaining last is used as the testing dataset. The model is trained with (k-2) parts in the next iteration, and the two are used for the testing phase. The procedure runs till the last iteration where the first part is the testing, and the nine parts are used as the training dataset. Depicts the split of the dataset based on the cross-validation technique.

5. Machine Learning Models

In the Machine Learning Models section, some machine learning models have been discussed:

1) Linear Regression: This algorithm is used when you want to predict a continuous target variable based on one or more input features. It models the relationship between the dependent variable (the target) and one or more independent variables (the features) by fitting a linear equation to the observed data.

2) Random Forest Classifier: Multiple decision trees are used in this ensemble learning technique to provide predictions. Each tree is trained using a random subset of data elements and characteristics, resulting in a forest of decision trees. Individual trees' forecasts are combined by voting or average to get the final projection. The resilience, scalability, and capacity for handling high-dimensional data of Random Forests are well recognized. The accuracy was much improved using the Random forest classifier.

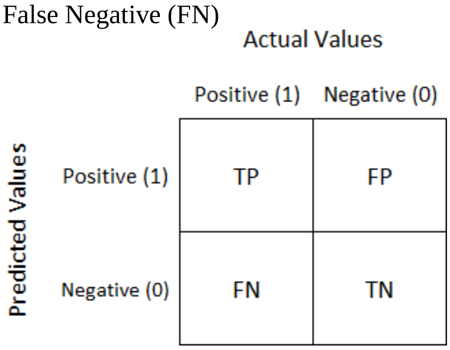
3) Polynomial Classifier: Polynomial regression is not typically used as a classifier; it's primarily used for regression tasks. In polynomial regression, you model the relationship between a dependent variable and one or more independent variables by fitting a polynomial equation to the data. This allows you to capture nonlinear relationships between the variables.

4) Decision Tree Classifier: It is a well-liked technique for both regression and classification tasks. It builds a model of choices and potential outcomes that resembles a tree. A test performed on a feature is represented by each of the internal nodes of the tree, the test result is represented by each branch, and a class labelling or value is represented by each leaf node. Decision Trees can manage both category and numerical data and are simple to grasp and analyze.

6. Evaluation

The cross-validation technique gives the model merit for using the complete dataset for training and testing. The evaluation is grouped with the accuracy of the model. In general, the assessment can be accomplished using a confusion matrix which is formed from the following, shown in Fig. 21:

○ True Positive (TP)
○ False Positive (FP)
○ True Negative (TN)
○ False Negative (FN)



		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

Fig. 21: Confusion Matrix

Relying on the aforementioned we can calculate the precision of the prototype.

IV-PERFORMANCE ANALAYSIS

The ML classifier, such as the logistic regression, random forest classifier, support vector classifier, decision tree classifier, and KNN classifier, gives an accuracy of 99%

range. Compared to any other research paper, we provided the cumulative comparison between the various Machine learning models with proper accuracy. Our results show that the logistic regression, random forest classifier, support vector classifier, decision tree classifier, and KNN classifier model can achieve an accuracy of 99.91%, 99.959%, 99.93%, 99.92%, 99.95% in detecting fraudulent transactions.

In the above performance analysis, we provided a better figure of merit in the context of fraud detection, and the algorithm we worked on is far better than any work. These works indicate a better F-1 score towards the analysis of a particular function on which we implemented. Performance evaluation in these works depends upon the accuracy factor, in which the accuracy range is increased from 94% scale to 99% scale.

IV-CONCLUSION

In conclusion, the present study has shown that applying ML techniques to CC fraud detection can yield highly accurate results. By comparing the various Machine learning models, we achieved an accuracy rate of 99%, which suggests that this approach can significantly contribute to preventing fraudulent transactions. Implementing such a model can greatly benefit financial institutions and their customers, as it allows for faster and more reliable detection of suspicious activities. Moreover, by minimizing false positives, this method can help reduce the workload of fraud analysts, thus enabling them to focus on more complex cases that require human attention. Overall, the findings of this research demonstrate the potential of ML algorithms for fraud detection in the financial sector. Further research can explore the effectiveness of other techniques and models and the generalizability of our approach to different types of fraud. Ultimately, such advancements can create a more secure and trustworthy financial system for everyone.

In the future, we intend to improve the model's working and increase the security of data considering practical occurrences. As we applied the multiple machine learning models, we needed to conduct precise research on indentation towards data analysis.

REFERENCES

[1] A. Mahajan, V. S. Baghel and R. Jayaraman, "Credit Card Fraud Detection using Logistic Regression with Imbalanced Dataset," 10th International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, India, 2023, pp. 339-342.

[2] Ryll, Lukas, and Sebastian Seidens. "Evaluating the performance of machine learning algorithms in financial market forecasting: A comprehensive survey." arXiv preprint arXiv:1906.07786 (2019).

[3] Lorenzo Famiglini, Giorgio Bini , Anna Carobene, Andrea Campagner, Federico Cabitza. " Prediction of ICU admission for COVID-19 patients: a machine learning approach based on Complete Blood Count Data". The paper was published in (2021).

[4] Eunbi Kim, kap Su Han, Taesu Cheong, Sung Woo Lee, Joonyup Eun, Su Jin Kim. " Analysis on Benefits and Costs of Machine learning -Based Early Hospitalization Prediction." The paper was published in March 29,2022.

[5] MD Sadique Hasan, Chad Sundberg, Hasibul Hasan, Yordan Kostov, Xudong Ge, Fow-Sen Choa, Govind Rao."Rapid Bacterial Detection and Identification of bacterial strains using Machine learning methods Integrated with a portable multichannel Fluorimeter" The paper was published in 17 August(2023).

[6] "National Health Accounts", National Health Systems Resource Centre, [online] Available: https://nhsrcindia.org/national-health-accounts records..

[7] Global Expenditure on Health", *WHO annual report*, 2021, [online] Available

[8] "Health Insurance of India's missing middle", *Niti Ayog*, Oct 2021, [online] Available.

[9] L. Moran, P. J. Solomon, A. R. Peisach and J. Martin, "New models for old questions: generalized linear models for cost prediction", *Journal of evaluation in clinical practice*, vol. 13, no. 3, pp. 381-389, 2007

[10] S. Sushmita, S. Newman, J. Marquardt, P. Ram, V. Prasad, M. D. Cock, A. Teredesai et al., "Population cost prediction on public healthcare datasets", Proceedings of the 5th International Conference on Digital Health 2015, pp. 87-94, 2015.].

[11] Medical Cost Prediction Dataset, [online] Available: https://www.kaggle.com/hely333/eda-regression/data.

[12] Donald W. Marquardt, Ronald D. Snee et al., "Ridge Regression in Practice", *"The American Statistician"*, vol. 29, pp. 3-20, 2012

[13] S. V. S. S. Lakshmi, S. D. Kavilla "Machine Learning for Credit Card Fraud Detection System", International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 24 (2018) pp. 16819-16824.

[14] N. Malini and M. Pushpa, "Analysis on credit card fraud identification techniques based on KNN and outlier detection," 3rd International Conference on Advances in Electrical, Electronics, Information, Communication and Bioinformatics (AEEICB), Chennai, India, 2017, pp. 255-258.

[15] C. Navamani, S. Krishnan, "Credit Card Nearest Neighbor Based Outlier Detection Techniques", International Journal of Computer Techniques -– Volume 5 Issue 2, 2018, pp. 54-60.

[16] A. Maurya and A. Kumar, "Credit card fraud detection system using machine learning technique," International Conference on Cybernetics and Computational Intelligence (CyberneticsCom), Malang, Indonesia, 2022, pp. 500-504.

[17] A. Thennakoon, C. Bhagyani, S. Premadasa, S. Mihiranga and N. Kuruwitaarachchi, "Real-time Credit Card Fraud Detection Using Machine Learning," 2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, 2019, pp. 488-493.