इलाहाबाद विश्वविद्यालय
UNIVERSITY OF ALLAHABAD
(A Central University)

Health Insurance Premium Prediction Using Machine Learning Models

(Minor Project Report)

Submitted by

Priya Rai

M.A 3rd Sem

Roll No:- 24355067

Enrolment No:- U2273002

Under the Supervision of

Dr. Prashant Verma Sir

Submitted to

Department of Statistics

Faculty of science

University of Allahabad

Certificate

This is to certify that Miss Priya Rai has completed her minor project on the topic "**Health Insurance Premium Prediction** " under the supervision of **Dr. Prashant Verma sir** . This report is submitted to the Department of Statistics, University of Allahabad for the partial fulfilment of requirement for the award of M.A Degree.

Project Supervision **Head of Department**

(Dr Prashant Verma sir) (Dr P. S. Pundir sir)

Acknowledgement

I would like to express my sincere gratitude to all those who contributed to the successful completion of the Health Insurance Premium Prediction project. This endeavor would not have been possible without the support, guidance, and expertise of my Project Supervisor Dr Prashant Verma sir Thank you sir for your valuable guidance, mentorship, and continuous support throughout the project. Your insights and leadership were instrumental in shaping the direction of our work. And my also help me lot .

Abstract

Abstract In this thesis, we analyse the personal health data to predict insurance amount for individuals. Four models naming Linear Regression ,SVM , RandomForest Regressor and Gradient Boosting Decision tree Regression have been used to compare and contrast the performance of these algorithms. Dataset was used for training the models and that training helped to come up with some predictions. Then the predicted amount was compared with the actual data to test and verify the model. Later the accuracies of these models were compared. It was gathered that gradient boosting algorithms performed better than the linear regression and decision tree. Gradient boosting is best suited in this case because it takes much less computational time to achieve the same performance metric.

Objective

The objective of health insurance cost prediction is to estimate the future healthcare expenses for individuals or groups.

Develop predictive models to improve the accuracy of health insurance premium predictions.

Leverage advanced data science and machine learning techniques for robust analysis.

Establish a comprehensive evaluation framework, including metrics such as mean absolute error, root mean squared error, and precision-recall curves.

1. INTRODUCTION

The goal of this project is to allows a person to get an idea about the necessary amount required according to their own health status. Later they can comply with any health insurance company and their schemes & benefits keeping in mind the predicted amount from our project. This can help a person in focusing more on the health aspect of an insurance rather than the futile part.

Health insurance is a necessity nowadays, and almost every individual is linked with a government or private health insurance company. Factors determining the amount of insurance vary from company to company. Also people in rural areas are unaware of the fact that the government of India provide free health insurance to those below poverty line. It is very complex method and some rural people either buy some private health insurance or do not invest money in health insurance at all. Apart from this people can be fooled easily about the amount of the insurance and may unnecessarily buy some expensive health insurance.

Our project does not give the exact amount required for any health insurance company but gives enough idea about the amount associated with an individual for his/her own health insurance.

Prediction is premature and does not comply with any particular company so it must not be only criteria in selection of a health insurance. Early health insurance amount prediction can help in better contemplation of the amount

needed. Where a person can ensure that the amount he/she is going to opt is justified. Also it can provide an idea about gaining extra benefits from the health insurance.

2.DATASET USED

The primary source of data for this project was from Kaggle user Dmarco. The dataset is comprised of 1338 records with 6 attributes. Attributes are as follow age, gender, bmi, children, smoker and charges .The data was in structured format and was stores in a csv file.

Dataset is not suited for the regression to take place directly. So cleaning of dataset becomes important for using the data under various regression algorithms.

Data Source:

1. Age (age of client)

2. BMI (Body Mass Index)

3.Children (number of children the client have)

4.Sex (Male or Female)

5.Smoker (Whether the client is smoker or not)

6.Region (where the client lives southwest, southeast, northwest , northeast)

7.Charges (Medical Cost the client pay per month)

Data Source Link:

https://www.kaggle.com/datasets/thedevastator/prediction-of-insurance-charges-using-age-gender/data

Software Used:

Python: Python is widely used for data analysis due to its rich ecosystem of libraries and tools tailored for handling, analyzing, and visualizing data. Here's a brief overview of Python's role in data analysis:

Libraries Used: Python boasts powerful libraries for data analysis, including NumPy for numerical operations, Pandas for data manipulation and analysis, Matplotlib and Seaborn for data visualization, and Scikit-learn for machine learning tasks.

1. **Pandas:** The Pandas library provides data structures like DataFrames and Series, making it easy to manipulate and analyze structured data. It supports tasks such as filtering, grouping, merging, and handling missing data.
2. **NumPy:** NumPy is a fundamental library for numerical computing in Python. It provides support for large, multi-dimensional arrays and matrices, along with mathematical functions to operate on these arrays efficiently.
3. **Matplotlib and Seaborn:** These libraries offer a variety of tools for creating static, animated, and interactive visualizations. Matplotlib is a versatile plotting library, while Seaborn builds on it to provide a high-level interface for statistical graphics.
4. **Jupyter Notebooks:** Jupyter Notebooks allow interactive and collaborative data analysis. These notebooks integrate code, visualizations, and narrative text, making it easy to share and reproduce analyses.
5. **Data Cleaning and Preprocessing:** Python facilitates data cleaning and preprocessing tasks through libraries like Pandas, allowing analysts to handle missing values, outliers, and other data quality issues.
6. **Statistical Analysis:** Python provides tools for statistical analysis, hypothesis testing, and modeling. The statsmodels library, for example, supports various statistical models.
7. **Machine Learning:** Python's ecosystem includes Scikit-learn, a powerful machine learning library for tasks such as classification, regression, clustering, and model evaluation.

8. **Integration with Big Data Tools:** Python integrates well with big data tools like Apache Spark, allowing analysts to scale their analyses to large datasets.

3.DESIGNING AND IMPLEMENTATION

A. Data Preparation & Cleaning :

The data has been imported from kaggle website. The website provides with a variety of data and the data used for the project is an insurance amount data. The data included various attributes such as age, gender, body mass index, smoker and the charges attribute which will work as the label for the project. The data was in structured format and was stores in a csv file format. The data was imported using pandas library. The presence of missing, incomplete, or corrupted data leads to wrong results while performing any functions such as count, average, mean etc. These inconsistencies must be removed before doing any analysis on data. The data included some ambiguous values which were needed to be removed.

B. Training

Once training data is in a suitable form to feed to the model, the training and testing phase of the model can proceed. During the training phase, the primary concern is the model selection. This involves choosing the best modelling approach for the task, or the best parameter settings for a given model. In fact, the term model selection often refers to both of these processes, as, in many cases, various models were tried first and best performing model (with the best performing parameter settings for each model) was selected.

C. Prediction

The model was used to predict the insurance amount which would be spent on their health. The model used the relation between the features and the label to predict the amount. Accuracy defines the degree of correctness of the predicted value of the insurance amount. The model predicted the accuracy of model by using different algorithms, different

features and different train test split size. The size of the data used for training of data has a huge impact on the accuracy of data. The larger the train size, the better is the accuracy. The model predicts the premium amount using multiple algorithms and shws the effect of each attribute on the predicted value.

4. Models Used :

Regression: Regression analysis allows us to quantify the relationship between outcome and associated variables. Many techniques for performing statistical predictions have been developed, but, in this project, three models – Multiple Linear Regression (MLR), Decision tree regression and Gradient Boosting Regression were tested and compared.

1.Linear Regression

linear regression analysis is used to predict the value of a variable based on the value of a variable based on the value of the another variable. and multiple linear regression can be defined as extended simple linear regression. It comes under usage when we want to predict a single output depending upon multiple input or we can say that the predicted value of a variable is based upon the value of two or more different variables. The predicted variable or the variable we want to predict is called the dependent variable and the variables being used in predict of the value of the dependent variable are called the independent variables (or sometimes, the predicto, explanatory or regressor variables).

2.Support Vector Machine:

- **Type**: Supervised machine learning algorithm.
- **Objective**:

    - **Classification**: Finds a hyperplane to separate data into classes.
    - **Regression**: Predicts continuous outcomes.
- **Kernel Trick**: Maps data into higher dimensions for non-linear boundaries.
- **Support Vectors**: Key data points influencing the decision boundary.
- **Kernels**: Linear, polynomial, RBF, and sigmoid for different data types.
- **Parameters**: Trade-off between smooth decision boundary and correct classification.it is Influences the decision boundary shape in
- RBF kernel. Maximizes distance between hyperplane and nearest data points. then Hinge loss penalizes misclassifications. Image classification, text categorization, bioinformatics. Effective in high-dimensional spaces, versatile. Computationally intensive for large datasets, sensitive to kernel and parameters.

3.RandomForest Regression:

- **Type**: Ensemble learning algorithm.
- **Objective**:
    - **Classification**: Builds multiple decision trees and combines their predictions.
    - **Regression**: Aggregates predictions for continuous outcomes.
- **Ensemble Approach**: Constructs a forest of decision trees and merges their outputs for more robust predictions.
- **Decision Trees**: Base learners in the forest; each tree is trained on a random subset of the data.
- **Randomization**:

- Randomly selects features for each tree during training.
- Bootstrap sampling: Randomly selects data points with replacement.

Combines predictions from individual trees. Bootstrap Aggregating; combines results from multiple models to improve overall performance and Estimates model performance using data not used during training. and Provides a measure of the importance of each feature in making predictions. Classification, regression, feature selection. Robust, handles high-dimensional data, less prone to overfitting. May be computationally expensive, complex models can be hard to interpret.

4.Gradient Boosting Regression

This algorithm for Boosting Trees came from the application of boosting methods to regression trees. The basic idea behind this is to compute a sequence of simple trees, where each successive tree is built for the prediction residuals of the preceding tree. For predictive models, gradient boosting is considered as one of the most powerful techniques.

Gradient boosting involves three elements:

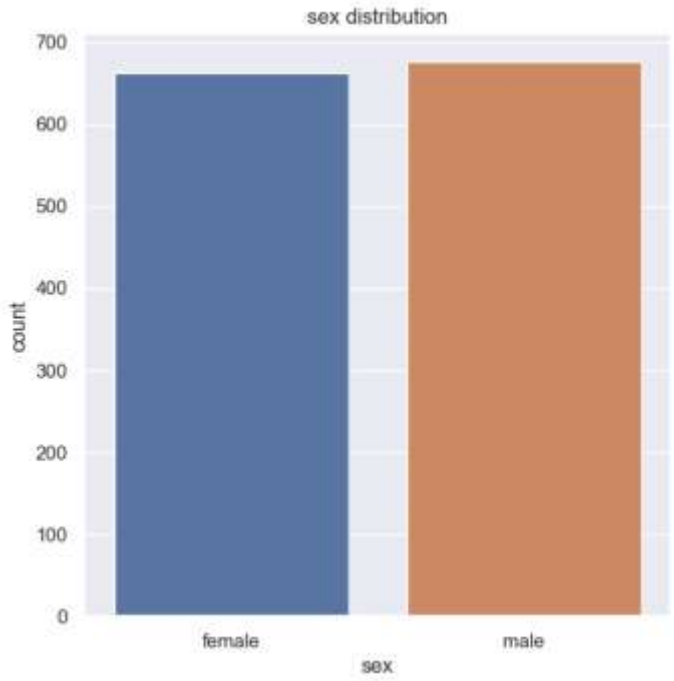
An optimized loss function.

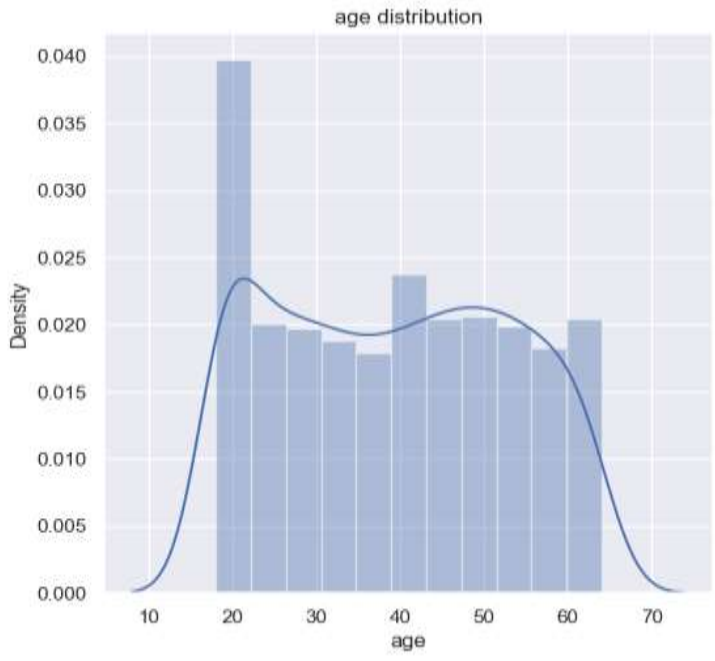
An additive model to add weak learners to minimize the loss function.

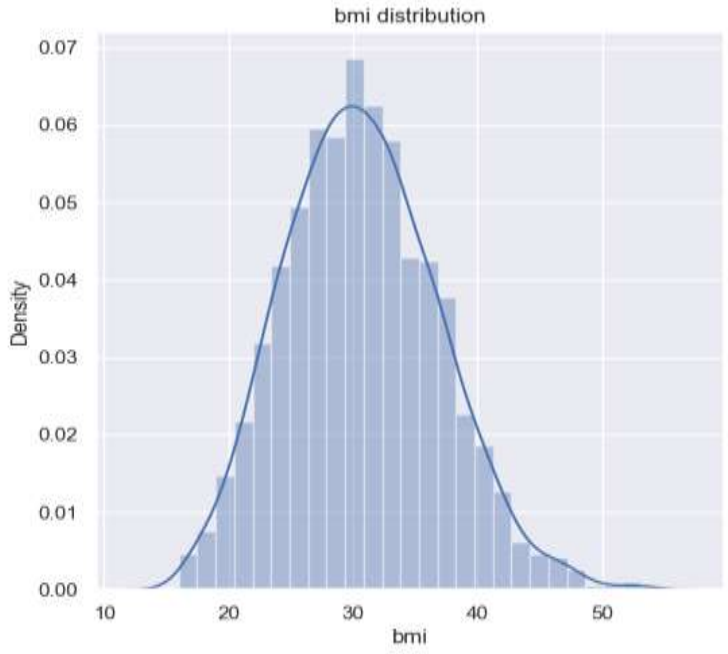
A weak learner to make predictions

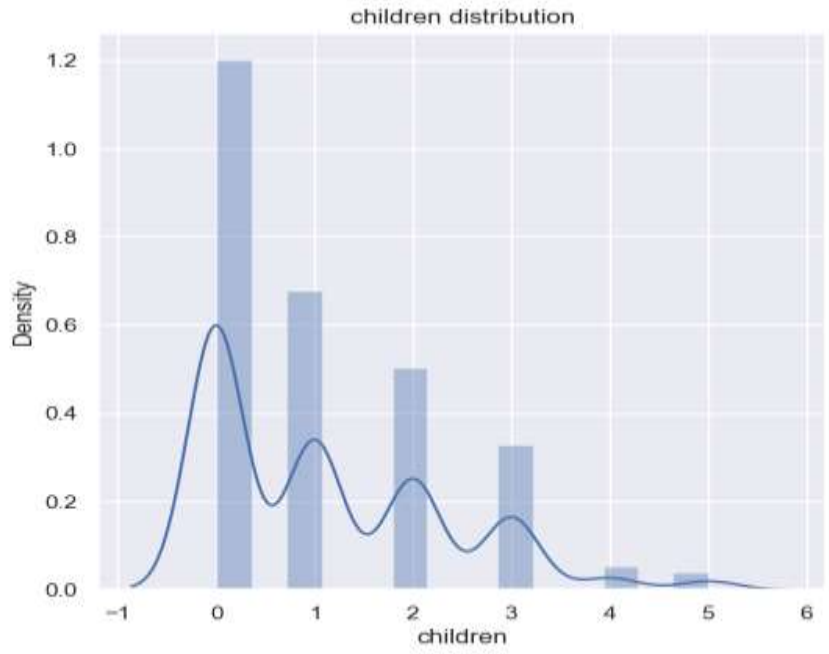
5.Result and Discussion:

Histogram Plots of Feature Variables :



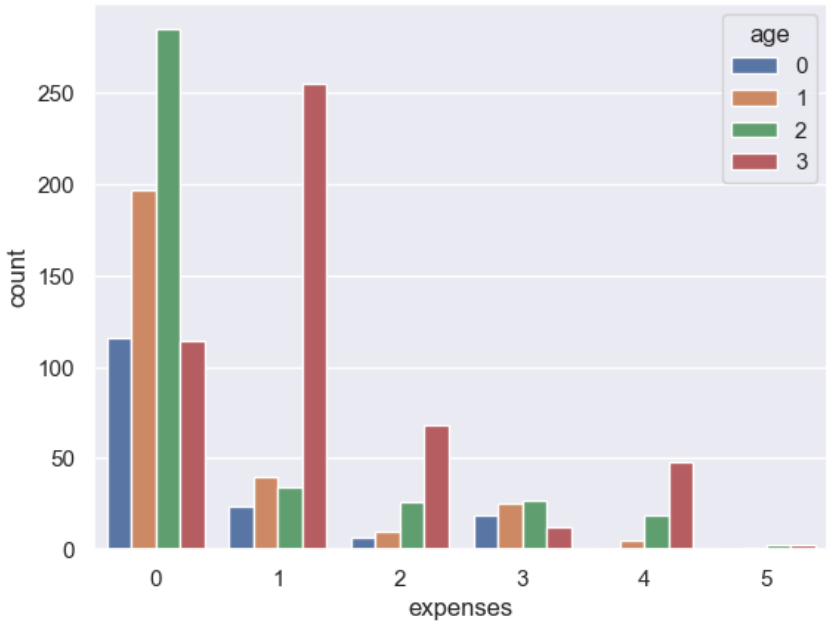



bmi distribution
Density
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
10
20
30
40
50
bmi

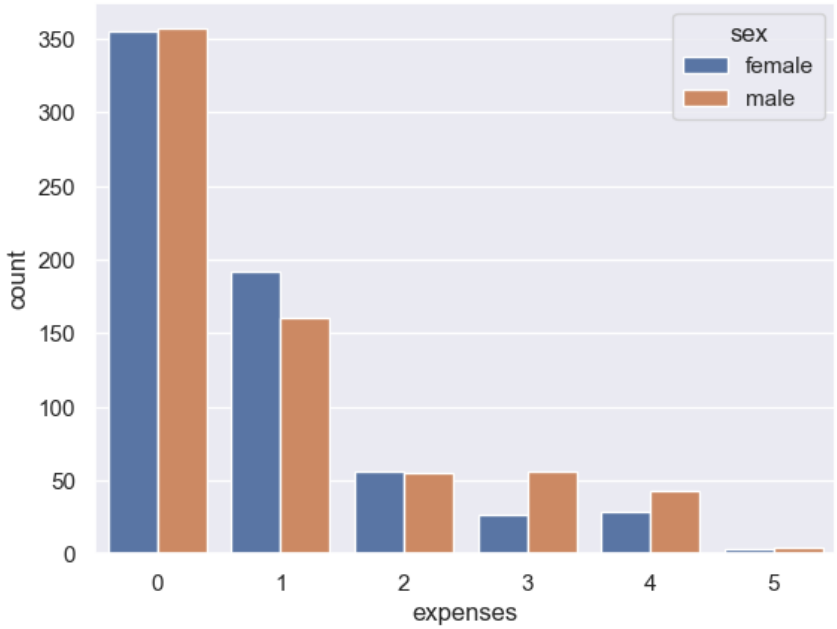

children distribution
Density
1.2
1.0
0.8
0.6
0.4
0.2
0.0
−1
0
1
2
3
4
5
6
children

Bi-variate Analysis of Feature with Target Variable:

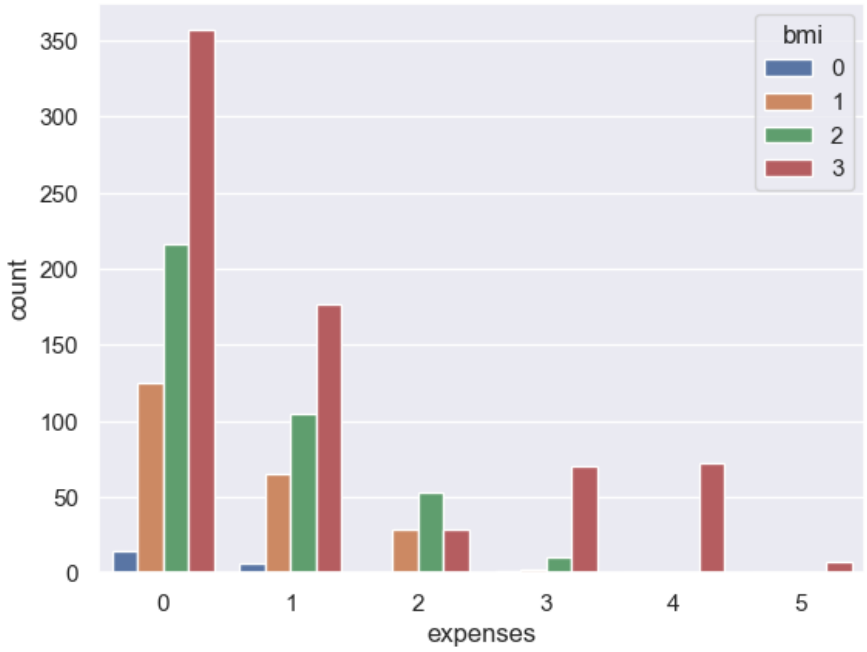
Age with respect to Expenses



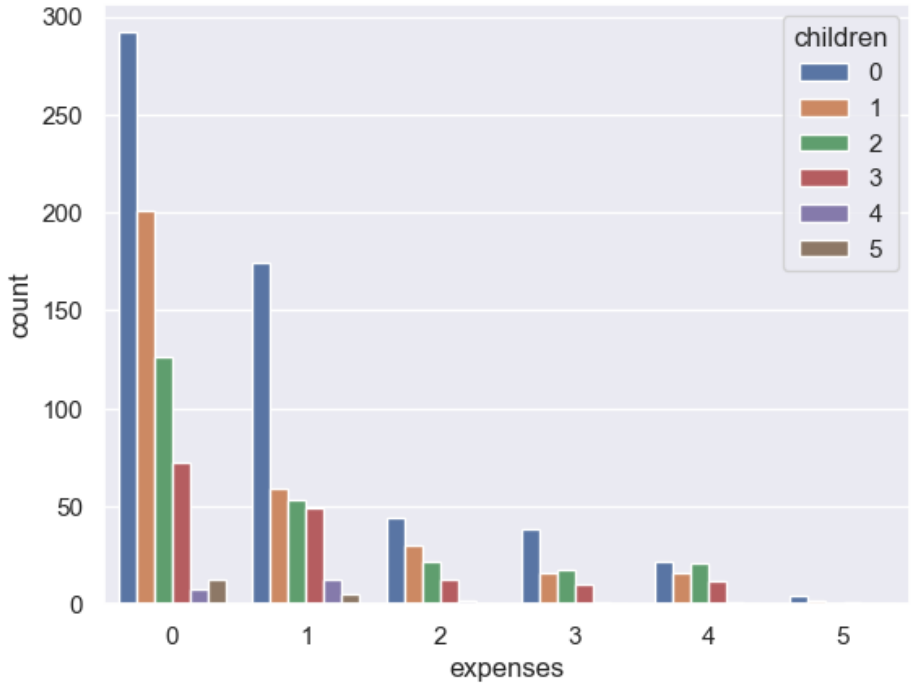
Sex with Respect to Expenses



Body Mass Index with Respect to Expenses

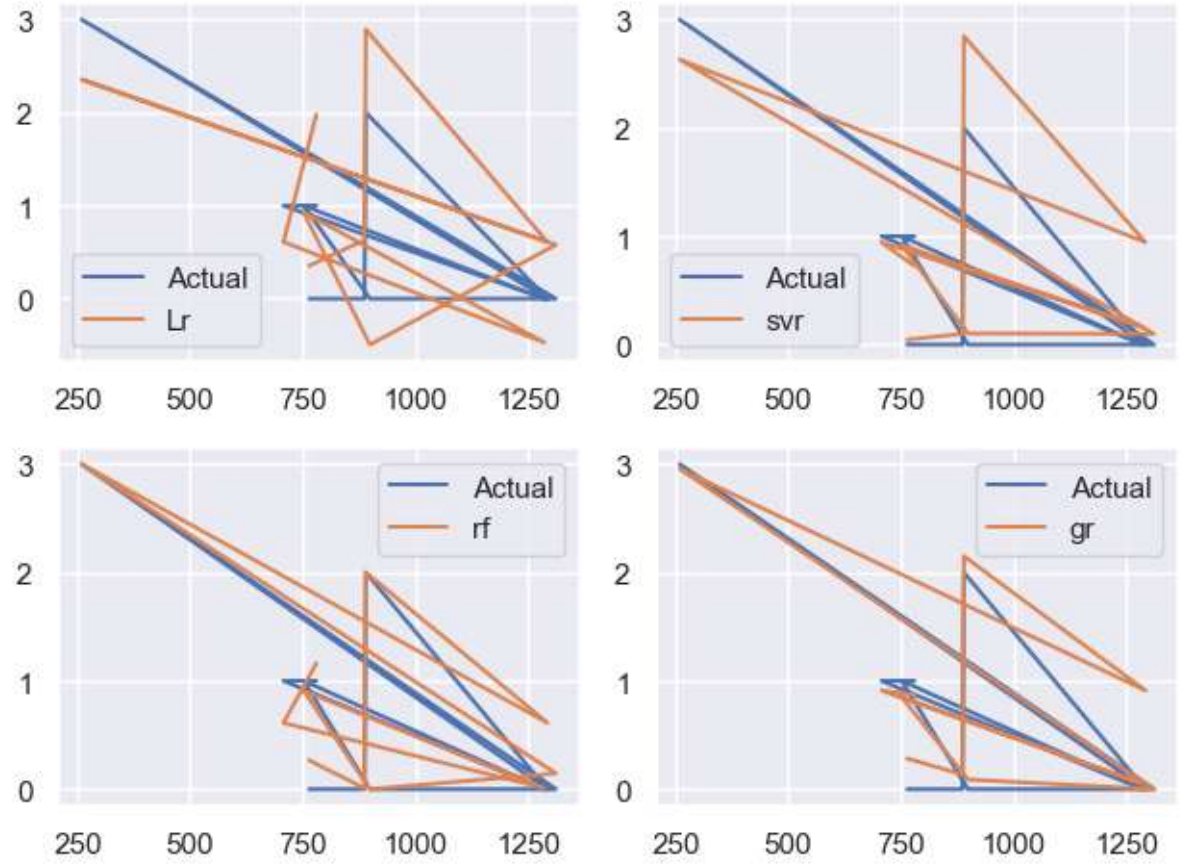


Children with Respect to Expenses



Attributes which had no effect on the prediction were removed from the features.

Training and Evaluating Models:



RESULT:

	Model1	Model2	Model3	Model4
R-squared score	0.688	0.784	0.765	0.812
Mean Absolute Error:	0.523	0.323	0.335	0.320

We see that the accuracy of predicted amount was seen best

i.e. 82.2% in gradient boosting decision tree regression. Other Three regression models also gave good accuracies about 70% In their prediction. Fig 4 shows the best accuracy percentage of all Four models With mean absolute Error 32.02% . Model giving highest percentage of accuracy taking input of all four attributes was selected to be the best model which eventually came out to be Gradient Boosting Regression

CONCLUSION & FUTURE SCOPE :

Background In this project, four regression models are evaluated for individual health insurance data. The health insurance data was used to develop the four regression models, and the predicted premiums from these models were compared with actual premiums to compare the accuracies of these models. It has been found that Gradient Boosting Regression model which is built upon decision tree of Random Forest is the best performing model.

Various factors were used and their effect on predicted amount was examined. It was observed that a persons age and smoking status affects the prediction most in every algorithm applied. Attributes which had no effect on the prediction were removed from the features.

The effect of various independent variables on the premium amount was also checked. The attributes also in combination were checked for better accuracy results.

Premium amount prediction focuses on persons own health rather than other companys insurance terms and conditions. The models can be applied to the data collected in coming years to predict the premium. This can help not only people but also insurance companies to work in tandem for better and more health centric insurance amount

References:

1.Academic Journals and Research Papers:

https://www.ijert.org/health-insurance-amount-prediction

Nidhi Bhardwaj and Rishabh Anand "Health Insurance Premium Prediction Using Machine Learning Algorithms," in 2020.

2.Books:

- "Machine Learning for Healthcare Analytics Projects" by N. L. S. Samarasimha Reddy and B. M. Baveja.
- "Healthcare Analytics: From Data to Knowledge to Healthcare Improvement" by Laura Shang and Hui Yang.

3.Online Courses:

Consider online courses on platforms like Coursera , Udacity that focus on healthcare analytics, machine learning, or predictive modeling.

4.GitHub Repositories:

Explore GitHub repositories related to health insurance premium prediction or healthcare analytics. These repositories may contain code, models, and insights shared by researchers and practitioners.