INSURANCE COST OPTIMIZATION:

A PREDICTIVE AND PRESCRIPTIVE STATISTICAL ANALYSIS

A PROJECT BY

TUSHAR DEEPAK KSHIRSAGAR

M.Sc. (Statistics)

EXL-Certified Data Analyst

INDEX

ABSTRACT

The “Insurance Cost Optimization: A Predictive and Prescriptive Statistical Analysis” project is a comprehensive exploration of the multifaceted domain of insurance charges, blending predictive and prescriptive analytics to revolutionize the insurance industry's pricing strategies. In an era of increasing complexity and data availability, understanding the factors influencing insurance charges and providing prescriptive guidance to optimize these charges is of paramount importance.

This project embarks on a data-driven journey, guided by cutting-edge statistical and machine learning techniques, to unlock valuable insights, predict insurance charges with remarkable accuracy, and prescribe strategies for cost optimization. The key elements of this project are as follows:

A diverse and extensive dataset, encompassing policyholder profiles, medical histories, lifestyle choices, and more, is meticulously collected and preprocessed. Data cleaning and transformation techniques are employed to ensure data quality and consistency. Advanced machine learning models, including linear regression, decision trees, and ensemble techniques, are applied to develop predictive models for insurance charges. These models are trained, validated, and fine-tuned to achieve high predictive accuracy, enabling insurers to estimate charges with precision.

In addition to prediction, the project explores prescriptive analytics to provide actionable insights. Optimization algorithms and strategies are employed to recommend adjustments to risk factors and policy features that can lead to reduced insurance charges while maintaining profitability for insurance companies. The best-performing predictive and prescriptive models are deployed as practical tools for insurance companies, providing the means to estimate and optimize insurance charges. These tools can help insurers make data-driven decisions and enhance competitiveness in a dynamic market.

The “Insurance Cost Optimization: A Predictive and Prescriptive Statistical Analysis” project embodies a forward-thinking approach to insurance pricing. By combining the predictive power of machine learning with the strategic insights of prescriptive analytics, this project aims to revolutionize the insurance industry's approach to pricing, making it more equitable, data-driven, and customer-centric. It represents a significant stride towards a future where insurance charges are not only predicted accurately but also optimized intelligently, benefitting both insurance providers and policyholders alike.

KEYWORDS: Random Forest, Decision tree, K-nearest neighbours, Prescriptive Analysis.

INTRODUCTION

In the contemporary landscape of insurance, the dynamics of premiums and charges have witnessed a transformation. The intersection of healthcare, lifestyle, and financial factors has introduced unprecedented complexity into the calculation of insurance charges. This transformation necessitates the development of advanced statistical methodologies that not only predict insurance charges with precision but also offer actionable insights to both policyholders and insurance providers. It is in this context that the project titled "Predictive and Prescriptive Statistical Analysis of Insurance Charges Prediction" emerges as a beacon of innovation and informed decision-making.

The pivotal role of insurance in safeguarding individuals and families against unforeseen financial burdens cannot be overstated. Whether it pertains to health, life, auto, or property, insurance charges often serve as a pivotal determinant for coverage decisions. The understanding and accurate estimation of these charges not only provide clarity but also enable sound financial planning.

The project is designed as a holistic exploration of the complex interplay between insurance charges and multifaceted policyholder attributes. It leverages the power of predictive and prescriptive analytics, delving deep into historical data and harnessing the capabilities of advanced statistical and machine learning techniques. The core objective is twofold: to develop predictive models that can estimate insurance charges with an unprecedented degree of accuracy and to prescribe strategic insights that empower stakeholders to optimize their insurance coverage.

The “Insurance Cost Optimization: A Predictive and Prescriptive Statistical Analysis” project stands as a testament to the power of data-driven decision-making. By fusing the capabilities of predictive and prescriptive analytics, it aims to equip policyholders with greater financial control and insurance providers with the tools to offer fairer, more precise, and customizable insurance solutions. In an age of evolving insurance landscapes, this project serves as a pioneering effort to reshape the understanding and application of insurance charges.

MOTIVATION

The motivation behind undertaking the project on “Insurance Cost Optimization: A Predictive and Prescriptive Statistical Analysis” is driven by a convergence of crucial factors that address the needs and challenges within the insurance industry. Insurance charges significantly impact the financial well-being of policyholders. The lack of transparency in pricing models can lead to skepticism and confusion. This project aims to bring transparency and fairness to the forefront by developing predictive models that provide clear and accurate estimates of insurance charges.

In today's complex insurance landscape, individuals and businesses are presented with a multitude of coverage options. Empowering policyholders with the ability to make informed decisions about their insurance coverage is paramount. Accurate predictions of future insurance costs enable effective financial planning. Insurance providers face the challenge of pricing policies competitively while effectively managing risks and costs. Predictive analytics can aid insurance companies in fine-tuning their pricing strategies, improving underwriting, and ultimately enhancing their operational efficiency.

The insurance industry is not static. Evolving healthcare systems, demographic shifts, and economic fluctuations necessitate flexible and data-driven approaches to pricing and risk assessment. This project seeks to create adaptable models that respond to these ever-changing dynamics. By providing not only predictive insights but also prescriptive recommendations, this project aims to empower policyholders and insurance providers. The goal is to offer actionable tools for cost optimization, policy adjustment, and well-informed decision-making, fostering a more collaborative and data-driven insurance ecosystem.

In conclusion, this project is motivated by the desire to contribute to a more equitable, transparent, and data-informed insurance landscape. Through advanced statistical analysis, we aspire to promote fair pricing, financial prudence, and trust within the insurance industry, ultimately benefiting both policyholders and insurance providers.

OBJECTIVES

- To propose an innovative machine learning model to predict insurance charges using predictive and prescriptive analysis.
- To determine the factors like Age, Gender, BMI, Children, Smoker, Region etc. which affects insurance charges.
- To find the best fit model on the basis of R-squared, Testing R-squared and Mean Squared Error (MSE) among various other models.

METHODOLOGY

1.Data Collection:

- Initiated the project by acquiring the complex dataset from the designated insurance company. Ensured that data sources are reliable, and the dataset encompasses a wide range of variables, including customer demographics, personal information.

2.Data Preprocessing:

- Began with data preprocessing to enhance data quality and suitability for analysis. This phase includes addressing missing data, handling outliers, and removing duplicate records. Standardize data formats, encode categorical variables, and ensure consistency in data representations.

3.Exploratory Data Analysis (EDA):

- Conducted an exploratory data analysis to gain initial insights into the dataset's characteristics. Generated summary statistics, visualize data distributions (e.g., histograms, box plots), and create scatter plots or heatmaps to identify potential correlations and trends.

4.Descriptive Statistical Analysis:

- Performed a detailed descriptive statistical analysis to further understand the dataset. Calculated central tendency measures (mean, median), measures of dispersion (standard deviation, range), and correlation coefficients. Summarize findings in tables and charts for clarity.

5.Diagnostic Statistical Analysis:

- Diagnosed potential issues in the dataset that might affect subsequent analyses. Used correlation plot matrices to detect multicollinearity among variables and regression plots to identify heteroscedasticity or non-linear relationships.

6.Statistical Testing:

Employed two specific statistical tests to assess the data:

- Shapiro-Wilk's Test: Evaluated the normality of data distributions to determine the applicability of parametric statistical tests.
- Chi-squared Test: Assessed the dependency of attributes, particularly in categorical data, to identify significant relationships.

7.Predictive Statistical Model (Machine Learning):

- Developed a predictive statistical model using machine learning techniques. Choosed appropriate algorithms, such as linear regression, decision trees, random forest, based on the project's objectives. Split the dataset into training and testing subsets for model validation.

8.Model Comparison:

Trained and evaluated multiple predictive models, assessing their performance using key metrics, including:

- R-squared: Measured the proportion of variance explained by the model.
- Adjusted R-squared: Accounted for the number of predictors in the model.
- Testing R-squared: Measured the model's accuracy on test set.
- Mean Squared Error (MSE): Quantified the model's predictive accuracy.

9.Best-Fitted Model Selection:

- Selected the best-fitted model based on the model comparison results. Choosed the model that demonstrates the highest predictive accuracy and aligns with the project's goals.

10.Prescriptive Statistical Analysis:

- Utilized the selected predictive model to conduct prescriptive analysis. Generated actionable recommendations and strategies based on the model's predictions. These recommendations aim to optimize operational efficiency, enhance customer satisfaction, or drive revenue growth for the insurance company.

11.Documentation and Reporting:

- Compiled a comprehensive project report that documents all phases of the analysis. Included detailed explanations of the methodology, data preprocessing steps, EDA findings, descriptive and diagnostic statistical analyses, results of statistical tests, model comparison details, and prescriptive recommendations. Utilized visualizations, tables, and charts to aid in data interpretation and presentation.

Throughout the project, utilized statistical software tools like Python, or specialized machine learning libraries to execute the analysis and generate visualizations. This methodical approach ensures a rigorous analysis of the complex insurance company dataset, delivering valuable insights and actionable recommendations to address specific business challenges.

MATERIALS AND METHODS

Software, Libraries and Packages Used:

Google Colab	1) Predictive and Prescriptive analysis and Exploratory Data Analysis (EDA) is performed using Python language. 2) Basically, Google Colab is an application which provides us with an environment where we can write and execute python code. 3) Importing data in python and performing data pre-processing methods on it, so that data becomes ready for analysis and we get more accurate analysis. 4) After performing pre-processing on the data, we can also implement different machine learning algorithms in order to classify the sentiments. 5) And finally, Google Colab will also make it easier to visualise and analyse the data. 6) Code scripts, research-related text, visualisation plots and graphs, machine learning models, and other materials may all be kept in one Notebook or document and shared effortlessly across several platforms. 7) These are all important uses of the Google Colab and these are the reasons behind choosing Google Colab .
Pandas	The data which was collected from the Hospitals was stored in CSV files or formats. So, in order to work with CSV files and import them in the python we will require a Special library of python which is PANDAS LIBRARY. So, basically when we want to work with the datasets, we will be using Pandas library. It is not just used to import the CSV files but there are many other uses such as: I. This library can clean up unreadable and irrelevant data sets. Relevant data is most important thing for a Data Analyst. II. This library can be used to answer some basic questions about data such as, • If there are two columns, so is there any correlation between these two columns? • What is an average value of a particular column? • What is the minimum or maximum value that is occurring in a dataset?
Sklearn	Sklearn basically stands for Scikit Learn Library. As we are aware of the fact that there are many Machine Learning Libraries used in Python. Scikit learn is one of them, in fact this library is one the best known. This library is responsible to support both of the machine learning approaches, Supervised and Unsupervised. This library also provides

	many different algorithms for the classification, dimensionality reduction, clustering and regression purposes. The combination of two libraries i.e., NumPy and SciPy is Sklearn library. Additionally, it functions nicely with other libraries like Pandas and Seaborn. We can use this library for pre-processing such as feature encoding, feature extraction and we can also use this library to split the data into train and test, then we can use this library to implement different machine learning models and finally in order to check the accuracy of these models we can use this library.
Pickle	The pickle module implements binary protocols for serializing and de-serializing a Python object structure. *"***Pickling***"* is the process whereby a Python object hierarchy is converted into a byte stream and *"***unpickling***"* is the inverse operation, whereby a byte stream (from a binary file or bytes-like object) is converted back into an object hierarchy. Pickling is alternatively known as "serialization", "**marshalling**," 1 or "flattening"; however, to avoid confusion, the terms used here are "**pickling**" and "**unpickling**".

STATISTICAL TERMS

LINEAR REGRESSION - In machine learning, linear regression is a supervised learning algorithm used for predicting a continuous target variable based on one or more independent predictor variables. It's one of the simplest and most widely used algorithms for regression tasks. Linear regression models the relationship between the input features (independent variables) and the output (dependent variable) as a linear equation.

Types of Linear Regression:

- **Simple Linear Regression:** In simple linear regression, there is only one independent variable (predictor) used to predict a single dependent variable. The relationship is modeled as a straight line: y = b0 + b1 * x, where y is the dependent variable, x is the independent variable, and b0 and b1 are coefficients to be learned from the data.

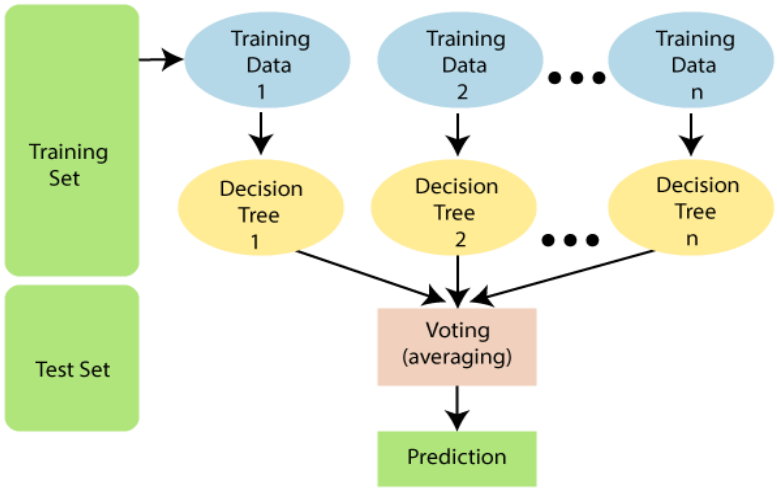
- **Multiple Linear Regression:** In multiple linear regression, there are multiple independent variables (predictors) used to predict a single dependent variable. The relationship is represented as a hyperplane in multi-dimensional space: y = b0 + b1 * x1 + b2 * x2 + ... + bn * xn, where there are n independent variables.

The objective of linear regression is to find the best-fitting line or hyperplane that minimizes the sum of squared differences (residuals) between the predicted values and the actual target values. Linear regression models are evaluated using various metrics, such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared (R2), to measure the quality of predictions.

Linear regression is widely used in fields such as economics, finance, social sciences, and machine learning for tasks like price prediction, demand forecasting, trend analysis, and more. Ridge Regression and Lasso Regression: These are variations of linear regression that include regularization terms to prevent overfitting.

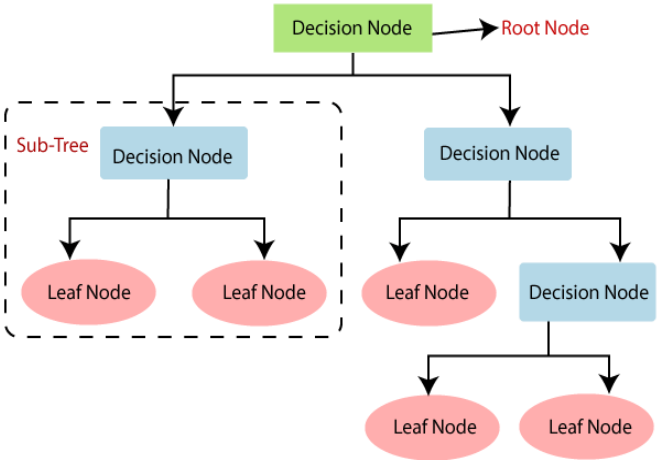
Linear regression is an essential tool in machine learning and statistics. It provides a solid foundation for understanding regression analysis and serves as a baseline for many more advanced regression techniques.

RANDOM FOREST - A random forest is a Machine Learning technique that is used to solve regression and classification problems. It utilizes ensemble learning technique, that combines many classifiers to provide solutions to complex problems. A random forest algorithm consists of many decision trees. The ‘forest’ generated by the random forest algorithm is trained through bagging or bootstrap aggregating. Bagging is an ensemble meta-algorithm that improves the accuracy of machine learning algorithms. The (random forest) algorithm establishes the outcome based on the predictions of the decision trees. It predicts by taking the average or mean of the output from various trees. Increasing the number of trees increases the precision of the outcome.



DECISION TREE- Decision tree is the most powerful and popular tool for classification and prediction. Decision tree is a type of Supervised Machine Learning where data is continuously split according to certain parameter. The tree can be explained by two entities, namely decision nodes and leaves. The leaves are the decisions or final outcomes and the decision nodes are where the data is split.

In its simplest form, a decision tree is a type of flowchart that shows a clear pathway to a decision. In terms of data Analysis, it is a type of algorithm that includes conditional 'control' statements to classify data. A decision tree starts at a single point (or 'node') which then branches (or 'splits') in two or more directions. Each branch offers different possible outcomes, incorporating a variety of decisions and chance events until a final outcome is achieved. When shown visually, their appearance is tree-like…hence the name!



- **Root nodes:**

  In the diagram above, the blue decision node is what we call a 'root node.' This is always the first node in the path. It is the node from which all other decision, chance, and end nodes eventually branch.

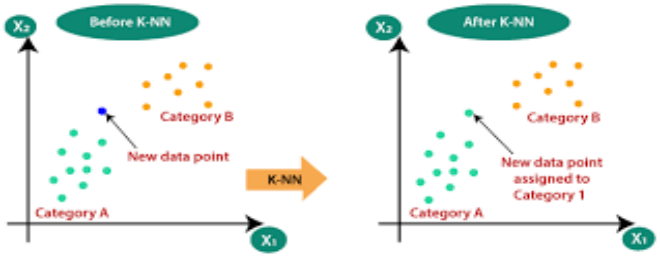
- **Leaf nodes:**

  In the diagram above, the end nodes are what we call ‘leaf nodes.’ These show the end of a decision path (or outcome). You can always identify a leaf node because it doesn’t split, or branch any further. Just like a real leaf!
- **Internal nodes**:

  Between the root node and the leaf nodes, we can have any number of internal nodes. These can include decisions and chance nodes. It’s easy to identify an internal node—each one has branches of its own while also connecting to a previous node.
- **Splitting:**

  Branching or ‘splitting’ is what we call it when any node divides into two or more sub-nodes. These sub-nodes can be another internal node, or they can lead to an outcome (a leaf/ end node.)

KNN- The k-nearest neighbours algorithm, also known as KNN or k-NN, is a non-parametric, supervised learning classifier, which uses proximity to make classifications or predictions about the grouping of an individual data point. It classifies the data point on how its neighbor is classified. KNN classifies the new data points based on the similarity measure of the earlier stored data points.



For example, if we have a dataset of Apple and Bananas. KNN will store similar measures like shape and color. When a new object comes it will check its similarity with the color (red or yellow) and shape. K in KNN represents the number of the nearest neighbors we used to classify new data points.

REPRESENTATION OF DATASET

	Index	Age	Gender	BMI	Children	Smoker	Region	InsuranceCharges
0	0	19	Female	27.900	0	Yes	southwest	16884.92
1	1	18	Male	33.770	1	No	southeast	1725.55
2	2	28	Male	33.000	3	No	southeast	4449.46
3	3	33	Male	22.705	0	No	northwest	21984.47
4	4	32	Male	28.880	0	No	northwest	3866.86
...	...	...	...	...	...	...	...	...
1333	1333	50	Male	30.970	3	No	northwest	10600.55
1334	1334	18	Female	31.920	0	No	northeast	2205.98
1335	1335	18	Female	36.850	0	No	southeast	1629.83
1336	1336	21	Female	25.800	0	No	southwest	2007.95
1337	1337	61	Female	29.070	0	Yes	northwest	29141.36

The Basic Description and understanding of Data

The Statistics of the dataset gives us the basic idea about the spread of the data. Therefore, we need to go for description data set like count, mean, standard deviation, min, max, quartiles, etc. This image shows the data description.

	count	mean	std	min	25%	50%	75%	max
Age	1338.0	39.207025	14.049960	18.00	27.00000	39.00	51.00000	64.00
Gender	1338.0	0.505232	0.500160	0.00	0.00000	1.00	1.00000	1.00
BMI	1338.0	30.663397	6.098187	15.96	26.29625	30.40	34.69375	53.13
Children	1338.0	1.094918	1.205493	0.00	0.00000	1.00	2.00000	5.00
Smoker	1338.0	0.204783	0.403694	0.00	0.00000	0.00	0.00000	1.00
Region	1338.0	1.515695	1.104885	0.00	1.00000	2.00	2.00000	3.00
InsuranceCharges	1338.0	13270.422414	12110.011240	1121.87	4740.28750	9382.03	16639.91500	63770.43

DATA PREPROCESSING

Removal of Unnecessary Columns:

There is no need of column 'Index' in our analysis. So, we will remove it.

Checking for Null Values:

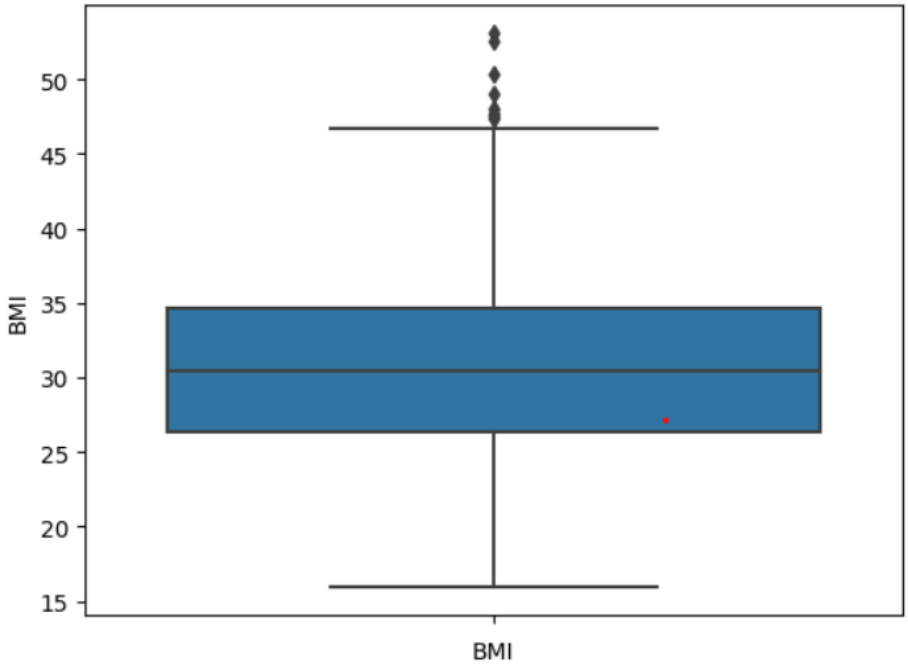
```
Index              0
Age                0
Gender             0
BMI                0
Children           0
Smoker             0
Region             0
InsuranceCharges   0
dtype: int64
```

Interpretation: There are no null values present in the data.

Checking for Duplicate Values:

There is no any duplicate value present in the data.

Checking for Outliers:

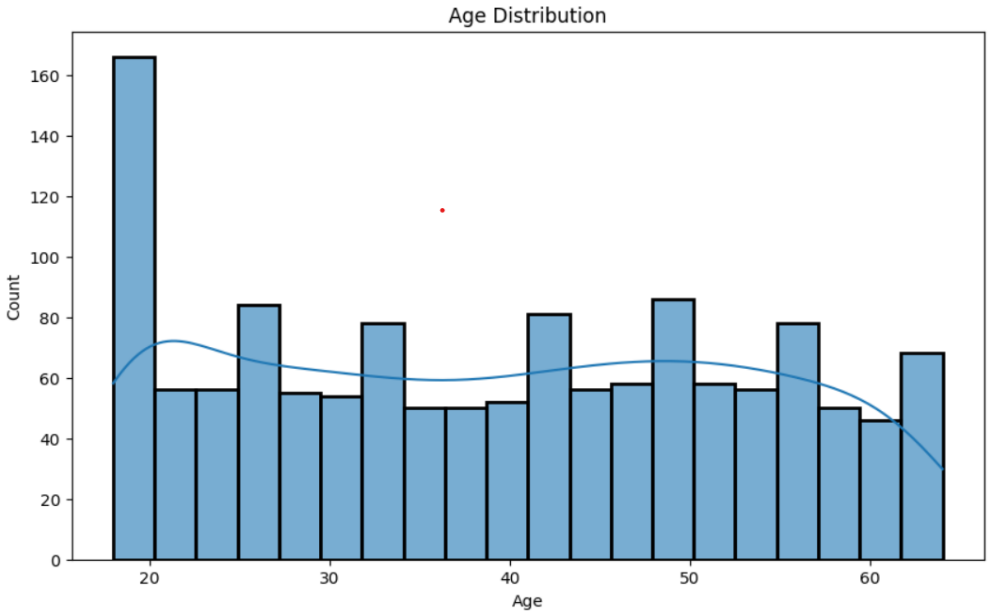


Conclusion: We will not do outlier treatment for BMI, we wish to maintain some realness in the data. Outlier observations are just 20 anyway, thus won't hurt our model.

EXPLORATORY DATA ANALYSIS

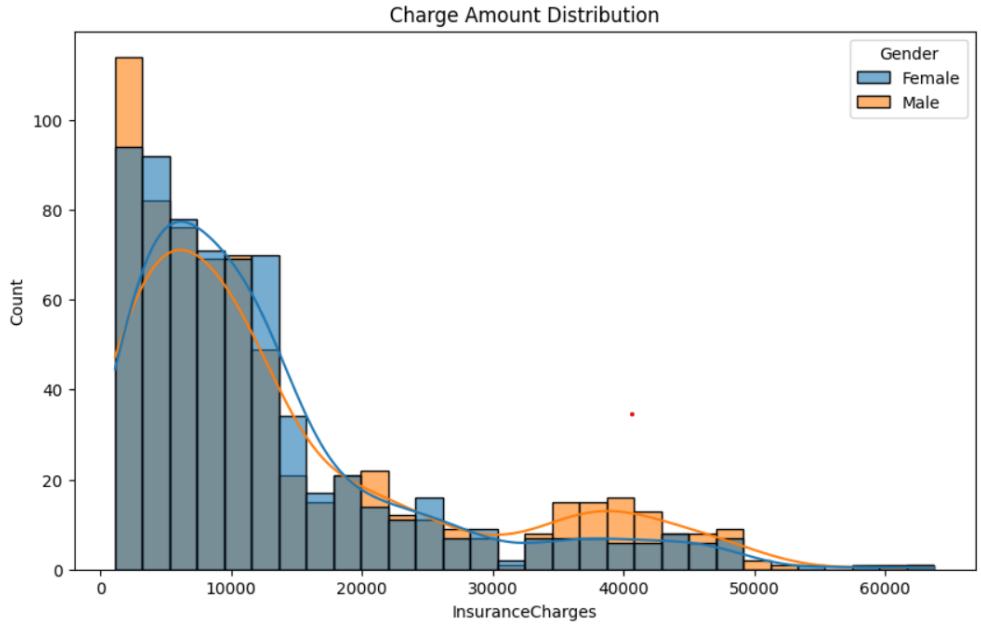
1) DESCRIPTIVE STATISTICAL MODEL:

Histogram of Age Distribution:



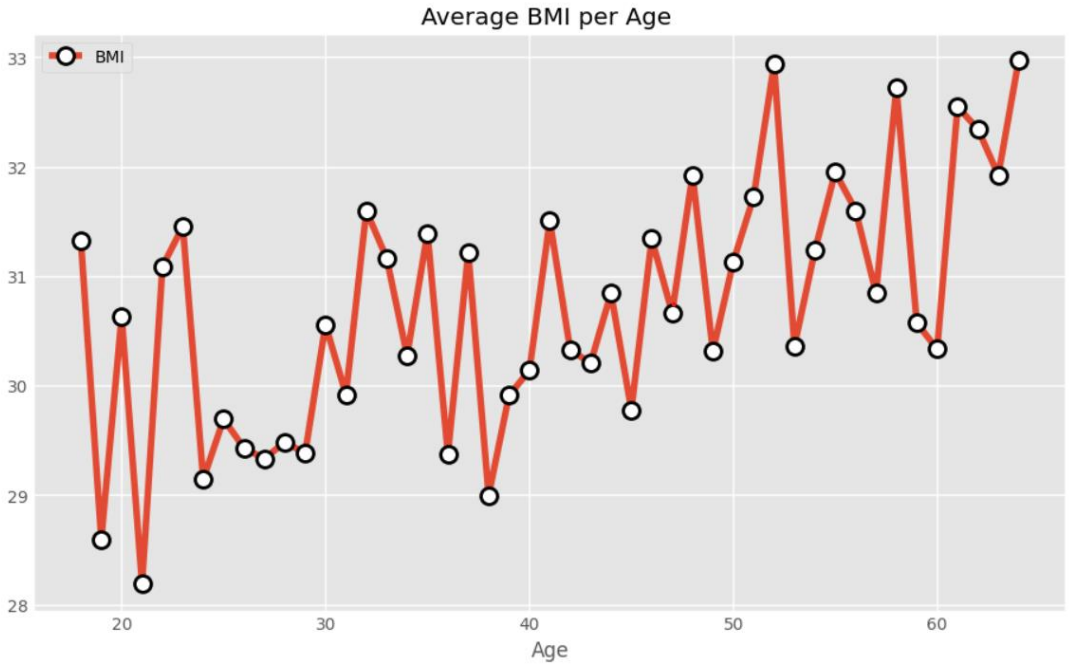
Conclusions: From above histogram we can see that, insurance customers of age 18-20 has higher proportion than others.

Histogram of Gender wise Charge Amount Distribution:



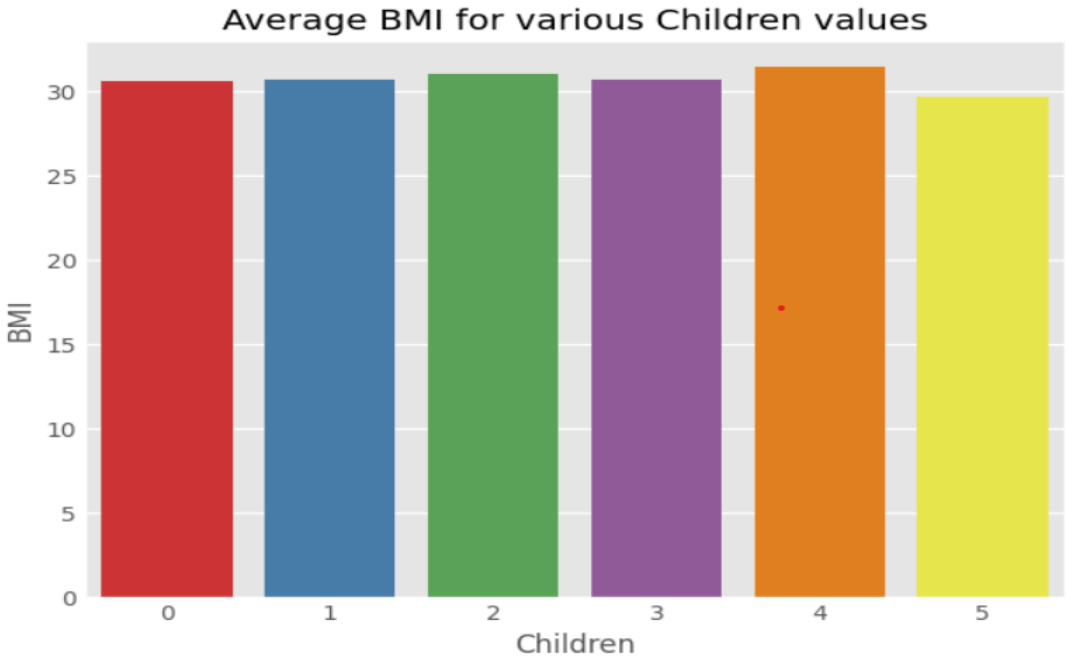
Conclusions: From above histogram we can see that, charge amount for Male and Female is nearly equal but it is higher for Males in a range 2500-17500 and higher for Females in range 30000-50000.

Line Chart of Average BMI per Age:



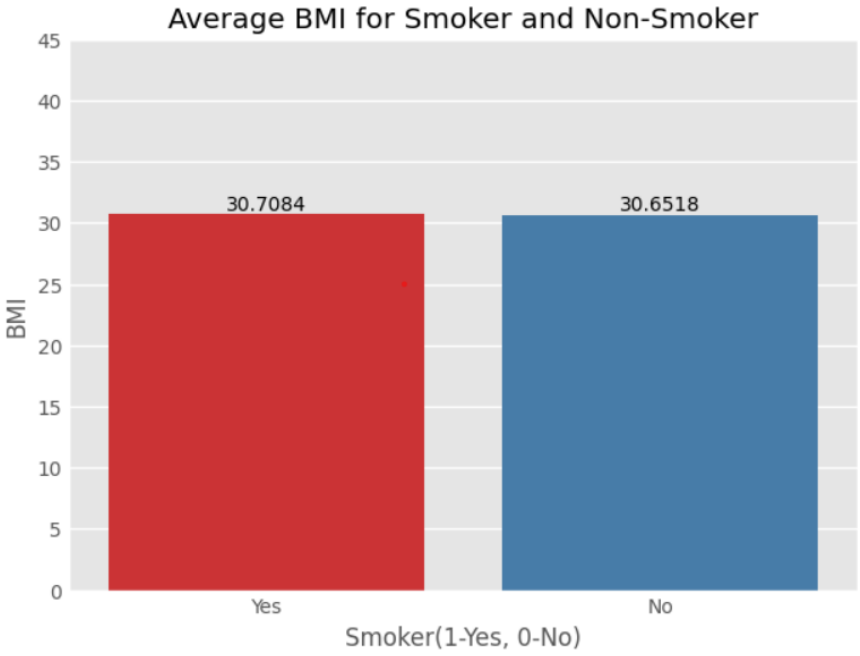
Conclusions: From above line chart we can see that, as age increases the average BMI score starts getting higher to unhealthier ranges.

Bar Plot of Average BMI for Various Children Values:



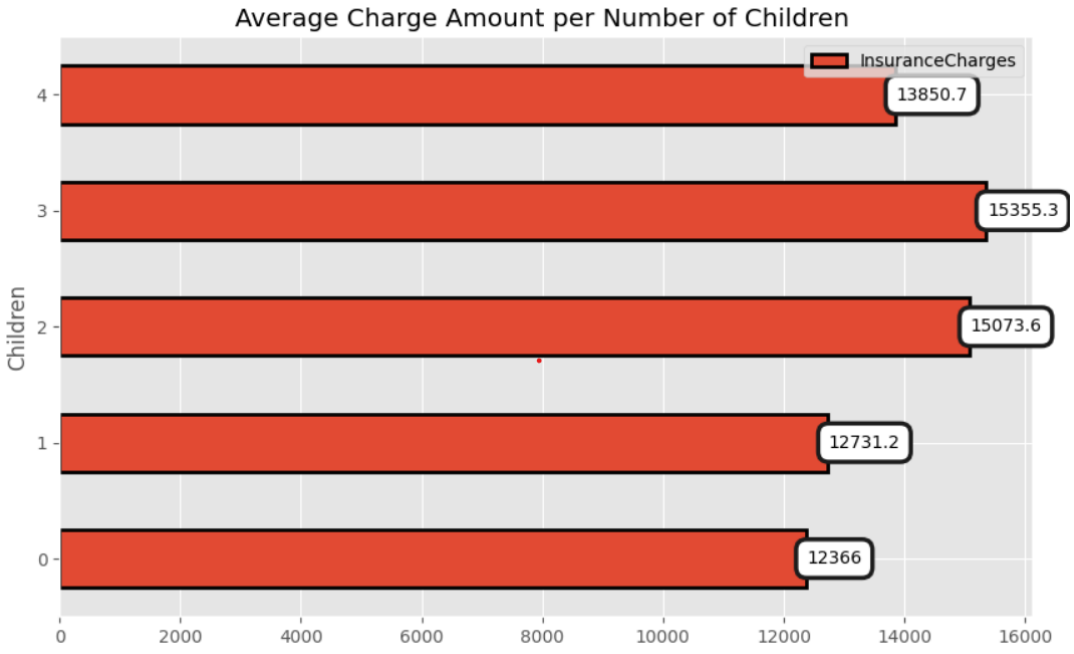
Conclusions: From above bar plot we can see that, average BMI for various children values is nearly same.

Bar Plot of Average BMI for Smoker and Non-Smoker:



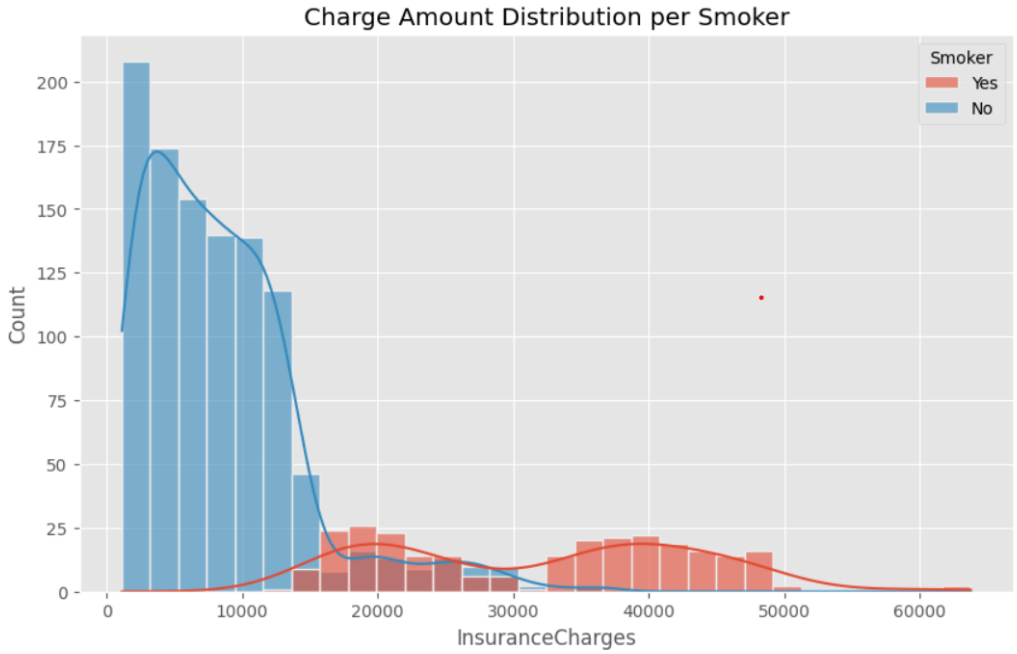
Conclusions: From above bar plot we can see that, average BMI value for Smoker and Non-Smoker is same.

Bar Plot of Average Charge Amount per No. of Children:



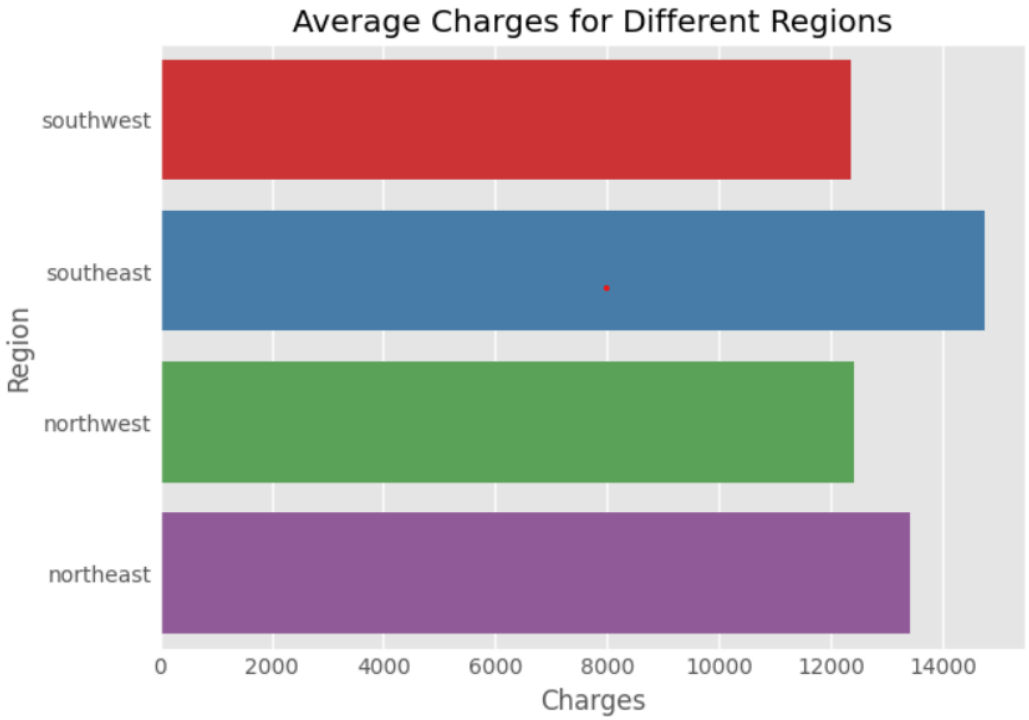
Conclusions: In general, average charge amount increases as no. of children increases.

Histogram of Charge Amount Distribution per Smoker:



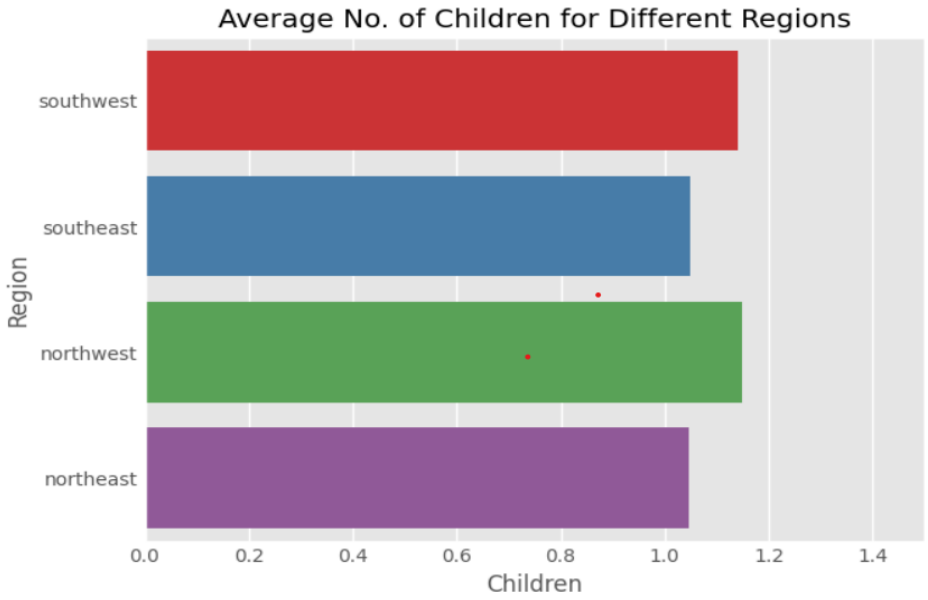
Conclusion: From above histogram we can see that, smoker people are getting charged more than non-smoker.

Bar Plot of Average Charges for Different Regions:



Conclusion: In general, there is a small difference between regions in terms of average charge amount, but southeast region has highest average charge amount among all regions.

Bar Plot of Average No. of Children for Different Regions:



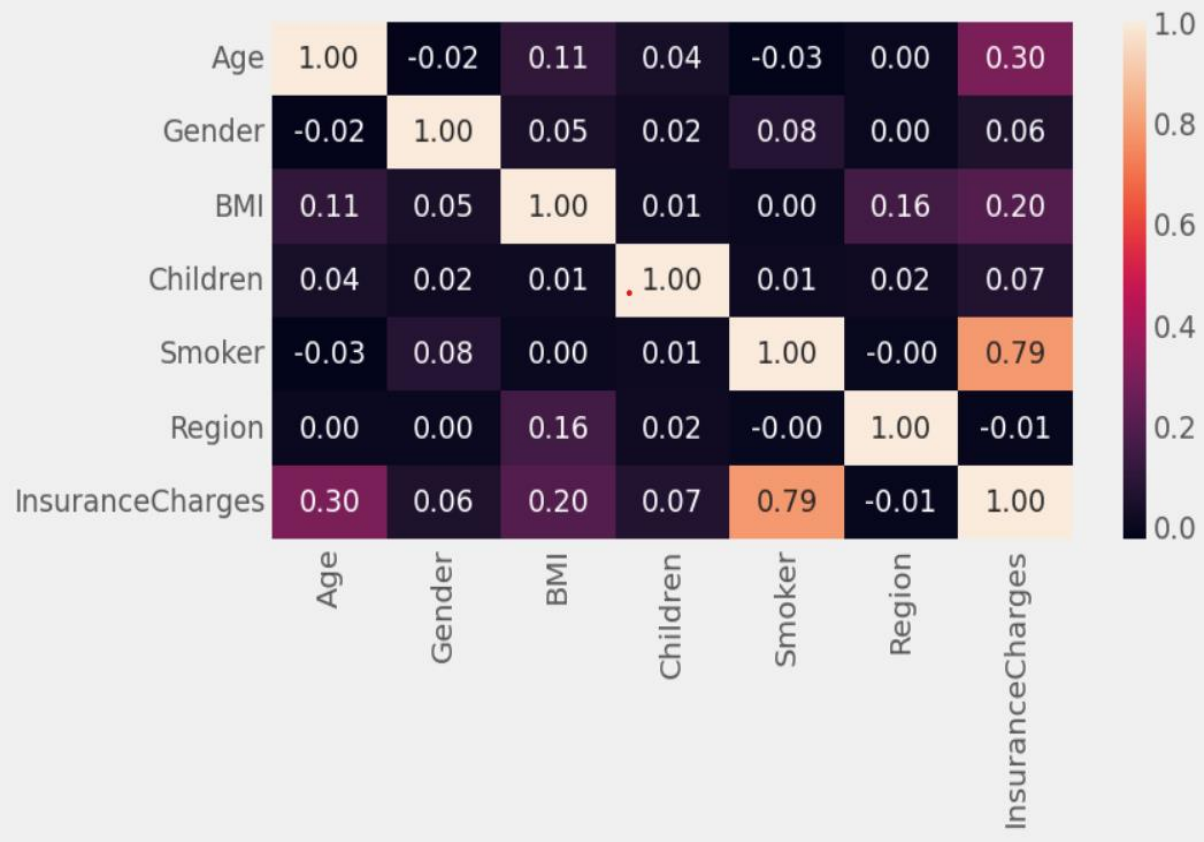
Conclusion: From above bar plot we can see that, average no. of children for different regions is same.

FEATURE ENGINEERING

FEATURES SELECTION:

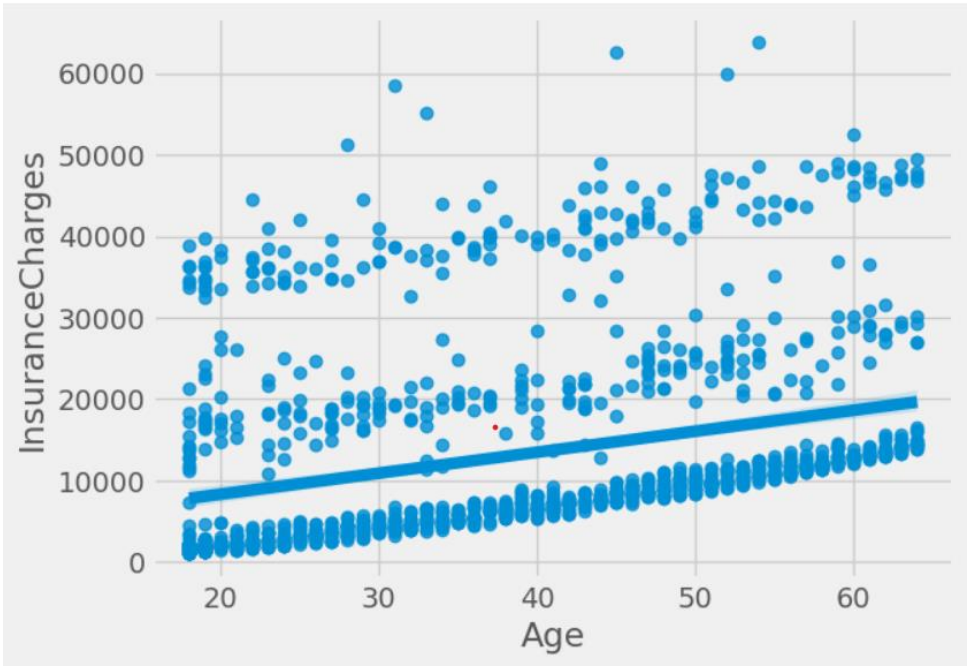
We will use techniques like Correlation Analysis and plots such as Regression Plots, Box Plot for Continous and Categorical Variables respectively for feature selection process.

Correlation Plot:



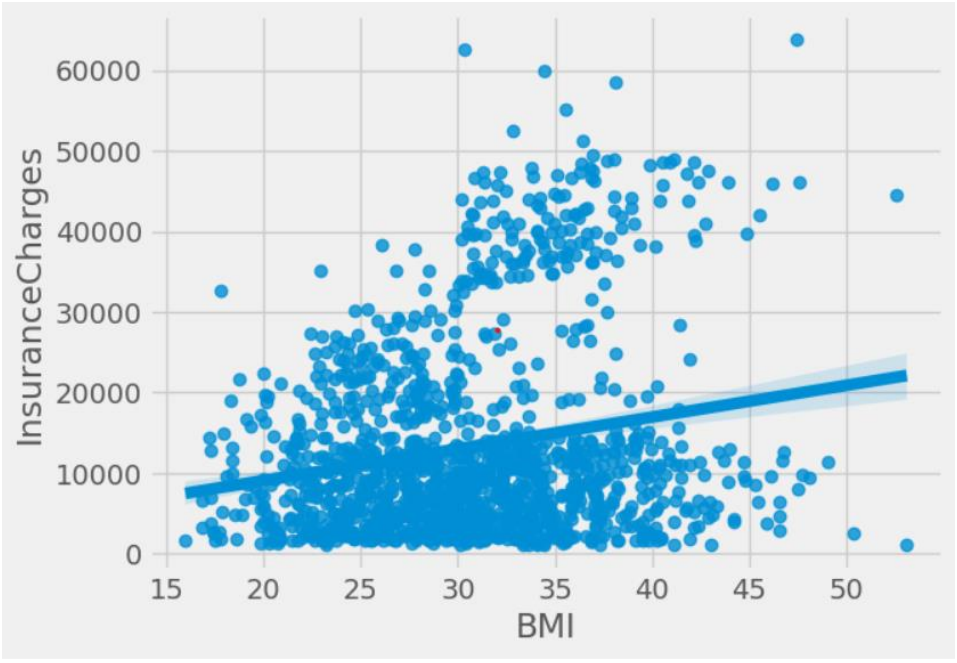
Interpretation: Based on the correlation plot, it is clear that "Age", "BMI" and "Smoker" are the features having higher correlation with Insurance Charges.

Regression Plot of Age:



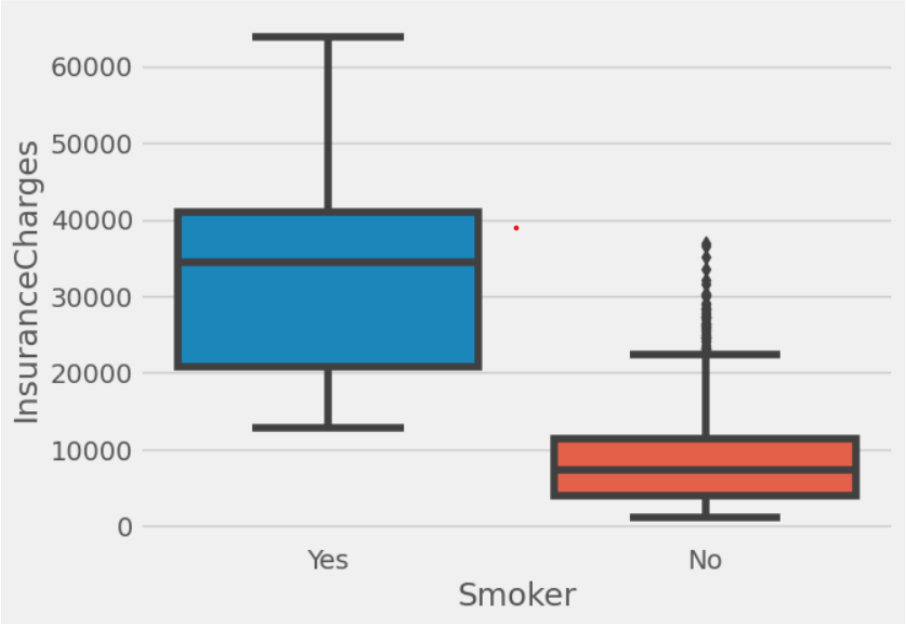
Conclusion: The data points are scattered in increasing manner with age. So, there is strong correlation between Age and Insurance Charges.

Regression Plot of BMI:



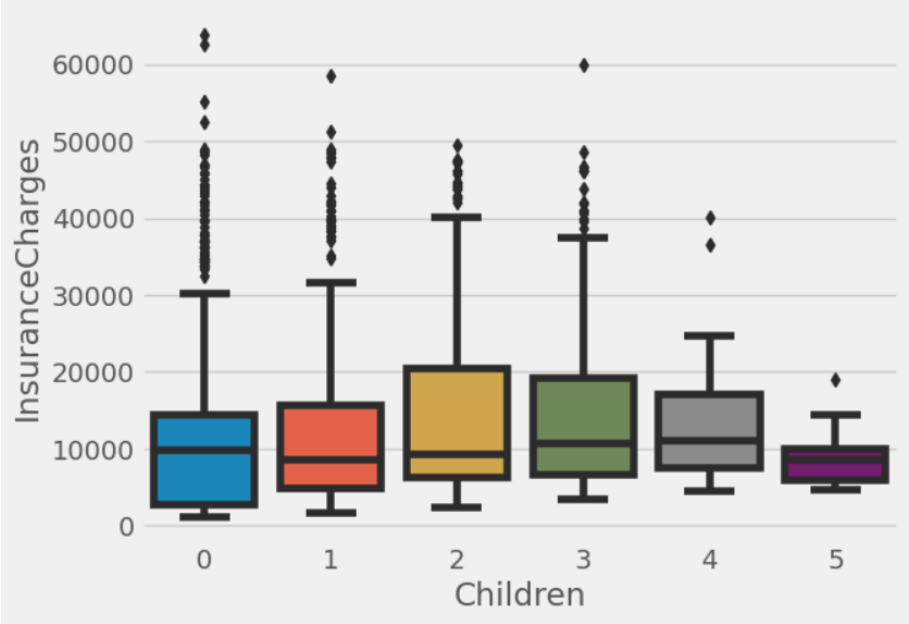
Conclusion: The data points shows somewhat increasing trend with BMI. So, there is considerable correlation between BMI and Insurance Charges.

Box Plot of Smoker:



Conclusion: The classes of feature for which the distribution of Insurance Charges is distinct enough, such feature can be potential good predictor of Insurance Charges.

Box Plot of Children:



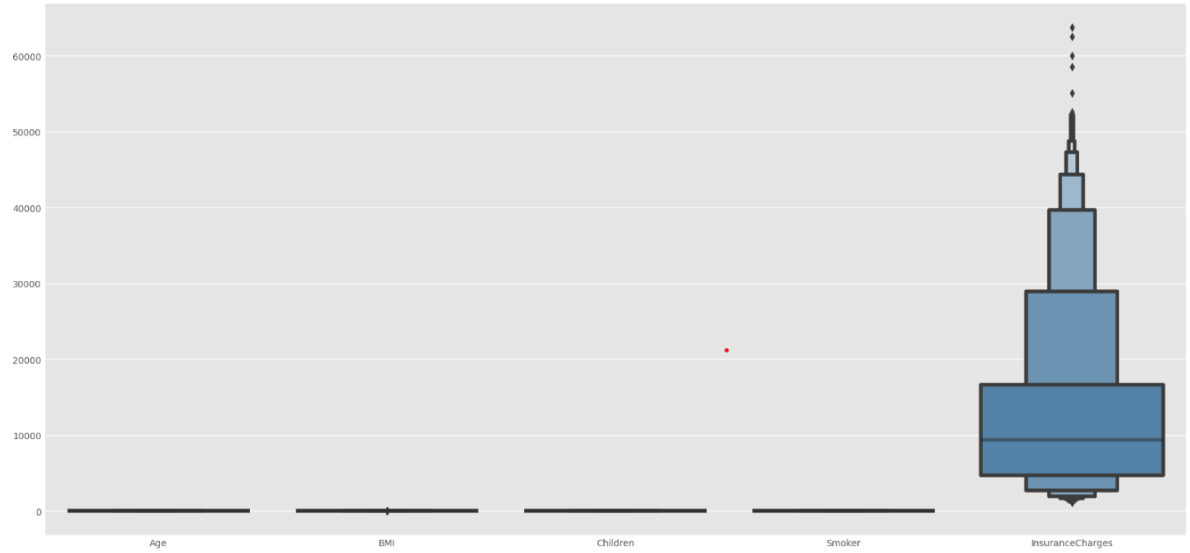
Conclusion: The classes of feature for which the distribution of Insurance Charges is somewhat distinct, such feature could be potential predictor of Insurance Charges.

CONCLUSION:

From above Correlation Analysis, Regression Plots and Box Plots we can conclude that, the variables ‘Age’, ‘BMI’, ‘Children’ and ‘Smoker’ can be good predictors for Insurance Charges. So, we will choose these variables for model building.

FEATURE SCALING:

Plotting the Features using Boxen plot:



Interpretation:

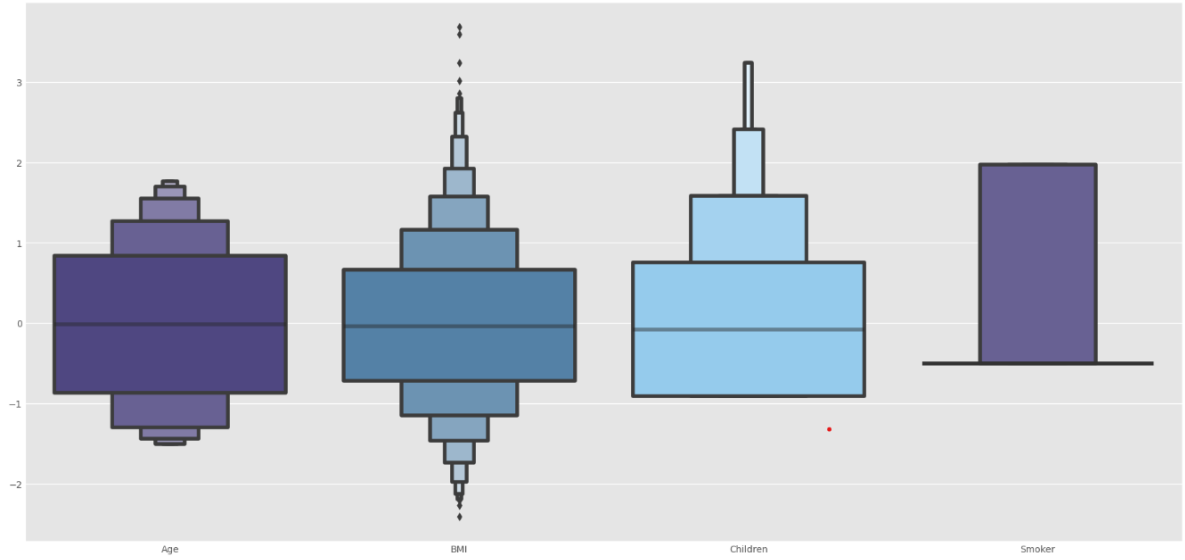
The above plot shows the range of features in the data. All the features are in different ranges.

To fit this in a model we must scale it to the same range.

Scaling the Features by setting up a Standard Scaler for the Features:

	Age	BMI	Children	Smoker
0	-1.438764	-0.453320	-0.908614	1.970587
1	-1.509965	0.509621	-0.078767	-0.507463
2	-0.797954	0.383307	1.580926	-0.507463
3	-0.441948	-1.305531	-0.908614	-0.507463
4	-0.513149	-0.292556	-0.908614	-0.507463
...	...	...	...	...
1333	0.768473	0.050297	1.580926	-0.507463
1334	-1.509965	0.206139	-0.908614	-0.507463
1335	-1.509965	1.014878	-0.908614	-0.507463
1336	-1.296362	-0.797813	-0.908614	-0.507463
1337	1.551686	-0.261388	-0.908614	1.970587

Plotting the Scaled Features using Boxen Plots:



Conclusion: The plot clearly indicates that, now all the features are in same range since we have scaled the data.

2) DIAGNOSTIC STATISTICAL MODEL:

Age	BMI	Children	Smoker	InsuranceCharges
19	27.900	0	1	16884.92
18	33.770	1	0	1725.55
28	33.000	3	0	4449.46
33	22.705	0	0	21984.47
32	28.880	0	0	3866.86

The above data consist of some important variables and their Description is explained below:

Age: It is an age (in years) of insurance customer whose record is considered for insurance.

BMI: In this feature there are BMI's of insurance customers whose records is considered for insurance.

Children: In this feature there are six classes on the basis of insurance customer whose record is considered for insurance having how many number of children.

No Child: - "0"

Single Child: - "1"

Two Childs: - "2"

Three Childs: - "3"

Four Childs: - "4"

Five Childs: - "5"

Smoker: In this feature there are two classes on the basis of whether insurance customer is smoker or not.

No: - "0"

Yes: - "1"

Insurance Charges: Insurance Charges for the customer in Rupees.

STATISTICAL ANALYSIS

SHAPIRO-WILK TEST FOR NORMALITY:

Hypothesis to be tested,

H_0: The data is normal.

v/s

H_1: The data is not normal.

```
ShapiroResult(statistic=0.4068381190299988, pvalue=0.0)
```

p-value: 0.0

Decision Criteria: p-value = 0.0 < 0.05

Decision:

We reject H_0 at 5% level of significance.

Conclusion: The data is not normal. We will go for non-parametric tests.

CHI-SQUARED TEST:

Here we are testing the dependency of one attribute on the other attribute.

1) Smoker and Gender

H_0: Smoker and Gender are independent.

v/s

H_1: Smoker and Gender are dependent.

Contingency table:

Smoker	No	Yes
Gender		
Female	547	115
Male	517	159

```
Chi-2 Statistic : 7.765921028604451
Degrees of Freedom : 1
p-value : 0.005324114164320548
```

Decision Criteria: p-value < 0.05

Decision:

We reject H_0 at 5% level of significance.

Conclusion:

Smoker and Gender are dependent.

2) Smoker and Children

H_0: Smoker and Children are independent.

v/s

H_1: Smoker and Children are dependent.

Contingency table:

Smoker	No	Yes
Children		
0	459	115
1	263	61
2	185	55
3	118	39
4	22	3
5	17	1

```
Chi-2 Statistic : 6.88771990494763
Degrees of Freedom : 4
p-value : 0.14194202277971024
```

Decision Criteria: p-value > 0.05

Decision:

We accept H_0 at 5% level of significance.

Conclusion:

Smoker and Children are independent.

3) Smoker and Region

H_0: Smoker and Region are independent.

v/s

H_1: Smoker and Region are dependent.

Contingency table:

Smoker	No	Yes
Region		
northeast	257	67
northwest	267	58
southeast	273	91
southwest	267	58

Chi-2 Statistic : 7.34347776140707
Degrees of Freedom : 3
p-value : 0.06171954839170546

Decision Criteria: p-value > 0.05

Decision:

We accept H_0 at 5% level of significance.

Conclusion:

Smoker and Region are independent.

CONCLUSION:

From the above chi-squared test, we can conclude that, attribute Gender has an association with attribute Smoker which is very important parameter in our analysis, but attributes Children and Region has no association with attribute Smoker.

3) PREDICTIVE STATISTICAL MODEL

Splitting Train & Test Set:

The whole dataset is divided into two parts, one is training data and the other is testing data. Taking randomly 85% dataset as training dataset and 15% dataset as testing dataset.

Model Building:

Various Machine Learning models are built on training dataset. The Test dataset is used further for evaluation of model performance.

MODEL SELECTION:

- **Linear Regression -**

  Output: Testing R-squared of Linear Regression is **0.7936.**

- **Decision Tree Regressor -**

  Output: Testing R-squared of Decision Tree Regressor is **0.6853.**

- **Random Forest Regressor -**

  Output: Testing R-squared of Random Forest Regressor is **0.8458.**

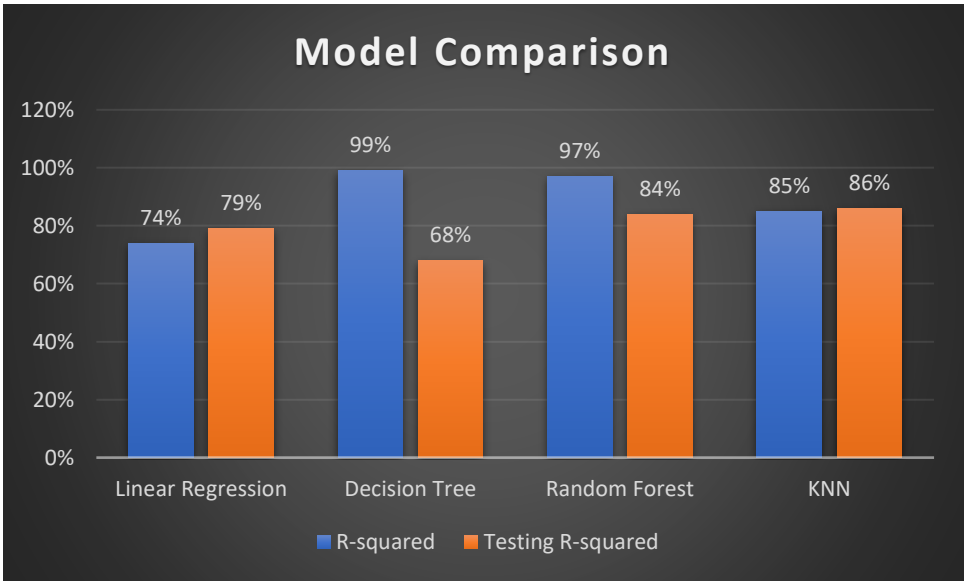
- **K-Neighbours Regressor**:

  Output: Testing R-squared of K-Neighbours Regressor is **0.8688.**

Conclusion: K-Neighbours Regressor has greater Testing R-squared than the other models. So, K-Neighbours Regressor does best amongst the models to be the most accurate. We have t o build a model which should be generalised well to new, unseen data.
Hence, we will fit Hypertuned K-Neighbours Regressor Model for predictions.

MODEL COMPARISON

Model	R-squared	Adjusted R-squared	Testing R-squared	RMSE
Linear Regression	74%	74%	79%	5605.02
Decision Tree	99%	99%	68%	6921.95
Random Forest	97%	97%	84%	4844.21
KNN	85%	85%	86%	4468.36



Interpretation:

The above graph also shows that K-Neighbours Regressor Model performs best among all the other models, with excellent R-squared and Testing R-squared.

K-NEIGHBOURS REGRESSOR
(Hypertuned)

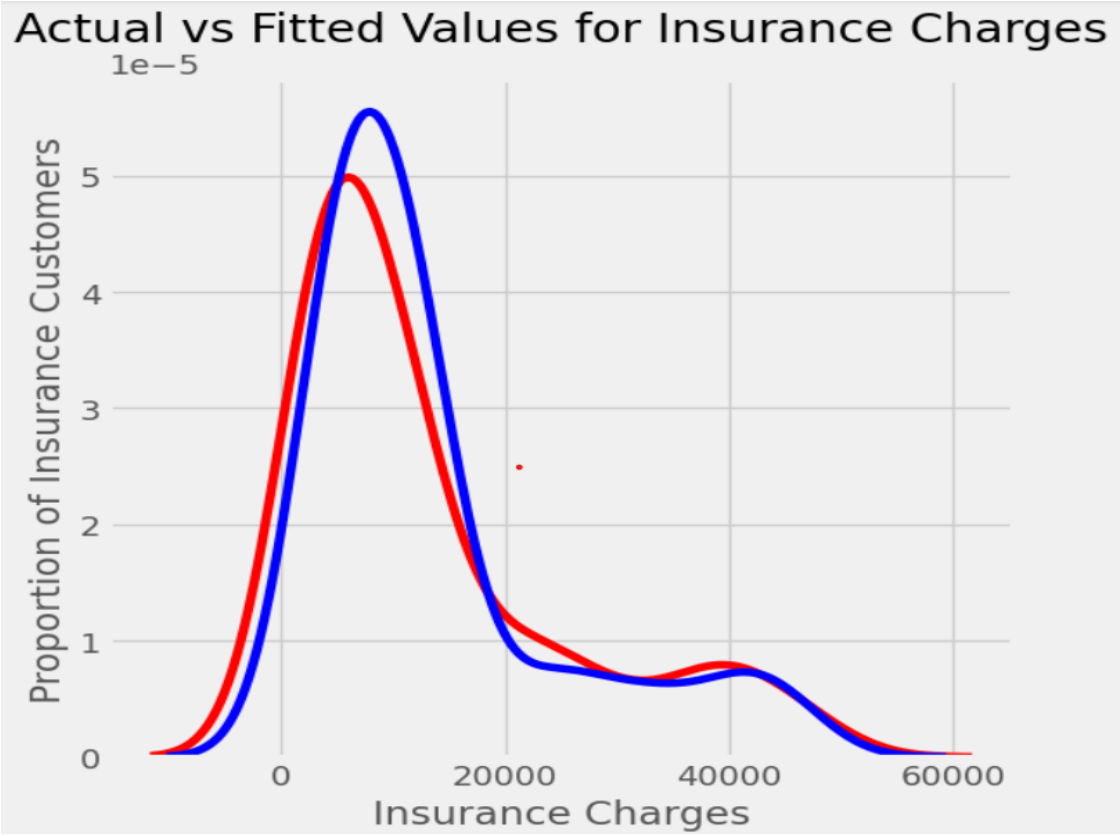
Model Summary:

```
********* K-Neighbours Regressor *********
R-squared              :  0.856649598292555
Adjusted R-squared     :  0.8561430597706206
Testing R-squared      :  0.8649116516286439
RMSE                   :  4535.371660958468
```

Interpretation:

- From the above R-squared, we can say that 85% of variation observed in target feature is explained by the model.
- From the above Testing R-squared, we can say that testing accuracy of fitted model is 86%.

K-Neighbours Regressor Model works as a Generalised Model:



Conclusion: From the above graph we can see that Actual and Fitted values for Insurance charges are significantly overlapped. Hence K-Neighbours Regressor Model works as Generalised Model.

4) PRESCRIPTIVE STATISTICAL ANALYSIS:

How does this model work?

To build this model we have used machine learning technique called input method. Model works on K-Neighbours Regressor. The model contains insurance customer's information like Age, Body Mass Index (BMI), Number of Children, Smoking Status etc. We have to provide these inputs to the model, and the output will show how much insurance charges would have such a customer. Output is in terms of Rupees. So, the Insurance Company and Hospital Management will get an idea about how much charges would be necessary to insure them.

Demonstration:

Suppose a case of a particular insurance customer having information as below:

Age: 42

BMI: 28.880

No. of Children: Single Child "1"

Smoking Status: Yes "1"

The user will provide these inputs to the machine learning model, then model will predict how much insurance charges would have such a customer.

Model will predict Insurance Charges for customer in terms of Rupees.

By taking output given by the model into consideration the Insurance Company and Hospital Management will get an idea about how much charges would be necessary to insure them. This prescriptive model will help Insurance Company and Hospital Management to avoid improper management, will also help to control over managing, allotting resources and save time and money.

Model works as given below:

```
Age :42
BMI :28.880
No. of Children :1
Smoking Status :1
Customer will have Insurance Charges Rs. 16443
```

From above output we can see that customer having information as **Age: 42, BMI: 28.880, No. of Children: Single Child "1", Smoking Status: Yes "1"** then Model Predicted Insurance Charges for customer to insure him is Rs.16443, and from this result the Insurance Company and Hospital Management can manage their resources which helps to save time and money.

CONCLUSION

- From the Bar Plot of Average Charge Amount per number of Children, it is clear that average charge amount increases as number of children increases.

- From the Histogram of Charge Amount Distribution per Smoker, it seems that smoker people are getting charged more than non-smoker.

- Based on the correlation plot, it is clear that "Age", "BMI" and "Smoker" are the features having higher correlation with Insurance Charges.

- From the Chi-squared test for dependence of different attributes, we get, attribute Gender has an association with attribute Smoker which is the most impactful feature in the analysis, but attributes Children and Region has no association with attribute Smoker.

- K-Neighbours Regressor has greater accuracy than the other models and it does best amongst the models to be the most accurate.

- The Testing R-squared of K-Neighbours Regressor Model is 0.86.

DISCUSSION

The project on "Insurance Cost Optimization: A Predictive and Prescriptive Statistical Analysis" holds substantial promise for reshaping the insurance landscape. By focusing on transparency and fairness, the project addresses a persistent concern by providing accurate and comprehensible estimates of insurance charges. This empowers policyholders to make well-informed decisions and plan their finances effectively. Additionally, insurance providers benefit from more precise risk assessment and pricing strategies, contributing to their competitiveness and efficiency. The adaptability of the models to dynamic variables and strict adherence to data privacy regulations underscore ethical and regulatory considerations. Most notably, the project introduces a prescriptive dimension, offering actionable insights for optimizing insurance costs and coverage. This not only enhances the role of data in insurance but also fosters a collaborative and data-driven insurance ecosystem. As the project unfolds, it opens doors for future research and applications, positioning itself at the forefront of innovative insurance analytics. Ultimately, it seeks to create a more equitable, transparent, and data-informed insurance industry to the benefit of both policyholders and insurance providers.

SUGGESTIONS

- From this study, we suggest Insurers to use predictive models to personalize insurance pricing for individual policyholders based on their unique profiles. This enables insurers to offer tailored coverage and more accurate pricing, aligning insurance charges with the actual risk profile of policyholders.

- On the basis of this study, we suggest Insurance companies to employ predictive analytics to assess and mitigate risks effectively. By identifying high-risk policyholders and offering prescriptive recommendations for risk reduction, insurers can enhance their risk management strategies and reduce claim payouts.

- By this study, we suggest Health Insurance providers to utilize predictive models to estimate healthcare costs for policyholders. Prescriptive insights can guide policyholders in making healthier lifestyle choices and adhering to preventive measures, ultimately reducing medical expenses.

- Based on this study, we suggest Insurance providers to gain a competitive edge by offering more accurate pricing and better-tailored coverage options to policyholders. This enhanced competitiveness can lead to increased market share and business growth.

SCOPE & LIMITATIONS

SCOPE

- In this study, we have proposed an innovative machine learning model which can help insurance providers identify potential churn among policyholders. Insurers can use prescriptive analytics to develop retention strategies, offering personalized incentives or coverage adjustments to enhance customer loyalty.

- This model can also be very helpful for Hospital Management to avoid improper management, will also help to control over managing, allotting resources and save time and money.

LIMITATIONS

- **Human Behavior Factors:** Predictive and prescriptive models may not account for unpredictable human behaviors that influence insurance charges, such as sudden lifestyle changes or unforeseen medical events.

- **Market Variability:** The insurance market is subject to economic fluctuations and competitive pressures. Market variability may introduce uncertainties in predictive and prescriptive recommendations.

REFERNCES

- Bertsimas, D., & Kallus, N. (2020). From predictive to prescriptive Analysis. Management Science, 66(3), 1025–1044. https://doi.org/10.1287/mnsc.2018.3253

- Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction. Springer.

- Data Mining by Mathew A. Russell

APPENDIX

#Code

Basic statistical Discription

```
df.info()
df.describe().T
```

Checking null values

```
df.isnull().sum()
```

#Exploratory Data Analysis

Histogram of Charge Amount Distribution per Smoker

```
plt.figure(figsize=(10,6))
plt.title('Charge Amount Distribution per Smoker')
sns.histplot(df,x='InsuranceCharges',hue="Smoker",kde=True,alpha=0.6,linewidth=1)
```

Bar Plot of Average Charges for Different Regions

```
fig, ax = plt.subplots()

sns.barplot(data=df, y='Region', x='InsuranceCharges', ax=ax, palette='Set1', ci=None)

ax.set_title("Average Charges for Different Regions")
ax.set_xlabel("Charges")

plt.show()
```

#Model Building

Splitting the Data

```
#Spliting the data into Train data and Test data
X_train, X_test, y_train,y_test = train_test_split(X_scaled,y,test_size=0.15,random_state=42)
```

#Table of all models

```
# table of all models

models = [LinearRegression(), LinearSVR(), DecisionTreeRegressor(), RandomForestRegressor(), KNeighborsRegressor()]

model_names = ['Linear Regression', 'SVR', 'Decision Tree', 'Random Forest', 'KNN']

R_squared = []

for i in models:
    i.fit(X_train, y_train)
    y_predict = i.predict(X_test)
    R_squared.append(i.score(X_train,y_train))

model_comparison = pd.DataFrame({'Model': model_names, 'R-squared': R_squared})
model_comparison.sort_values(by='R-squared', ascending=False)
```

#K-Neighbours Regressor with RandomSearchCV

```
# Define the parameter distribution for RandomizedSearch
param_dist = {
    'n_neighbors': np.arange(1, 20),  # Number of neighbors to consider
    'weights': ['uniform', 'distance'],  # Weighting function
    'p': [1, 2]  # Power parameter for Minkowski distance
}

# Create the KNN regressor
knn = KNeighborsRegressor()

# Perform RandomizedSearch with cross-validation
random_search = RandomizedSearchCV(knn, param_distributions=param_dist,
n_iter=10, cv=5, scoring='neg_mean_squared_error')
random_search.fit(X_train, y_train)

# Get the best hyperparameters and model
best_params = random_search.best_params_
best_model = random_search.best_estimator_

# Fit the best model to the entire training set
best_model.fit(X_train, y_train)

# Evaluate the model on the test set
y_pred = best_model.predict(X_test)
r_squared = best_model.score(X_train,y_train)
r2 = best_model.score(X_train,y_train)
n = X_train.shape[0]
p = X_train.shape[1]
adj_r2 = 1-(1-r2)*(n-1)/(n-p-1)
rscore=r2_score(y_test, y_pred)
mse = mean_squared_error(y_test, y_pred)
rmse = np.sqrt(mse)

print("\t********* K-Neighbours Regressor *********")
print("\tR-squared            : ", r_squared)
print("\tAdjusted R-squared   : ", adj_r2)
print("\tTesting R-squared    : ", rscore)
print("\tRMSE                 : ", rmse)
```