INNOMATICS®
RESEARCH LABS

INNOVATION. AUTOMATION. ANALYTICS

PROJECT ON

MEDICAL COST PREDICTION

About me

- Vijay sada
- B-tech in Electronics and Communications Engineering.
- Fresher
- LinkedIn - https://www.linkedin.com/in/vijay-sada/
- GitHub - https://github.com/vijaysada

Introduction

The project aims to create a predictive model for estimating medical costs, using advanced data analysis techniques to help healthcare providers and insurers anticipate and manage expenses more effectively. By examining past medical data, the goal is to uncover patterns and factors that influence healthcare costs, empowering stakeholders to make informed decisions.

Insurance, fundamentally, is a risk-sharing contract where one party agrees to bear the risk of another in exchange for a premium. Health insurance specifically pertains to covering or sharing the expenses associated with healthcare services. In India, the health insurance sector is rapidly growing due to factors like rising middle-class demographics, increased healthcare costs, and heightened awareness.

Accurate medical cost prediction is crucial for both healthcare providers and insurers as it impacts financial stability and service quality. By foreseeing potential expenses, providers can allocate resources efficiently, ensuring optimal patient care. Similarly, insurers can set fair premiums and develop tailored coverage plans, improving customer satisfaction and company profitability. Ultimately, this project seeks to enhance cost transparency and efficiency in the healthcare sector, benefiting both providers and patients alike.

Data Overview

The dataset was provided to us by the client.

The dataset includes the following features:

Age: Primary beneficiary's age.

Sex: Gender of the insurance contractor.

BMI: Body mass index indicating weight relative to height.

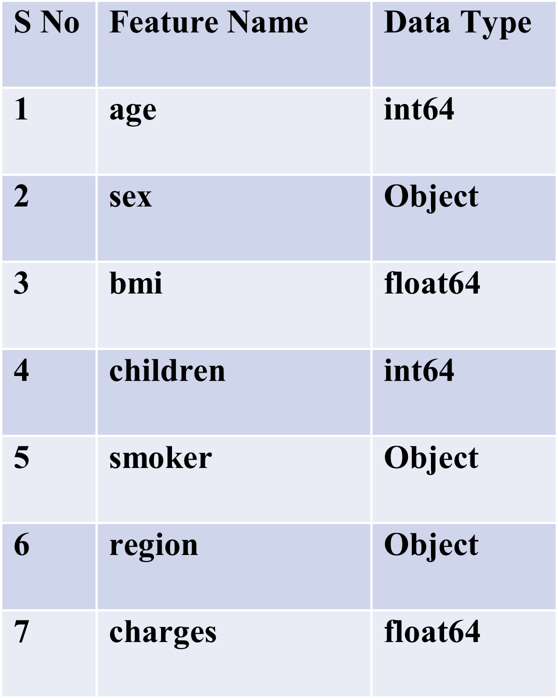
Children: Number of dependents.

Smoker: Categorical indicator of smoking status.

Region: Beneficiary's residential area in the country.

Charges: Individual medical costs billed by health insurance.

Our target column is ‘Charges’



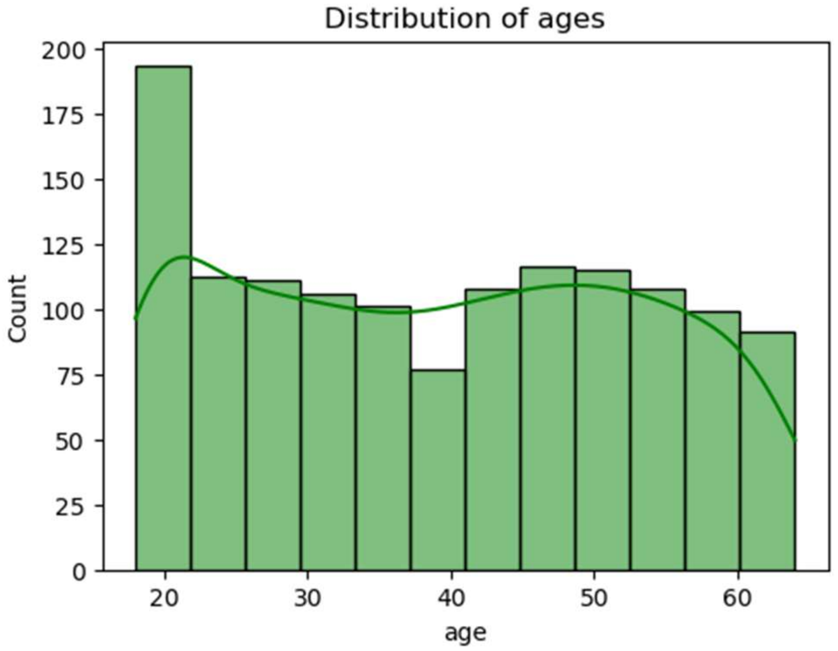
S No	Feature Name	Data Type
1	age	int64
2	sex	Object
3	bmi	float64
4	children	int64
5	smoker	Object
6	region	Object
7	charges	float64

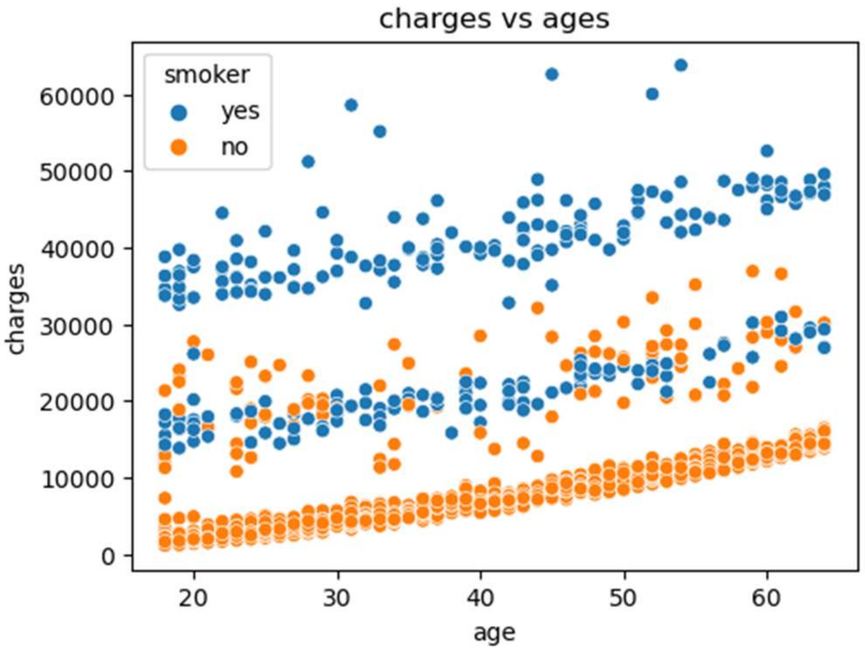
Pre-Processing

For the project, the following preprocessing steps were performed:

- **Data Cleaning:** Identified and handled inconsistencies like duplicate values, and outliers were found in the ‘bmi’ column the dataset.
- **Handling Duplicate Values:** Addressed the duplicate data point at row index 581 by removing to maintain the data consistency.
- **Standardization:** Standardized numerical columns (age, bmi, children) using Robust Scaling to ensure features are on a similar scale and address outliers in the bmi column.
- **Handling Categorical Variables:** Encoded categorical variables into numerical format using one-hot encoding as they represent nominal data, making them suitable for modeling.

Exploratory Data Analysis (EDA)

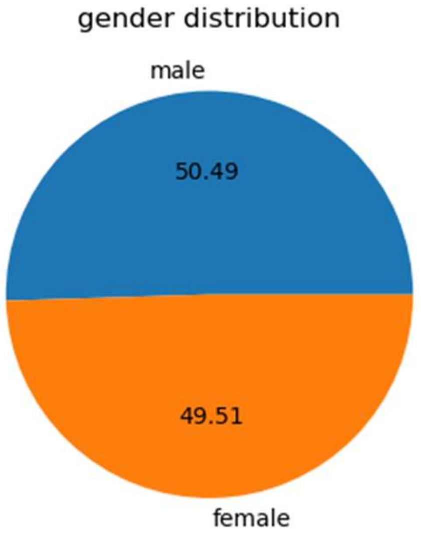


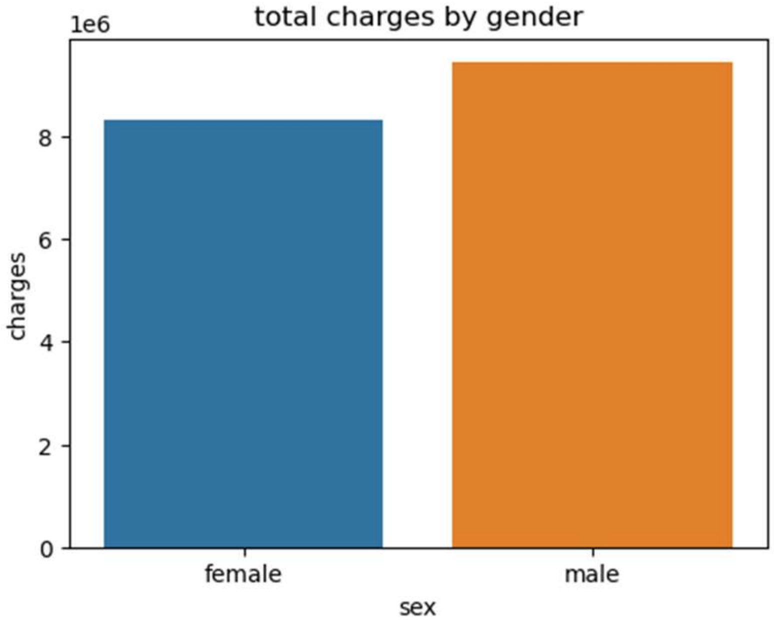


- From the above we can observe that most of the people are around the age of 20. We can also see that the charges increases along with the age of people, so we can say that the charges are quite high for people who smoke compared to those who don’t.

INNOMATICS RESEARCH LABS

EDA (Cont.):

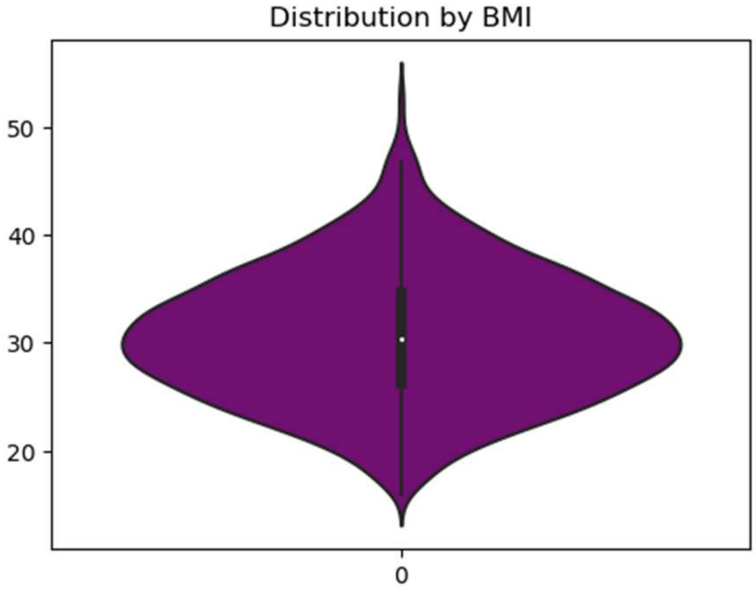


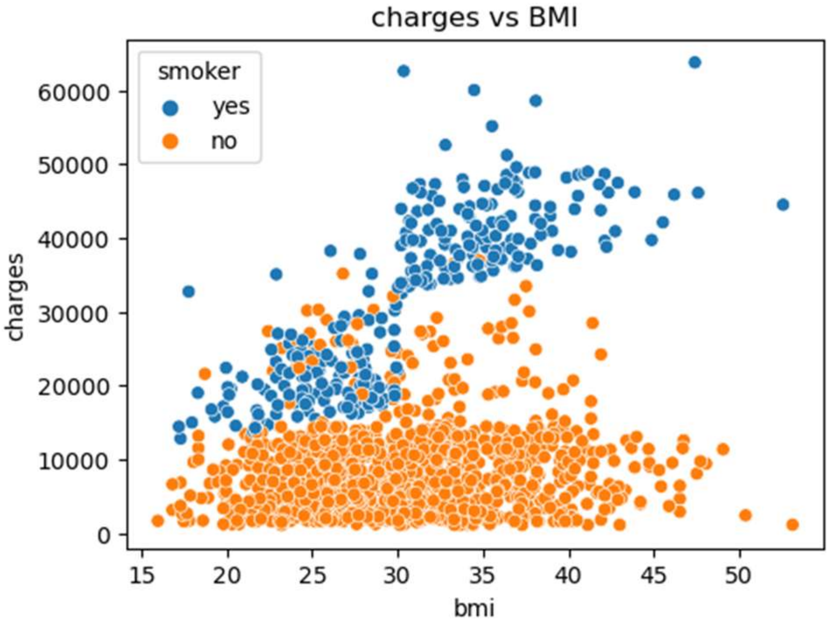


From the above we can observe that there are more male insurance holders in comparison to female insurance holders.

From the above we can observe that male are paying more for insurance in total by charges when compared with women.

EDA (Cont.):

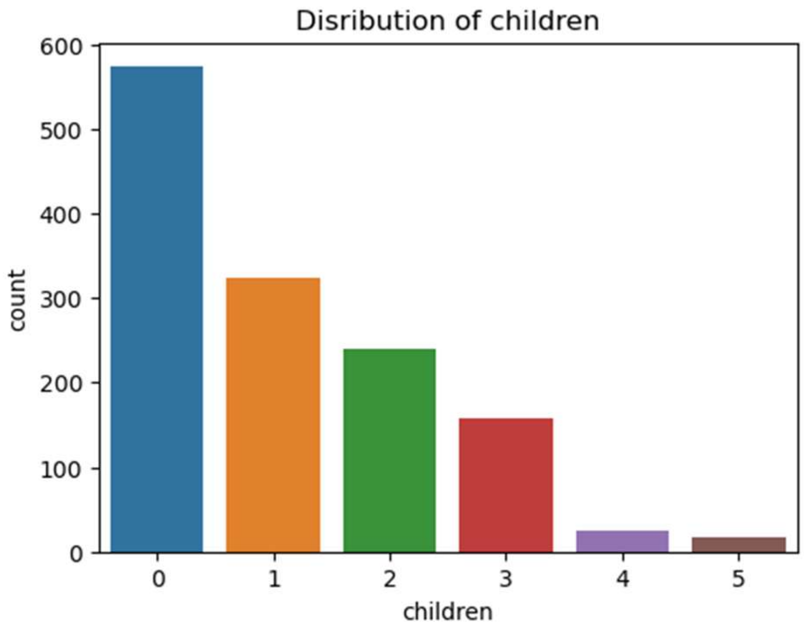


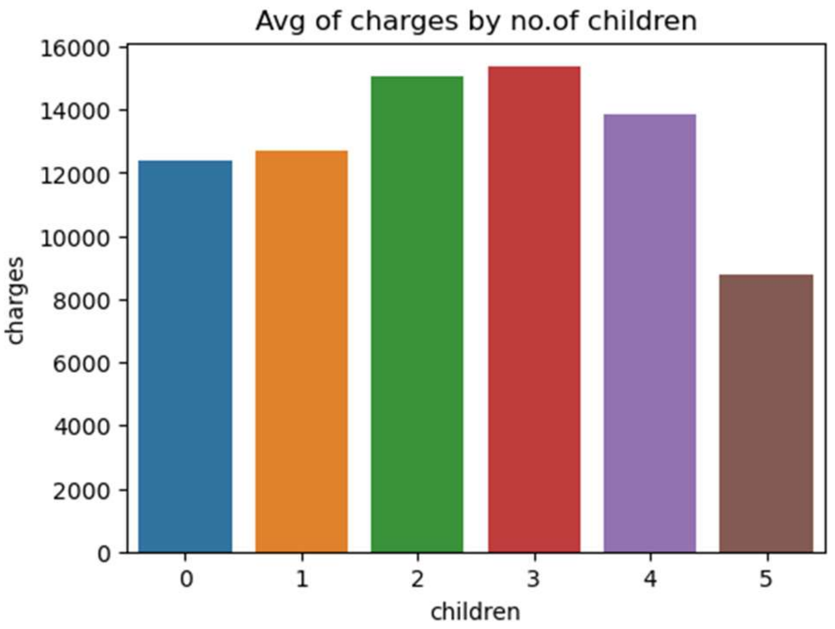


From the above we can say that most of the people have an average BMI of around 30.

From the above we can also observe that people with the same BMI but are smoking are getting charged more in comparison to people who don't smoke.

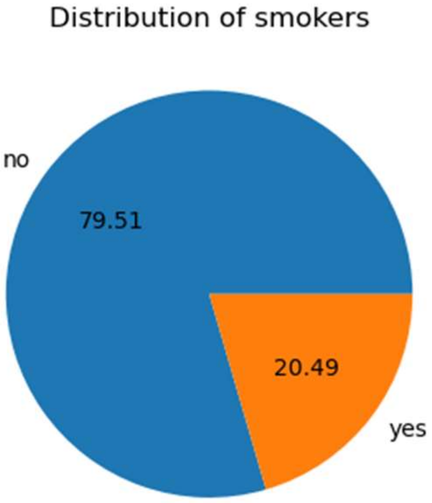
EDA (Cont.):

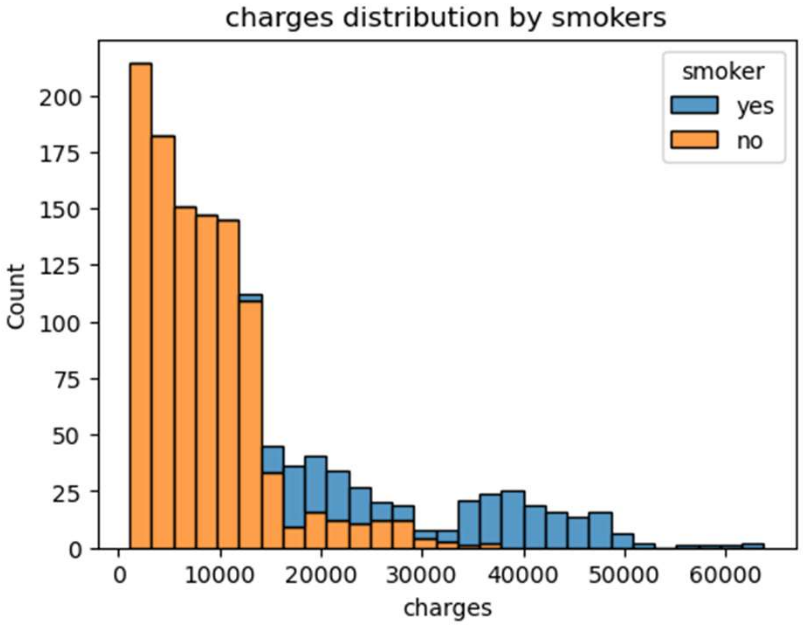

Disribution of children


Avg of charges by no.of children

- From the above observations we can say that people with no children are more in number and because of that they're the highest contributors to the total charges paid.
- We can observe that a families with 2 or 3 children are paying the highest charges on average when compared to rest of the families.

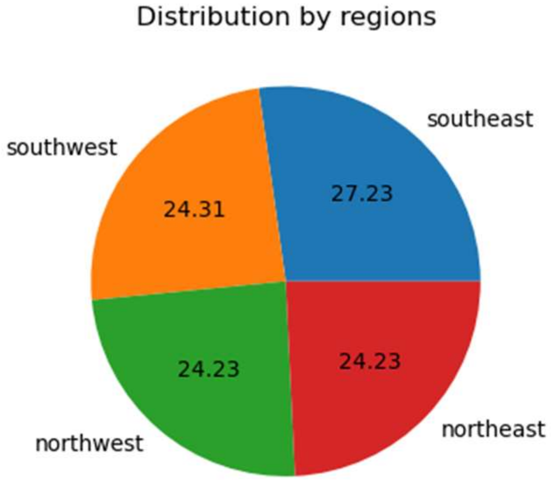
EDA (Cont.):

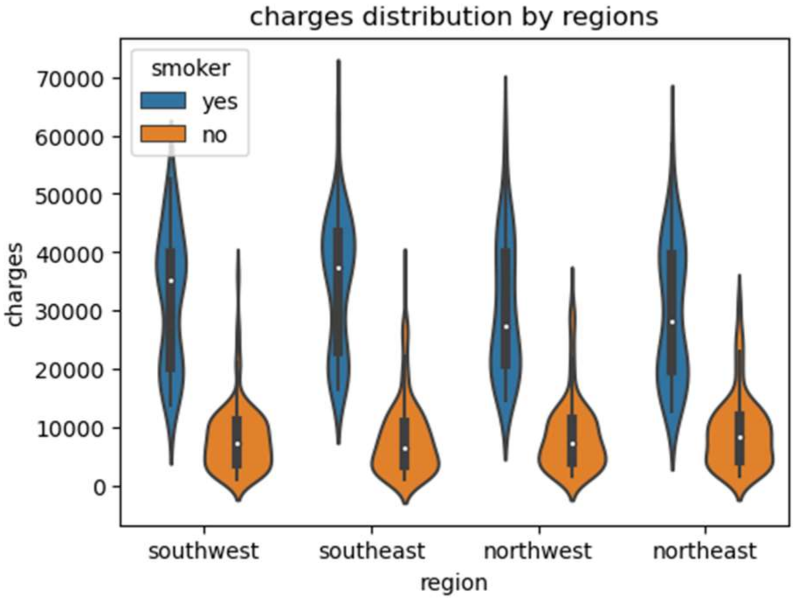




- From the above we can observe that around 80 percent of people are non smokers.
- We can observe that smoking individuals are paying more charges compared to non-smokers.

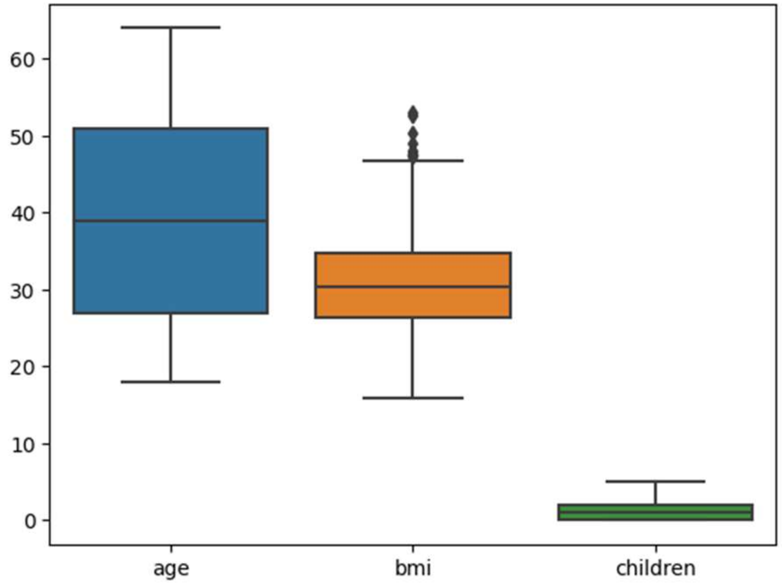
EDA (Cont.):



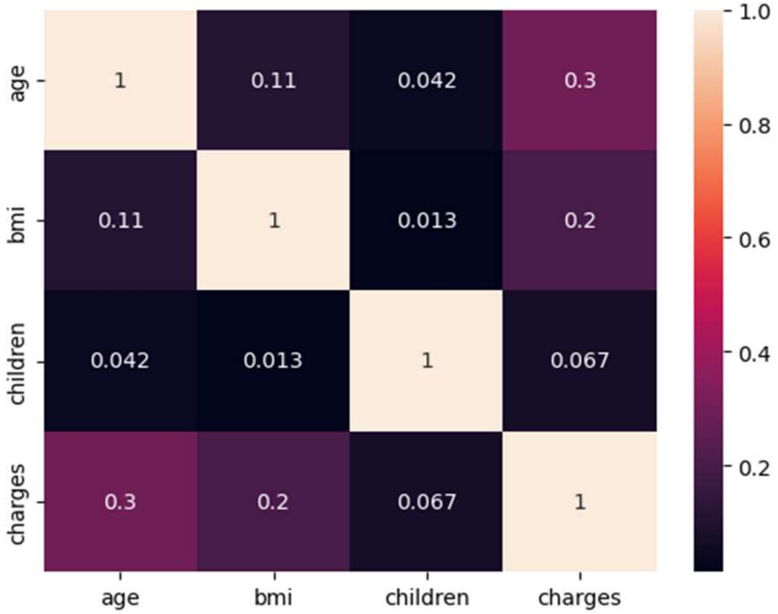


- From the above we can see that the southwest region is the dominant one by population percentage while the other 3 regions have a similar distribution of people.
- From the above we can observe that the people from the southeast region are paying more charges be it male or female compared to the rest of the regions. So we can say that the rest of the 3 regions have almost same charges.

EDA (Cont.):



From the above we can observe that there exists outliers in bmi column.



From the above we can observe that age and charges have a positive correlation.

Model Building

For the classification task of cost prediction, the below machine learning models were used:

- **Linear Regression:** A linear model that establishes a relationship between independent and dependent variables by fitting a straight line to the data, making it suitable for predicting continuous outcomes with a linear relationship.
- **KNeighbors Regressor:** A non-parametric regression algorithm that predicts the target variable based on the average of the values of its k-nearest neighbors, providing flexibility in capturing local patterns in the data but sensitive to the choice of k.
- **Decision Tree Regressor:** Utilizes a tree-like model of decisions to predict the target variable by recursively partitioning the feature space, making it interpretable and capable of capturing non-linear relationships but prone to overfitting.
- **Random Forest Regressor:** A powerful ensemble method that constructs multiple decision trees and aggregates their predictions to improve accuracy and generalization, effective for handling complex datasets and reducing overfitting.
- **SVR (Support Vector Regression):** A regression algorithm that finds the hyperplane in a high-dimensional space to minimize the error between predicted and actual values, suitable for datasets with clear margins of separation but less interpretable for complex relationships.

Model Training

The model training process involved:

- Splitting the dataset into training and testing sets (75:25 Ratio) to train the models on a subset of the data and evaluate their performance on unseen data.
- Training each model using the training set, where the model learns patterns and relationships between features and the target variable (charges).

S No	Type of Problem	Algorithm Name
1	Regression	Linear Regression
2	Regression	KNeighbors Regressor
3	Regression	Decision Tree Regressor
4	Regression	Random Forest Regressor
5	Regression	SVR

Model Evaluation

Evaluation metrics employed to assess model effectiveness was Mean Absolute Error :

- **Mean Absolute Error (MAE):** MAE measures the average absolute difference between the predicted and actual values, providing a straightforward indication of the model's performance in terms of prediction accuracy without considering the direction of errors. Lower MAE values indicate better predictive accuracy, with each absolute difference equally weighted in the calculation.

S No	Algorithm Name	Metric Score
1	Linear Regression	4323.963298
2	KNeighbors Regressor	3854.502497
3	Decision Tree Regressor	3081.074484
4	Random Forest Regressor	2748.811625
5	SVR	9114.819321

Challenges Faced

We’ve found that there are outliers in the input column “bmi”. As they are true outliers, we’ve retained the values and as we can’t use standard scaler for this data since it might create a bias due to the outliers, so in place of it we’ve used the robust scaler since it handles the outliers better.

Conclusion

We can observe that out of all the models that we've trained, the Random Forest Regressor model is giving the lowest mean absolute error, which means predictions made by this model are the closest to the real values in comparison to the other models. So, we can say that by observing the evaluation metric of all models that the Random Forest Regressor is the best algorithm for the Medical Cost Prediction problem.

THANK YOU

INNOMATICS RESEARCH LABS