ML End Term Presentation

Health Insurance Cost Prediction

Accurately Forecasting Future Healthcare Costs for Smarter Financial Planning

Presentation by
GROUP 08
CSE-04

Group 08

Our Team

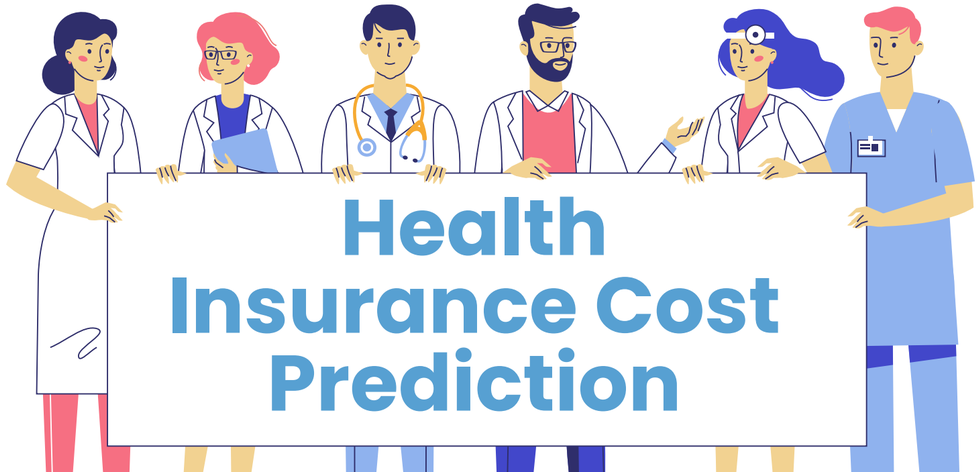
Radhika Bhati

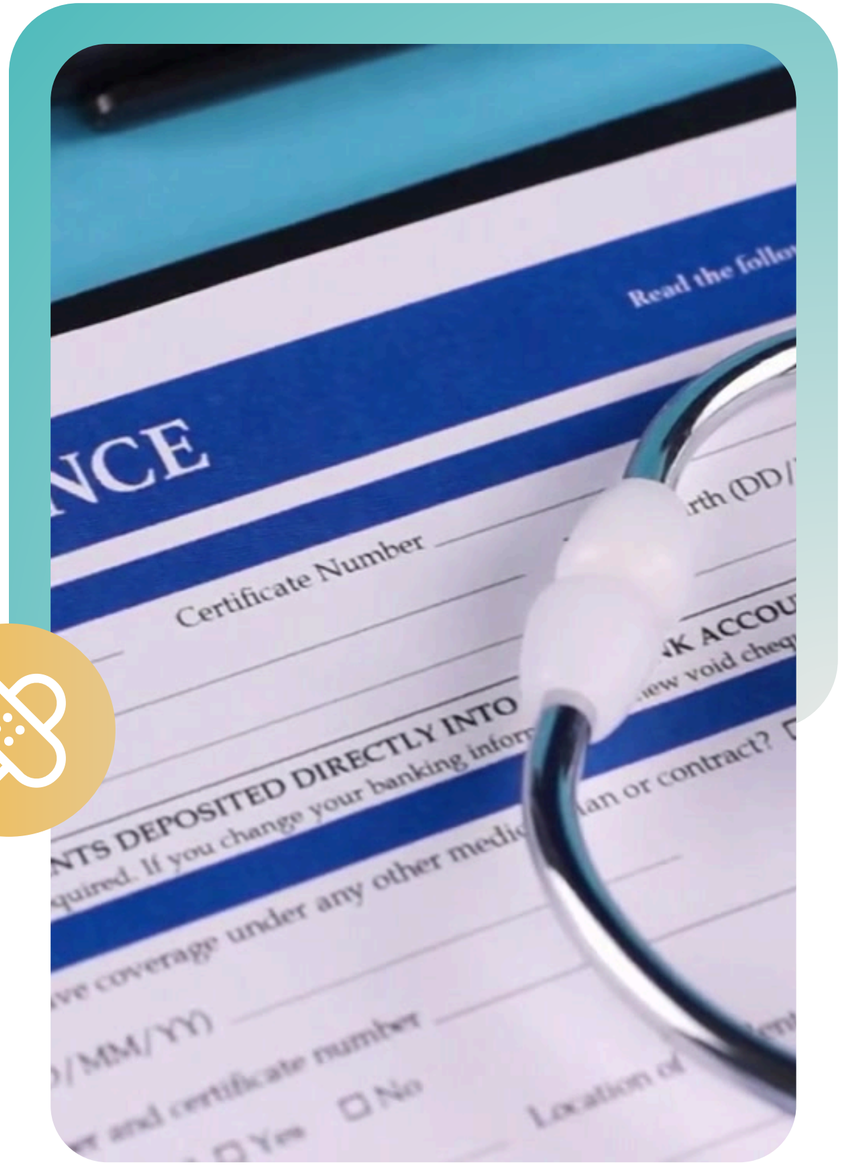
Deepanshu Aggarwal

Harshita Rupani

Ananya Srivasatava

Rohit





GROUP 08

Problem Statement

Healthcare costs are rising, and accurately predicting individual health insurance premiums is crucial for both insurers and policyholders. It remains a significant challenge due to the complex interplay of various factors such as age, gender, BMI, number of children, smoking status, and geographic region. Traditional methods often fail to account for these complexities, leading to inaccurate cost estimations and financial strain for both insurers and policyholders. This project addresses the need for a more precise and reliable prediction model to ensure fair pricing and better financial planning.

GROUP 08

Objective

The objective of this project is to create a reliable machine learning model that can accurately predict health insurance costs for individuals based on key factors such as age, gender, BMI, number of children, smoking status, and geographic region. By leveraging advanced algorithms and data analysis techniques, this project aims to provide a precise estimation tool that helps insurance companies determine fair premiums and enables individuals to anticipate their healthcare expenses more effectively. The ultimate goal is to enhance the transparency, accuracy, and fairness of the health insurance pricing process, leading to better financial planning and decision-making for all stakeholders involved.

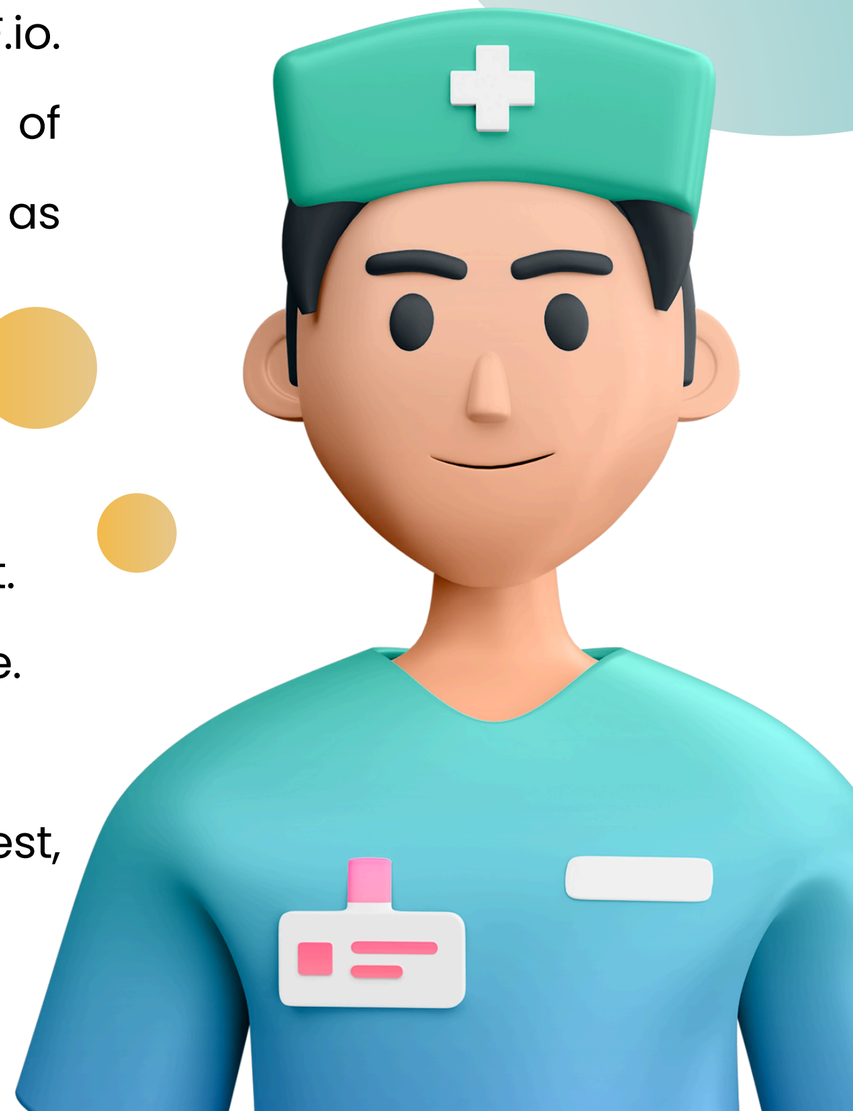
DataSet

Overview of the dataset used

GROUP 08

The dataset used for this project is the 'insurance.csv' file, sourced from OSF.io. It comprises 1338 rows and 7 columns, providing a comprehensive set of features relevant to health insurance cost prediction. The columns are as follows:

1. age: The age of the individual.
2. sex: The gender of the individual (male or female).
3. bmi: Body Mass Index, a measure of body fat based on height and weight.
4. children: The number of children or dependents covered by the insurance.
5. smoker: Smoking status of the individual (yes or no).
6. region: residential area of the individual in the U.S. (northeast, northwest, southeast, southwest).
7. charges: The medical costs billed to the insurance, target variable.



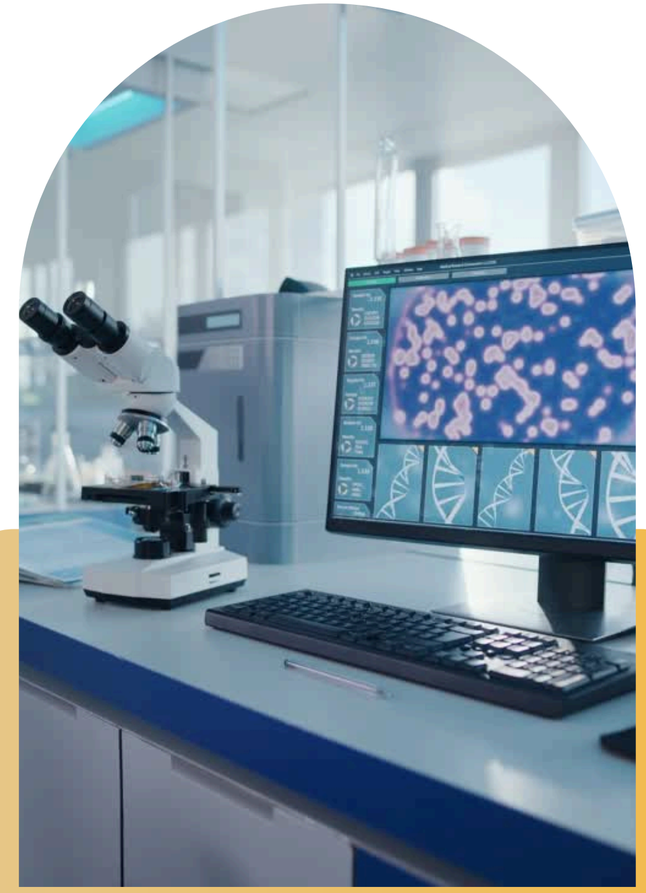
A Snapshot of the dataset:

	age	sex	bmi	children	smoker	region	charges
0	19	female	27.900	0	yes	southwest	16884.92400
1	18	male	33.770	1	no	southeast	1725.55230
2	28	male	33.000	3	no	southeast	4449.46200
3	33	male	22.705	0	no	northwest	21984.47061
4	32	male	28.880	0	no	northwest	3866.85520
...	...	...	...	...	...	...	...
1333	50	male	30.970	3	no	northwest	10600.54830
1334	18	female	31.920	0	no	northeast	2205.98080
1335	18	female	36.850	0	no	southeast	1629.83350
1336	21	female	25.800	0	no	southwest	2007.94500
1337	61	female	29.070	0	yes	northwest	29141.36030

1338 rows × 7 columns

GROUP – 08

Methodology



1

Exploratory Data Analysis (EDA):

First, loaded the dataset and analyzed its structure and summary statistics. Addressed missing or inconsistent data points. Converted categorical variables like 'sex', 'smoker', and 'region' into numerical format, for model compatibility.

2

Data Visualization:

Plotted a heatmap to visualize the correlation between independent features and the target variable 'charges' to identify significant relationships. Created scatter plots, histograms, and box plots to explore the relationships and distributions of the features.

3

Statistical Analysis:

Skew and Kurtosis: Analyze the skewness and kurtosis of the data to understand the distribution characteristics of the features and target variable.

Group -08

4

Data Preparation:

Apply feature scaling techniques to standardize numerical features for equal contribution to model training. Split the dataset into training and testing sets to assess the model's performance on unseen data.

5

Model Development:

Devloped regression models - Linear Regression, Support Vector Regression (SVR), Ridge Regression and Random Forest Regressor.

6

Hyperparameter Tuning:

Tuned hyperparameters of each regressor using techniques like Grid Search or Random Search to optimize performance.

7

Model Evaluation:

Evaluated models using metrics like Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared (R^2). Compared the performance of all models to identify the best-performing regressor.

8

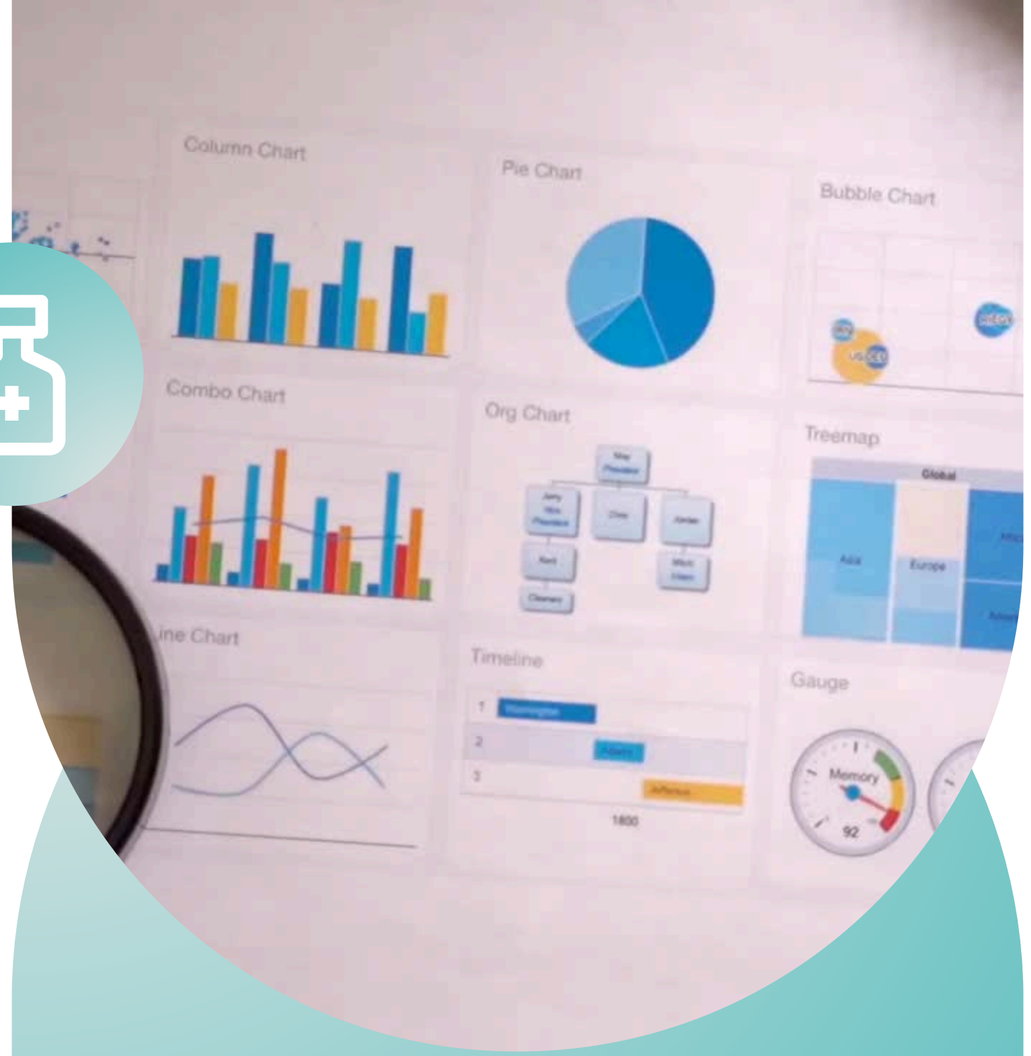
Model Deployment:

Selected the best-performing regressor (Random Forest Regressor) based on evaluation metrics. Deployed using Flask to create a user-friendly web application for real-time health insurance cost prediction.

Group - 08

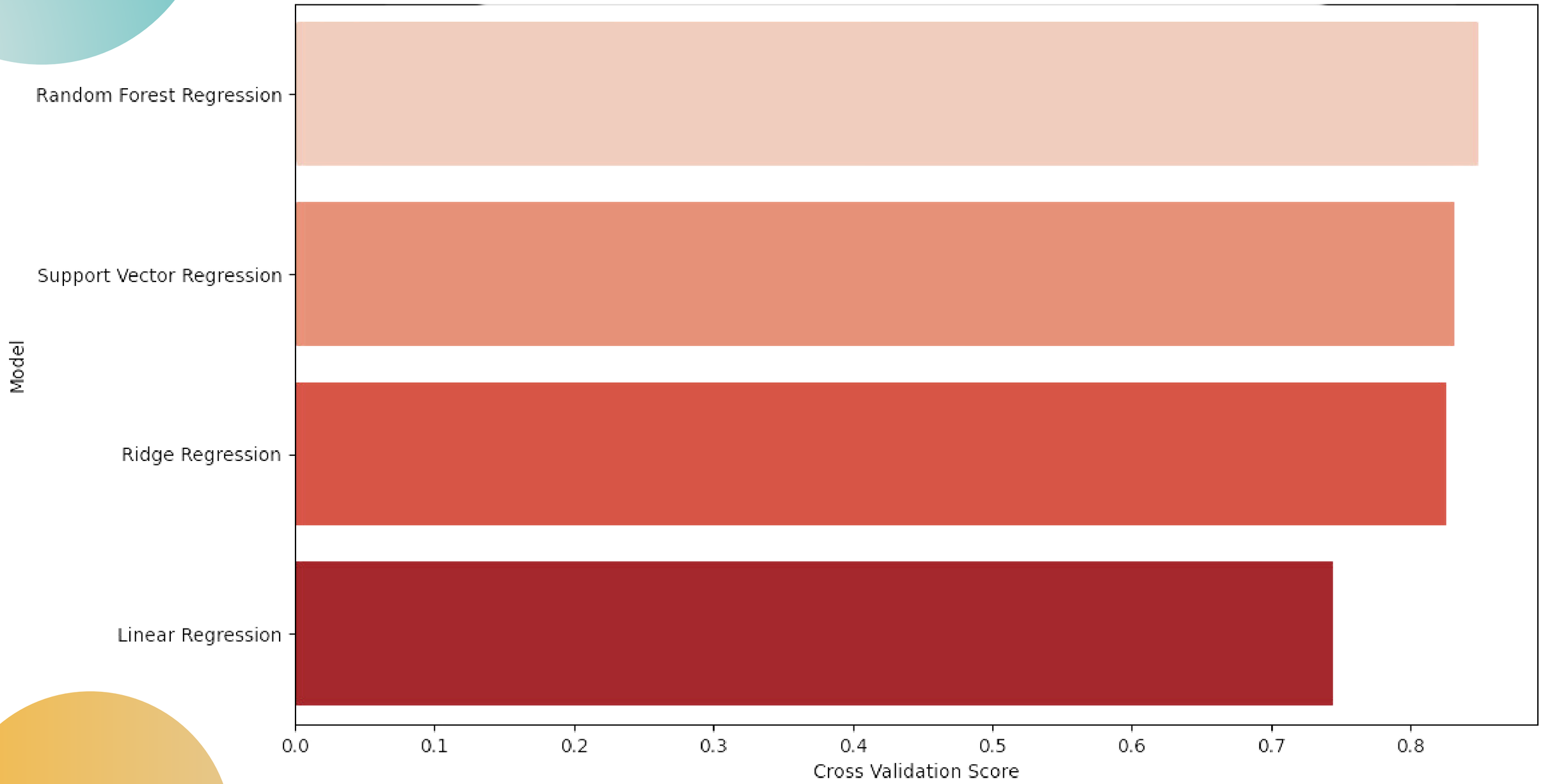
Results and Demonstration

In the further slides we would be showcasing the results as well demonstrating the working of our project



Medical

Comparing the performance of each model



Health Insurance Cost Prediction

Prepared by group 08 CSE-04
BML Munjal University

Predict the cost for your Medical Insurance!

Age
20

Gender
0

BMI
28

Children
0

Do you Smoke?
1

Which Region?
2

PREDICT PROBABILITY

Predicted Amount :

Expected amount is 19079.694

Model Deployment

Group -08

Conclusion

Our project lays a solid foundation for advancing the field of healthcare cost analysis and prediction. By harnessing the power of data science and machine learning, we aim to empower stakeholders with actionable insights.

In summary, this project successfully developed and deployed a Random Forest Regressor model for predicting health insurance costs based on demographic and lifestyle factors. Through thorough analysis and evaluation, the Random Forest Regressor demonstrated superior performance compared to other regression algorithms. The deployed model provides a user-friendly interface for real-time predictions, offering valuable insights for insurers and individuals to make informed decisions and plan finances effectively in the healthcare domain.

Group - 08

Future Scope

Looking ahead, our project on health insurance cost prediction has significant potential for future advancements. Integrating real-time data streams and IoT devices offers opportunities for dynamic monitoring and adaptive pricing. Collaboration with healthcare stakeholders can lead to more holistic approaches to healthcare delivery and resource allocation. These advancements will enable insurers to offer tailored plans, aligning with individual needs and empowering informed decision-making.

Group - 08

Thank You

MACHINE LEARNING END TERM PRESENTATION

Presentation by
GROUP -8 , CSE 4