Medical Insurance

GROUP 7 / Bhavans Vivekananda Degree College

Akshay Kumar| Khushi Bhansali | Paluri Sathvik | Pavan Prajapat

Abstract

- The world is full of uncertainties that expose people, families, and businesses to risks like mortality, illness, and property loss, with life and well-being being paramount. Although risks can't always be eliminated, the financial sector offers protective solutions to safeguard individuals and organizations. One such essential product is medical insurance

- The study explores machine learning algorithms like Linear Regression, Decision Tree, Random Forest, K-Nearest Neighbours to predict Medical Insurance Premium

Objective

To find the suitable machine learning model to predict the Medical insurance Premium

CONTENT

Introduction

- Medical insurance is a contract between an individual and an insurance provider where the insurer covers medical expenses in exchange for periodic premiums. It helps protect individuals and families from high healthcare costs, ensuring access to necessary medical services without financial strain.

- Medical insurance typically covers a wide range of healthcare services, including hospitalization, surgeries, prescription drugs, preventive care, and emergency treatments. It plays a crucial role in promoting public health, managing healthcare costs, and providing financial security during medical emergencies.

Objective

To find the suitable machine learning model to predict the Medical insurance Premium

DATA PREPROCESSING

Data

Dataset : Our Dataset consists of 7 variables and 1339 records

Link : https://drive.google.com/file/d/1KUgAirYsOeVeoL3JMITNddObGL3se3J0/view?usp=drive_link

Variables:

Categorical	**Numerical**
Region	Age
Sex	BMI
Smoker	Children
	Charges

	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

Data Cleaning

- Removing Columns: As there are same dates with different time stamps we deleted the column named date from the data.

- Checked for missing values and unique values

- Enlabled and Dummified the categorical variables: Enabled weekday and weekend from the variable WeekStatus to 0 and 1 respectively

Descriptive Statistics

	age	bmi	children	charges
count	1338.000000	1338.000000	1338.000000	1338.000000
mean	39.207025	30.663397	1.094918	13270.422265
std	14.049960	6.098187	1.205493	12110.011237
min	18.000000	15.960000	0.000000	1121.873900
25%	27.000000	26.296250	0.000000	4740.287150
50%	39.000000	30.400000	1.000000	9382.033000
75%	51.000000	34.693750	2.000000	16639.912515
max	64.000000	53.130000	5.000000	63770.428010

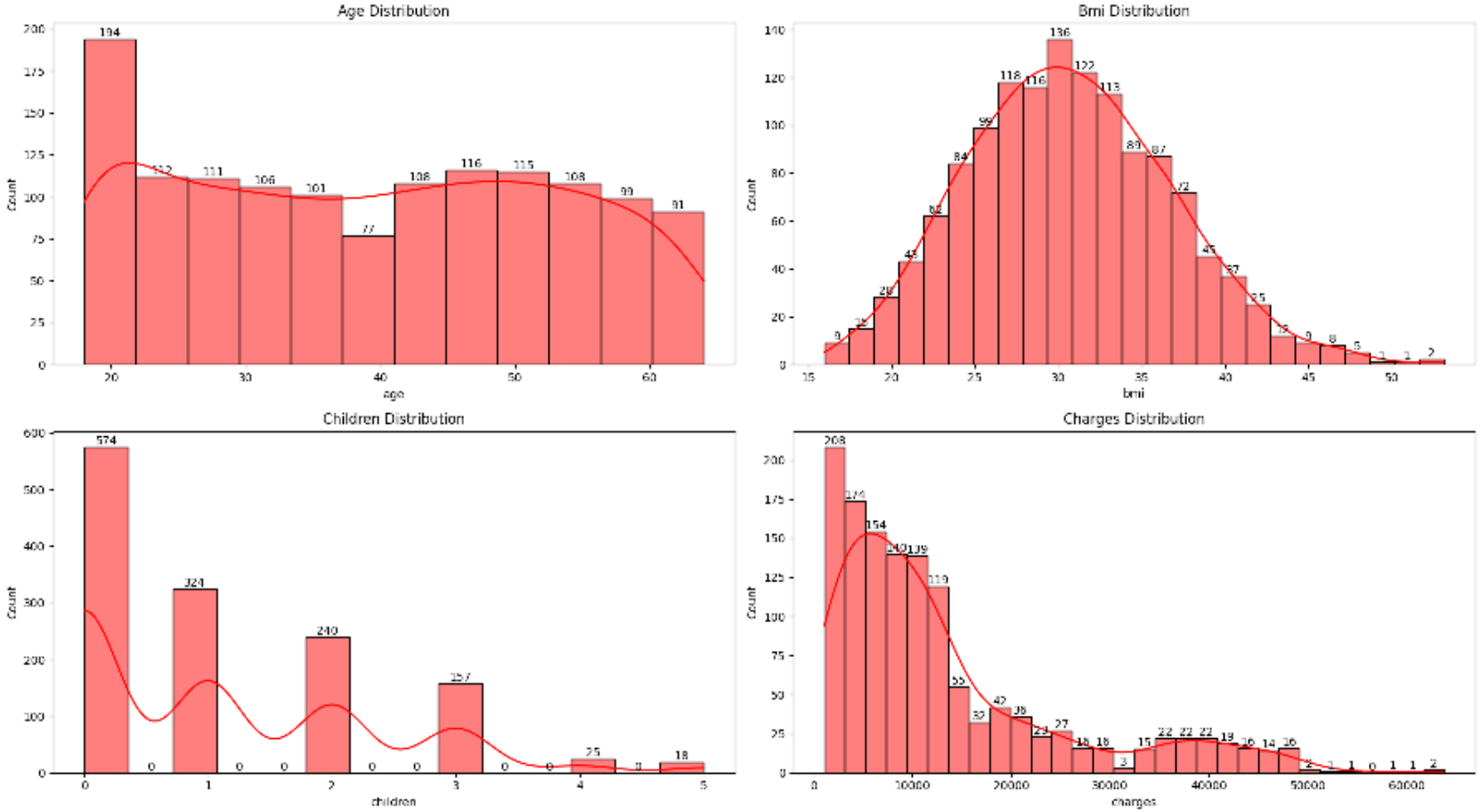
- To perform dummy variable encoding we divided the data into two sets
  - continuous data
  - categorical data

	age	bmi	children	charges	sex_male	smoker_yes	region_northwest	region_southeast	region_southwest
0	19	27.900	0	16884.92400	False	True	False	False	True
1	18	33.770	1	1725.55230	True	False	False	True	False
2	28	33.000	3	4449.46200	True	False	False	True	False
3	33	22.705	0	21984.47061	True	False	True	False	False
4	32	28.880	0	3866.85520	True	False	True	False	False
...	...	...	...	...	...	...	...	...	...
1333	50	30.970	3	10600.54830	True	False	True	False	False
1334	18	31.920	0	2205.98080	False	False	False	False	False
1335	18	36.850	0	1629.83350	False	False	False	True	False
1336	21	25.800	0	2007.94500	False	False	False	False	True
1337	61	29.070	0	29141.36030	False	True	True	False	False

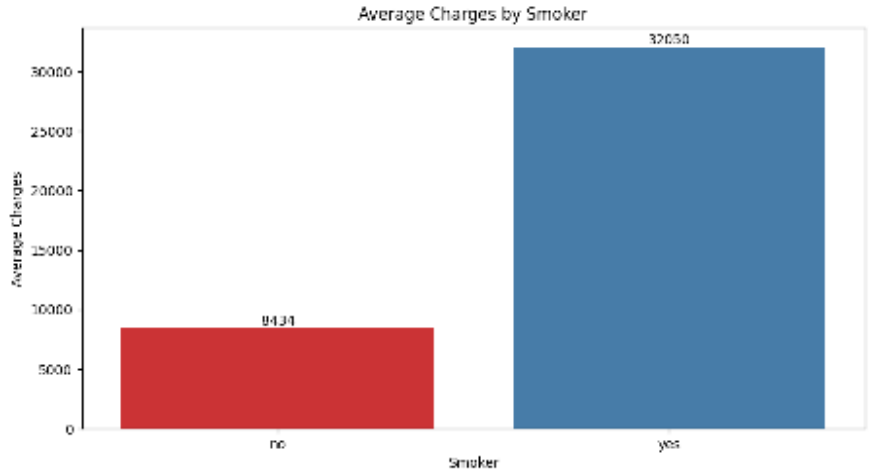
1338 rows × 8 columns

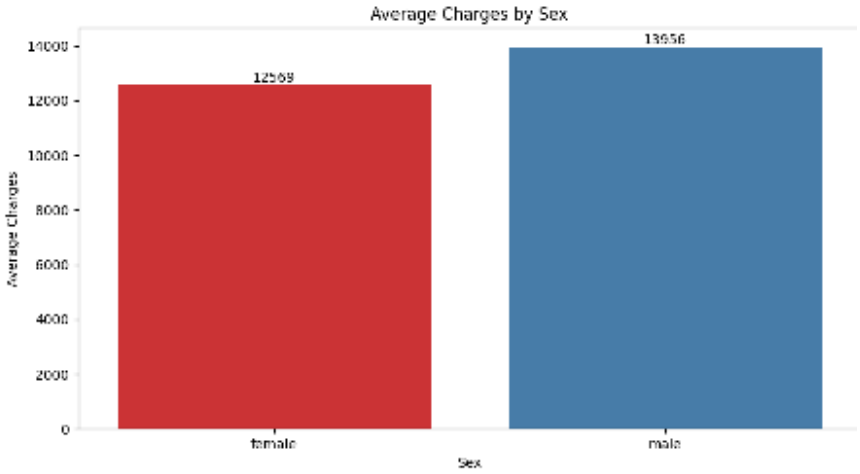
EXPLANATORY DATA ANALYSIS

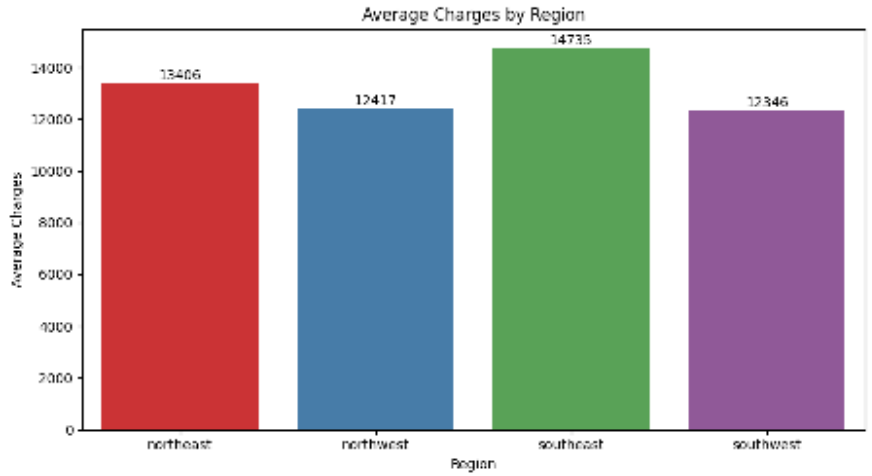
Histograms


Age Distribution
Count
age
Bmi Distribution
Count
bmi
Children Distribution
Count
children
Charges Distribution
Count
charges

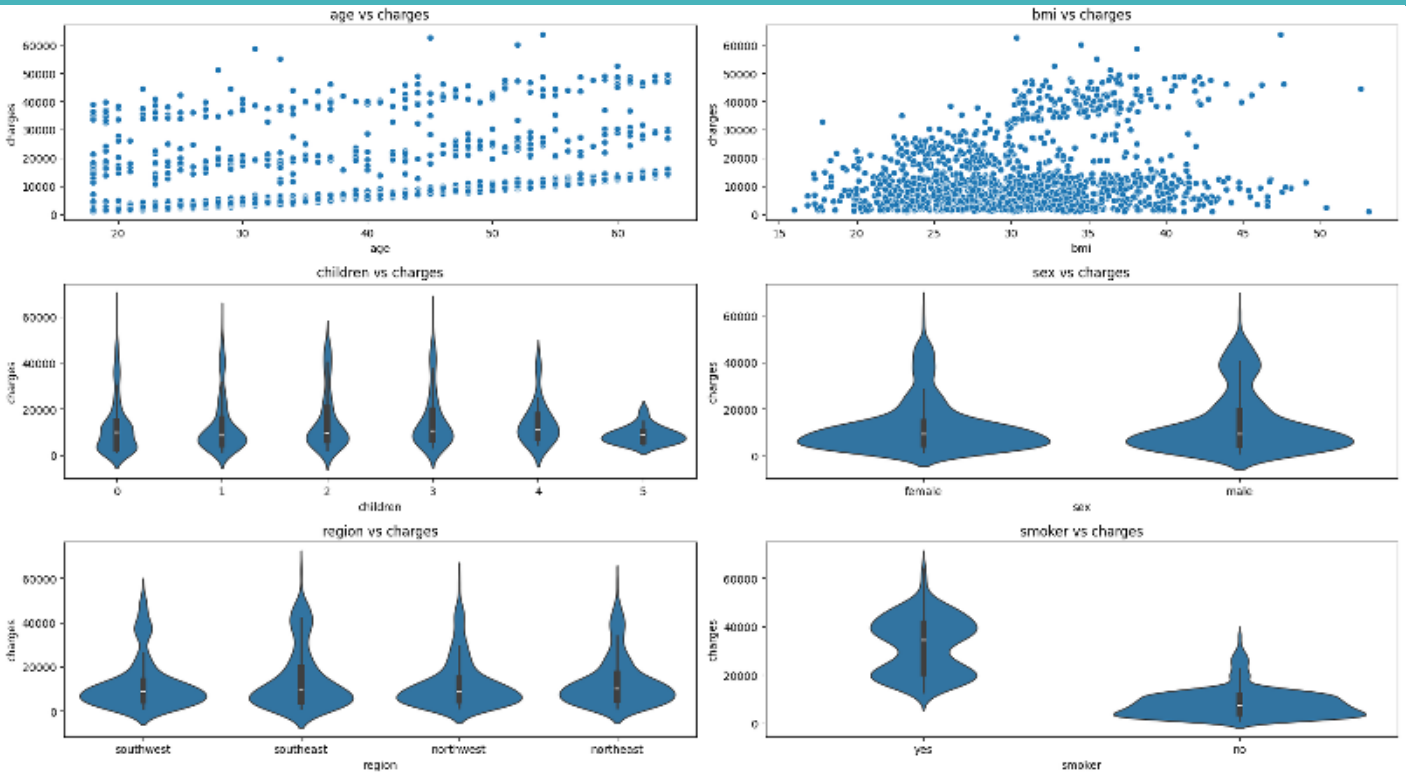
Bar Plots


Average Charges by Smoker
Average Charges
Smoker
no
yes
8434
32050


Average Charges by Sex
Average Charges
Sex
female
male
12569
13956


Average Charges by Region
Average Charges
Region
northeast
northwest
southeast
southwest
13406
12417
14735
12346

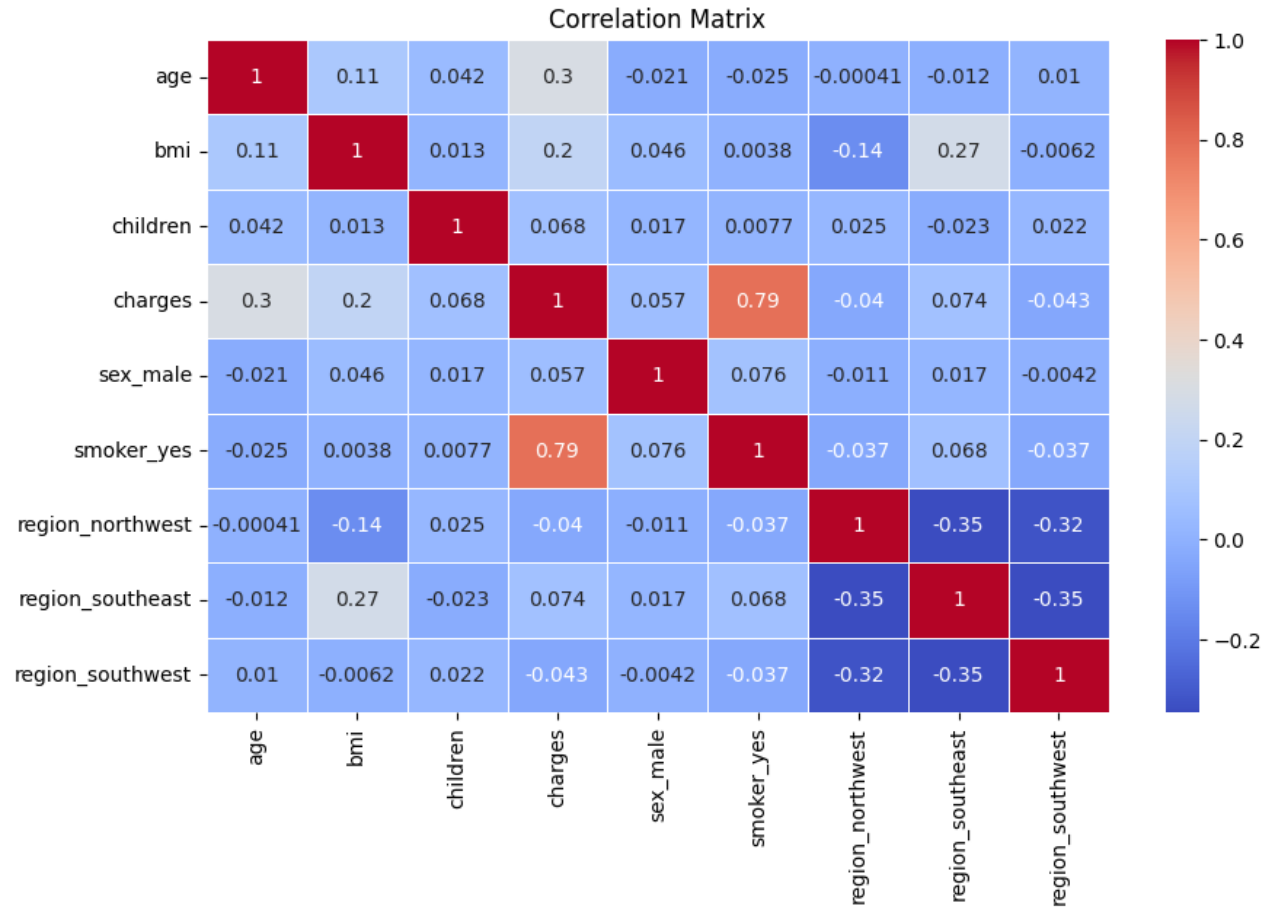
Scatter and Violin Plots



- **Age and BMI show a positive trend with charges: As age and BMI increase, insurance charges tend to increase. This is evident from the scatter plots for "age vs charges" and "bmi vs. charges".**

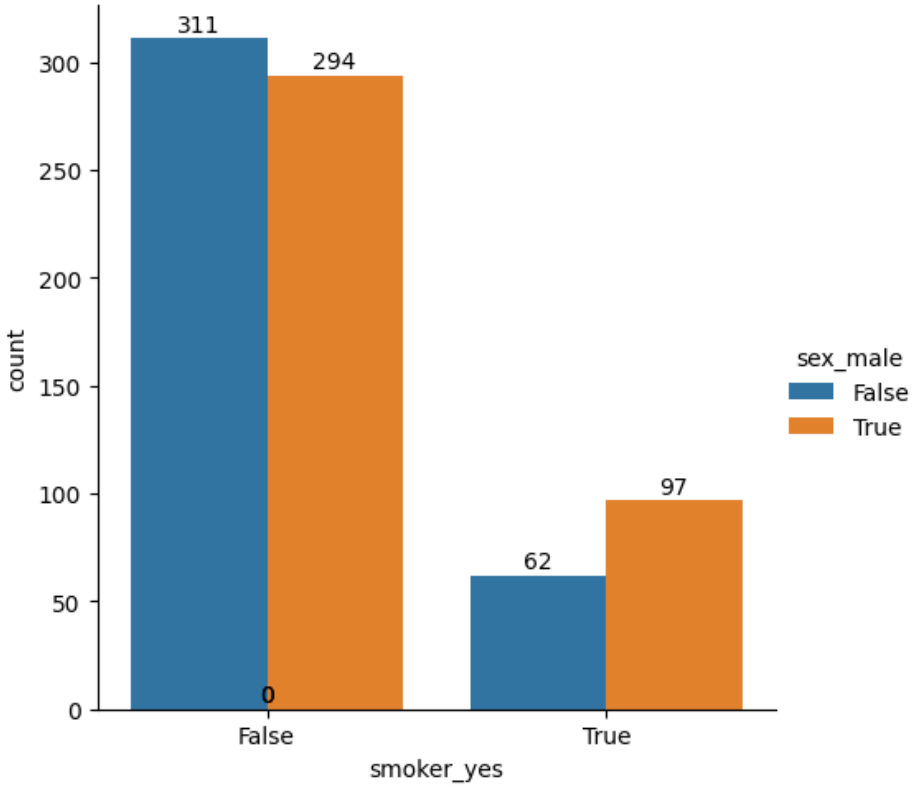
- **Smoking strongly influences charges: The "smoker vs charges" plot reveals a significant difference in charges between smokers and non-smokers. Smokers generally face much higher charges compared to non-smokers.**

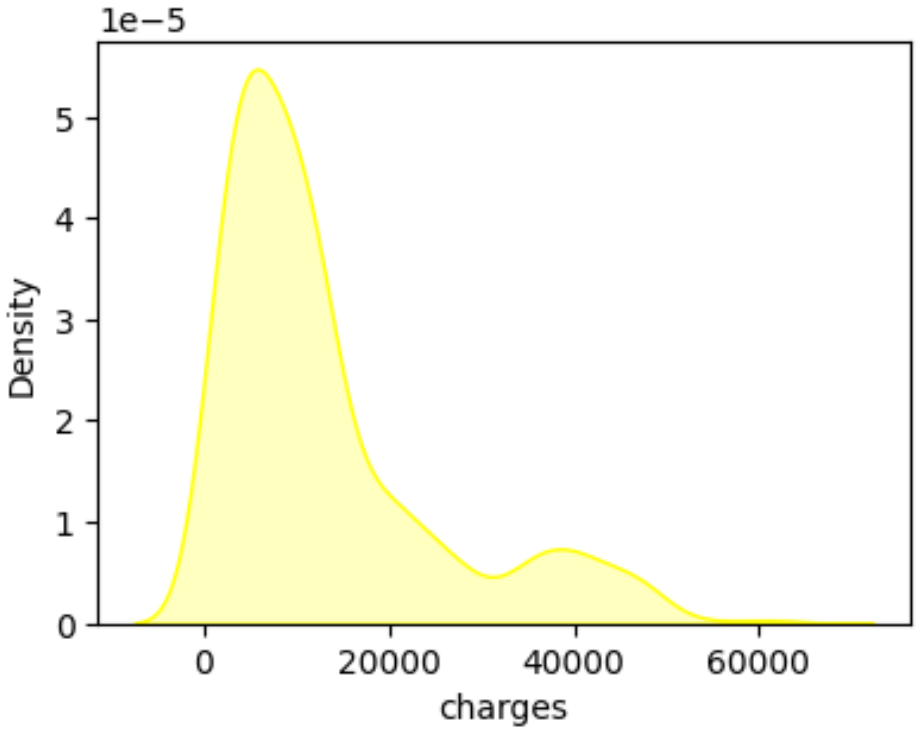
Correlation Matrix



We can observe Charges and Smoker_yes are most positively correlated

Plots


311
294
0
62
97
300
250
200
150
100
50
0
count
False
True
smoker_yes
sex_male
False
True


1e−5
5
4
3
2
1
0
Density
0
20000
40000
60000
charges

Multicollinearity Check

	Feature	VIF
0	age	8.098132
1	bmi	8.044400
2	children	1.800015
3	charges	2.473524

Regression

MACHINE LEARNING ALGORITHMS

(REGRESSION)

ML ALGORITHIMS

- Linear Regression
- K-Nearest Neighbors(KNN)
- Decision Tree

80:20 Train-Test Split

Algorithms	Model-1(r2 score)	Model-1(MAE)
Linear Regression	0.783	4181.1
KNN	0.188	7271.3
Decision Tree	0.533	0.115
Random Forest	0.755	0.125

75:25 Train-Test Split

Algorithms	Model-1(r2 score)	Model-1(MAE)
Linear Regression	0.76	3370.5
KNN	0.20	7271.3
Decision Tree	0.62	0.094
Random Forest	0.732	0.13

70:30 Train-Test Split

Algorithms	Model-1(r2 score)	Model-1(MAE)
Linear Regression	0.76	4145.4
KNN	0.174	7347.3
Decision Tree	0.512	0.119
Random Forest	0.72	0.133

60:40 Train-Test Split

Algorithms	Model-1(r2 score)	Model-1(MAE)
Linear Regression	0.76	4240.8
KNN	0.108	7884
Decision Tree	0.61	0.097
Random Forest	0.71	0.132

ALGORITHIM COMPARISON

Algorithms	Model-1(r2 score)	Model-1(MAE)
Linear Regression	0.783	4181.1
KNN	0.188	7271.3
Decision Tree	0.533	0.115
Random Forest	0.755	0.125

Summmary

- The purpose of this research is to determine the best-performing machine learning techniques to predict Insurance Premium

- Random Forest has the highest R^2 score (0.755) and the lowest MAE (0.125), indicating it performs the best among the four algorithms.of predictions in consideration to other regression models considered in this research
- Linear Regression also performs relatively well with a high R^2 score (0.783), but its MAE (4181.1) is higher than the others.

Classification

Machine Learning Algorithms (Classification)

ML ALGORITHIMS

- Logistic Regression
- K-Nearest Neighbors(KNN)
- Decision Tree

80:20 Train-Test Split

Algorithms	Model-1(r2 score)	Model-1(MAE)
Logistic Regression	0.910	0.208
KNN	0.79	0.2079
Decision Tree	0.917	0.082
Random Forest	0.94	0.059

75:25 Train-Test Split

Algorithms	Model-1(r2 score)	Model-1(MAE)
Logistic Regression	0.89	0.107
KNN	0.77	0.223
Decision Tree	0.910	0.089
Random Forest	0.91	0.865

70:30 Train-Test Split

Algorithms	Model-1(r2 score)	Model-1(MAE)
Logistic Regression	0.900	0.099
KNN	0.77	0.228
Decision Tree	0.898	0.101
Random Forest	0.93	0.119

60:40 Train-Test Split

Algorithms	Model-1(r2 score)	Model-1(MAE)
logistic Regression	0.900	0.995
KNN	0.77	0.228
Decision Tree	0.61	0.097
Random Forest	0.93	0.07

Algorithmic Comparison

Algorithms	Model-1(r2 score)	Model-1(MAE)
Logistic Regression	0.910	0.208
KNN	0.79	0.2079
Decision Tree	0.917	0.082
Random Forest	0.94	0.059

Summmary

- The purpose of this research is to determine the best-performing machine learning techniques to predict Insurance Premium

- Random Forest has the highest R^2 score (0.94) and the lowest MAE (0.059), indicating it performs the best among the four algorithms.of predictions in consideration to other regression models considered in this research
- Logistic Regression also performs relatively well with a high R^2 score (0.910), but its MAE (0.208) is higher than the others.

Thank You

Paluri Sathvik
Akshay Kumar
Khushi Bhansali
Pavan Prajapat