Insurance Charges Data Set – Supervised Machine Learning: Classification

Dataset Kaggle Link (https://www.kaggle.com/mirichoi0218/insurance).

The data consists of health insurance customers' information and the amount of charges they would pay for the insurance. A snippet of the first 5 rows is shown below:

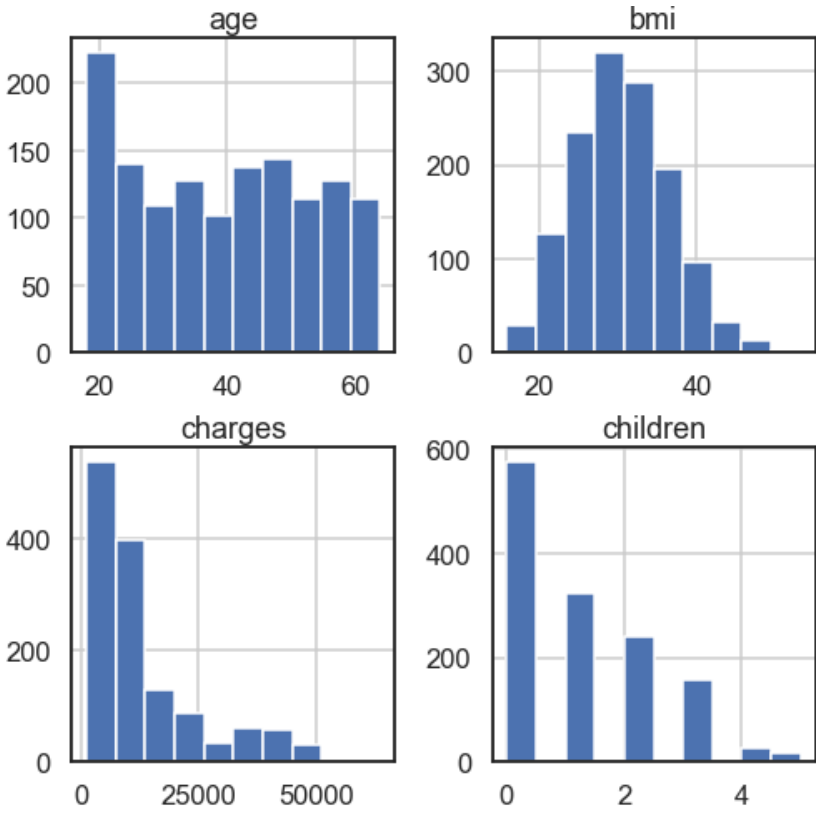
	age	sex	bmi	children	smoker	region	charges
0	19	female	27.900	0	yes	southwest	16884.92400
1	18	male	33.770	1	no	southeast	1725.55230
2	28	male	33.000	3	no	southeast	4449.46200
3	33	male	22.705	0	no	northwest	21984.47061
4	32	male	28.880	0	no	northwest	3866.85520

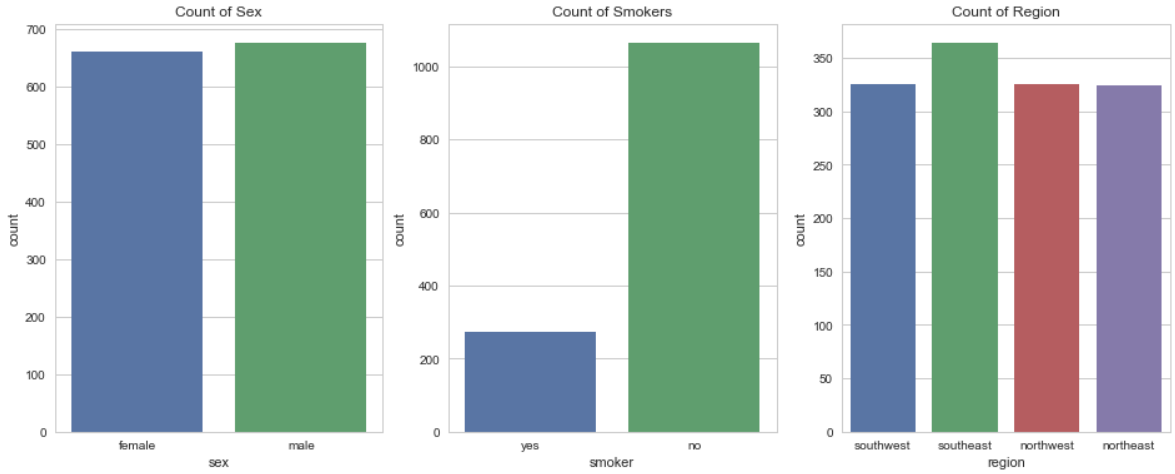
The data taken from Kaggle was clean and no values were missing. It consists of **1338 rows,** i.e., customers. A description of the features is given below:

- Age: Age of customer.
- Sex: Insurance customer's gender: Female or Male.
- bmi: Body mass index.
- Children: Number of children covered by health insurance / Number of dependents.
- Smoker: If Insurance customer smokes: Yes or No.
- Region: The customer's residential area: Northeast, Southeast, Southwest, or Northwest.
- Charges: cost of the health insurance

The aim of the classification model will be to **predict** whether a customer is a smoker or not based on customer information and charges, NOT interpretation.

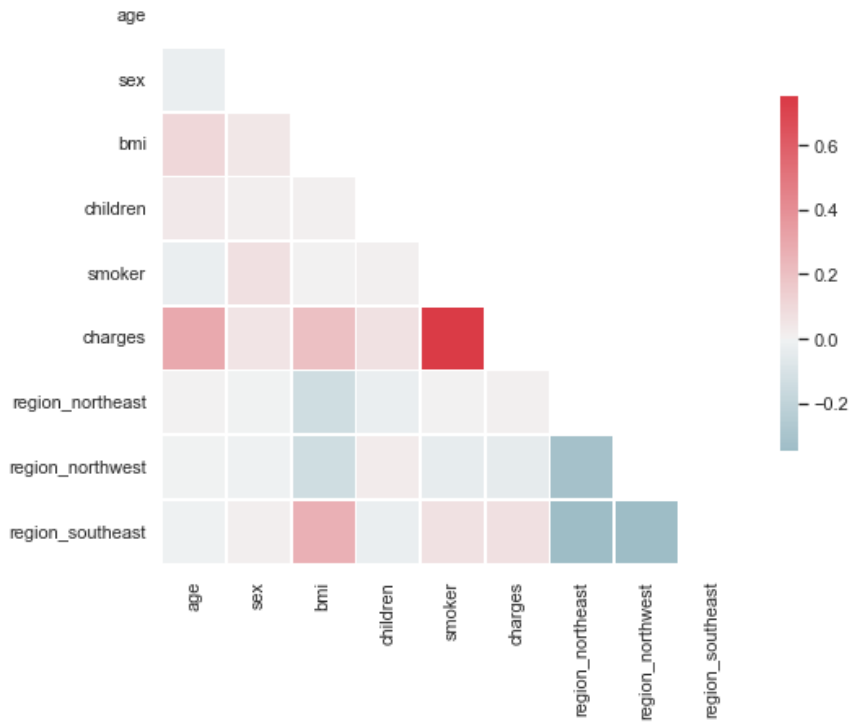
Distribution of Data





Since the proportion of smokers and non-smokers is not equal, stratified splitting will need to be done.

Correlation of Features



Smoker feature seems to have the highest correlation with the amount of charges which may make sense because a person paying a high health insurance charge may have health issues due to smoking. Although no other health information is available in the dataset to support this claim.

Feature Engineering

Categorical features (sex, smoker, region) were converted to numerical using Label Encoding:

- Sex: Male = 1, female = 0
- Smoker: Yes = 1, No =0
- Region: one-hot encoding

Age, charges and BMI features were scaled down using MinMax scaling. After these steps, the data set looked as follows,

	age	sex	bmi	children	smoker	charges	region_northeast	region_northwest	region_southeast
0	0.021739	0	0.321227	0	1	0.251611	0	0	0
1	0.000000	1	0.479150	1	0	0.009636	0	0	1
2	0.217391	1	0.458434	3	0	0.053115	0	0	1
3	0.326087	1	0.181464	0	0	0.333010	0	1	0
4	0.304348	1	0.347592	0	0	0.043816	0	1	0

Classification Models

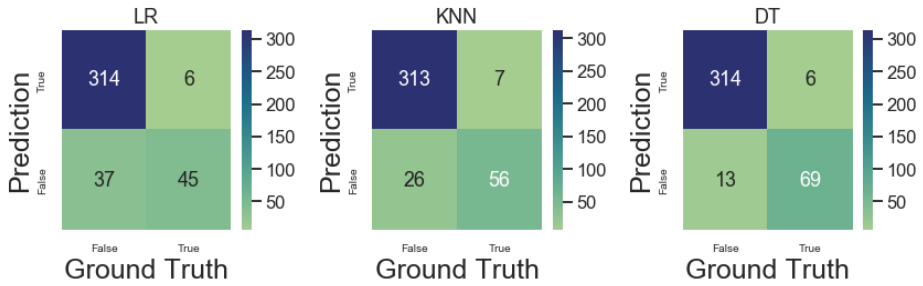
A total of 3 **models** were explored after splitting the data into **stratified training and testing** sets (70:30).

The models were as follows:

- Logistic Regression
- KNN with n = 3
- Decision Trees

The models were compared using precision, recall, accuracy, F1 score and AUC. The confusion matrix of each model is also shown.

	precision	recall	accuracy	f1score	auc
Logistic Regression	0.89	0.89	0.89	0.68	0.77
KNN	0.92	0.92	0.92	0.77	0.83
Decision Tree	0.95	0.95	0.95	0.88	0.91



Clearly, Decision Tree model performs the best in every metric. This model was found to have 73 nodes and a max depth of 12.

Conclusion

In conclusion, the Decision Tree model turned out to be the best of the chosen models. Soe more models such as Random Forests can also be explored to find better fitting models. Esemble models and boosting methods could also be explored. Furthermore, interpretation of results can also be done.