Project Report of
Auto Insurance Company Prediction

Research Scenario-

Here the dataset of an auto insurance company is taken for analysis. We analyze the data on various data fields such as credit score, size of family, vehicle type and visualize the patterns in the process. This research can answer the questions like which customer is more likely to claim an insurance in the future and which customer would not be able to claim the insurance given the credit score. It will help the insurance company to get an insight of their customer's insurance needs and they can further modify their insurance policy such that they both profit from it. Further the research can also provide an insight on which vehicle type the insurance was claimed the most. Finally, the research scenario will use data science and machine learning models to classify the data and come to a certain conclusion which can be used by the company to forecast the insurance claims of their customer and further build a strategy for the same.

Problem Statement- As an auto insurance owner, I want to be able to predict which customers are more likely to claim an insurance in the following years. Given the customer data, I want to find out which machine learning models turn out to be the most accurate for prediction.

Dataset Description & Preview-

The data taken here is medium size data of an Auto Insurance company. It has fields like –

a) Credit score- It is an important data field as insurance company maps the customer with their credit score availability.

b) Vehicle type- This data field gives the customer's car model such as Car, Van, Utility, Truck.

c) Size of Family- It gives the customer's family members size which will be used by the auto insurance company.

d) Age- Simple data field which will map the customer to their age column. It can be anywhere from 18 to 60 years old.

e) Gender- Simple data field which will map the customer to their gender category.

There are more data fields like engine HP etc in the data set which will give more insight on customer's details.

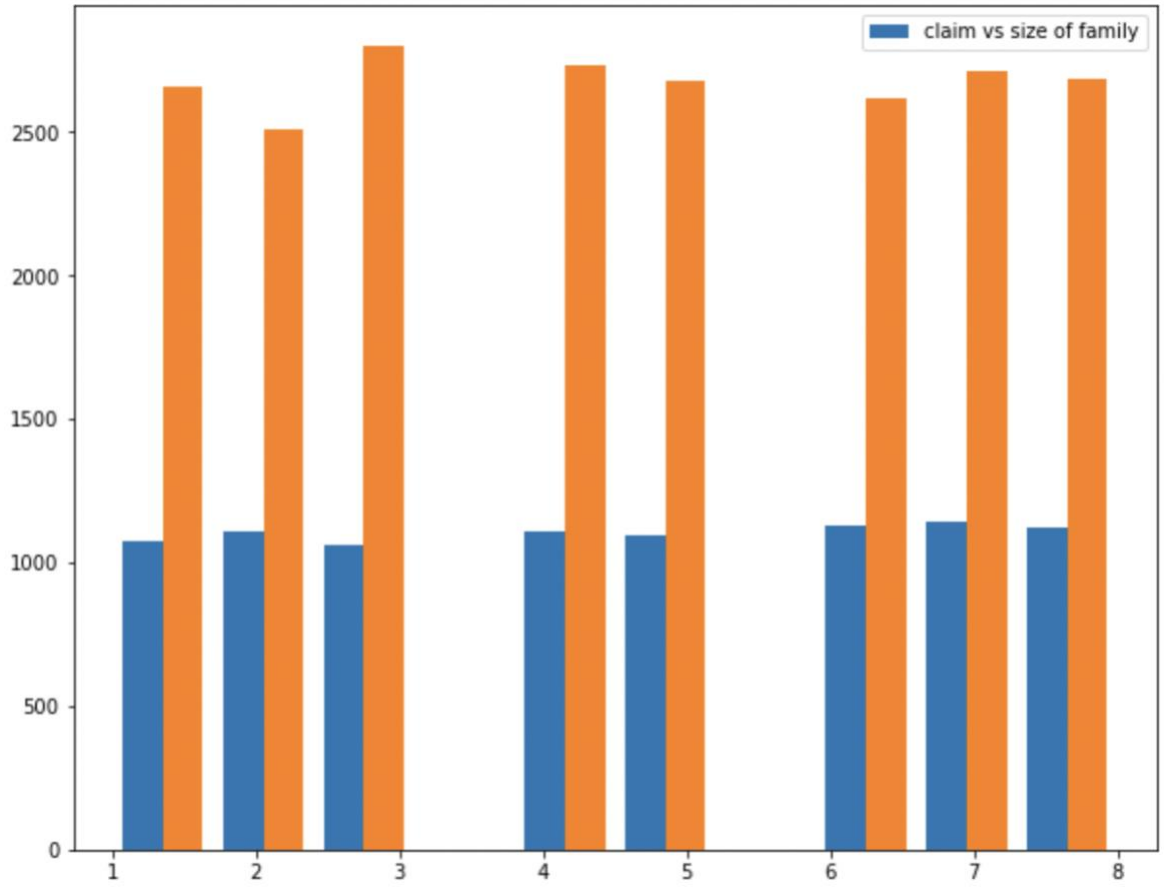
Preview-

ID	target	Gender	EngineHP	credit_history	Years_Experience	annual_claims	Marital_Status	Vehical_type	Miles_driven_annually	size_of_family	Age_bucket
1	1	F	522	656	1	0	Married	Car	14749.0	5	<18
2	1	F	691	704	16	0	Married	Car	15389.0	6	28-34
3	1	M	133	691	15	0	Married	Van	9956.0	3	>40
4	1	M	146	720	9	0	Married	Van	77323.0	3	18-27
5	1	M	128	771	33	1	Married	Van	14183.0	4	>40

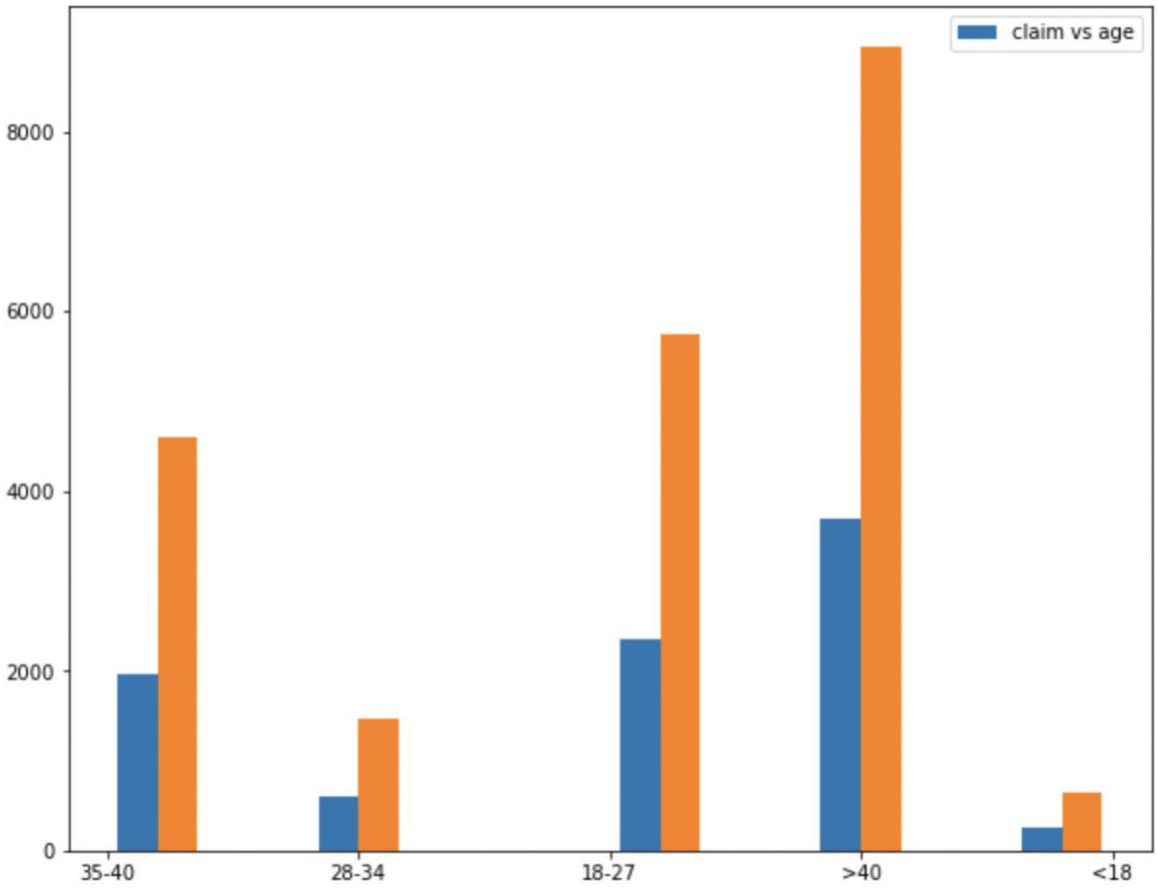
EngineHP_bucket	Years_Experience_bucket	Miles_driven_annually_bucket	credit_history_bucket	State
>350	<3	<15k	Fair	IL
>350	15-30	15k-25k	Good	NJ
90-160	15-30	<15k	Good	CT
90-160	9-14'	>25k	Good	CT
90-160	>30	<15k	Very Good	WY

Data Visualization-

1) Claims vs Size of family is visualized first to get the understanding between them.



2) Claim vs Age is further visualized.



Machine Learning models used for analysis-

1. K-Nearest Neighbor (KNN)
2. Random Forest
3. Decision Tree
4. Logistic Regression
5. Support Vector Machine (SVM)
6. Naive Bayes

Details of the learning models-

1. K-Nearest Neighbor (KNN)-

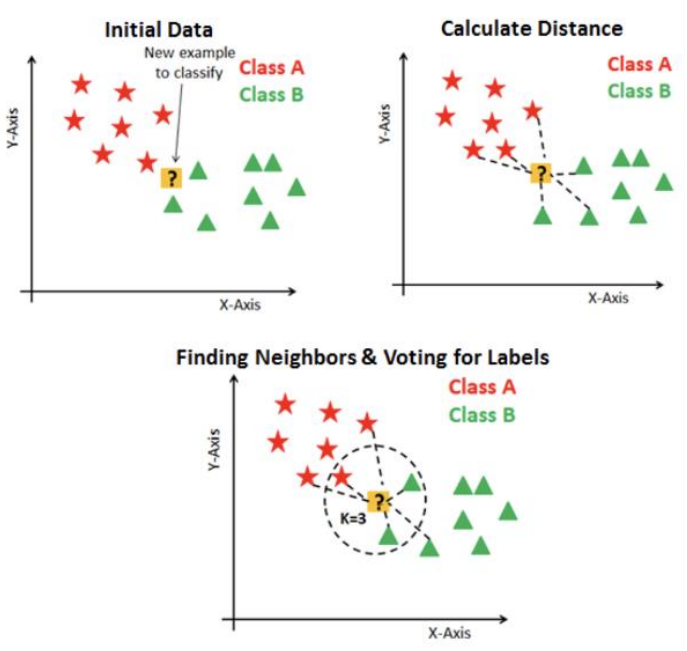
The KNN algorithm assumes that similar things exist in proximity. In other words, similar things are near to each other.

A given object gets assigned to the class most common among its k nearest neighbors where k is a positive integer and typically small.

For my analysis, I chose k = 7, so the values having least distance from nearest 7 values will be taken.

KNN has three basic steps:

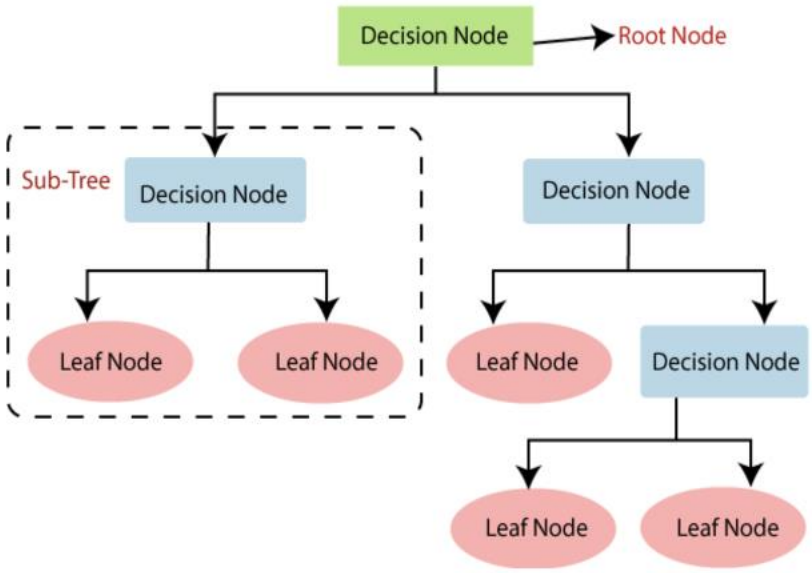
1. Calculate the distance
2. Find the k nearest neighbors
3. Vote for classes



2. Decision Tree-

A Decision Tree is a tree-like graph with nodes representing the place where we pick an attribute and ask a question, edges represent the answers to the questions asked, and the leaves represent the actual output or class label.

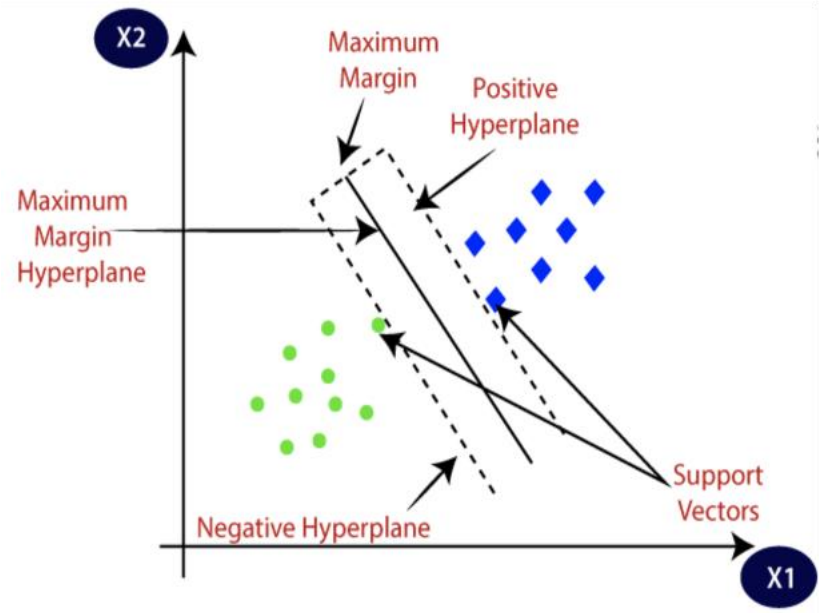
Based on the given features, I used a decision tree with a max depth of 10 to classify whether a given customer will claim insurance or not.



3. Support Vector Machine (SVM)-

The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes to easily classify data points into correct categories. This best decision boundary is called a hyperplane.

I used a Linear SVM with C=500, where ‘C’ tells the SVM optimization how much you want to avoid misclassifying each training example. For large values of C, the optimization will choose a smaller-margin hyperplane.



Accuracy Results-

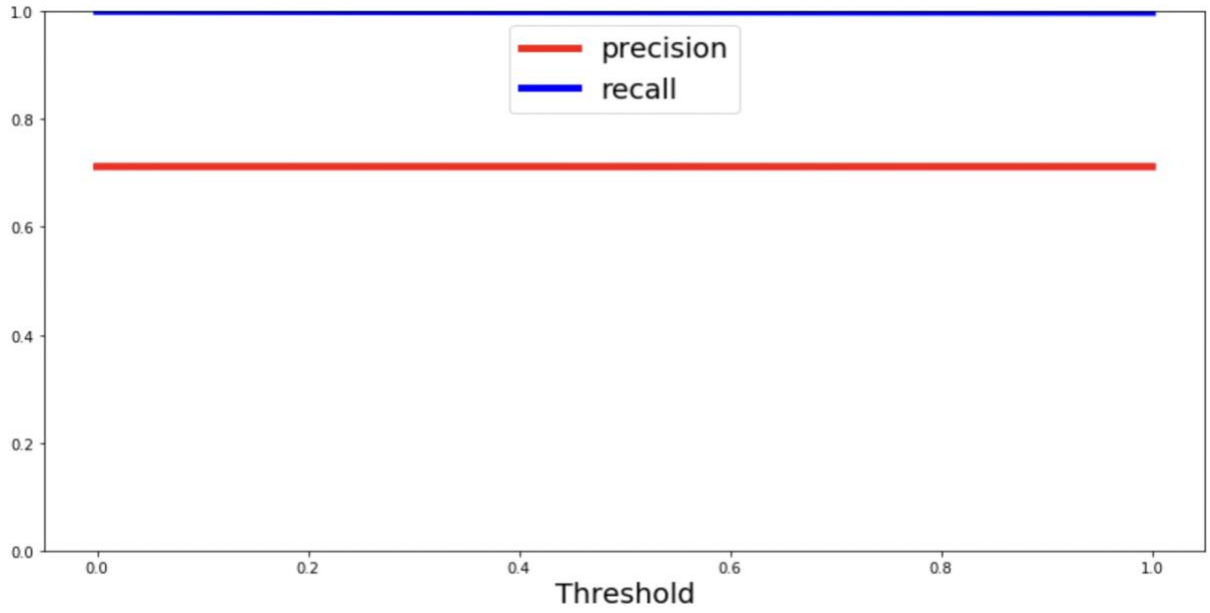
```
--------------------Random Forest trained--------------------
--------------------Logistic Regression trained--------------------
--------------------KNN trained--------------------
--------------------Naive Bayes trained--------------------
--------------------Support Vector Classifier trained--------------------
--------------------Decision Tree trained--------------------
              Model
Score
72.78           KNN
71.37  Decision Tree
70.92  Random Forest
70.64           SVM
70.64        Log Reg
70.64    Naive Bayes
```

Classification Report-

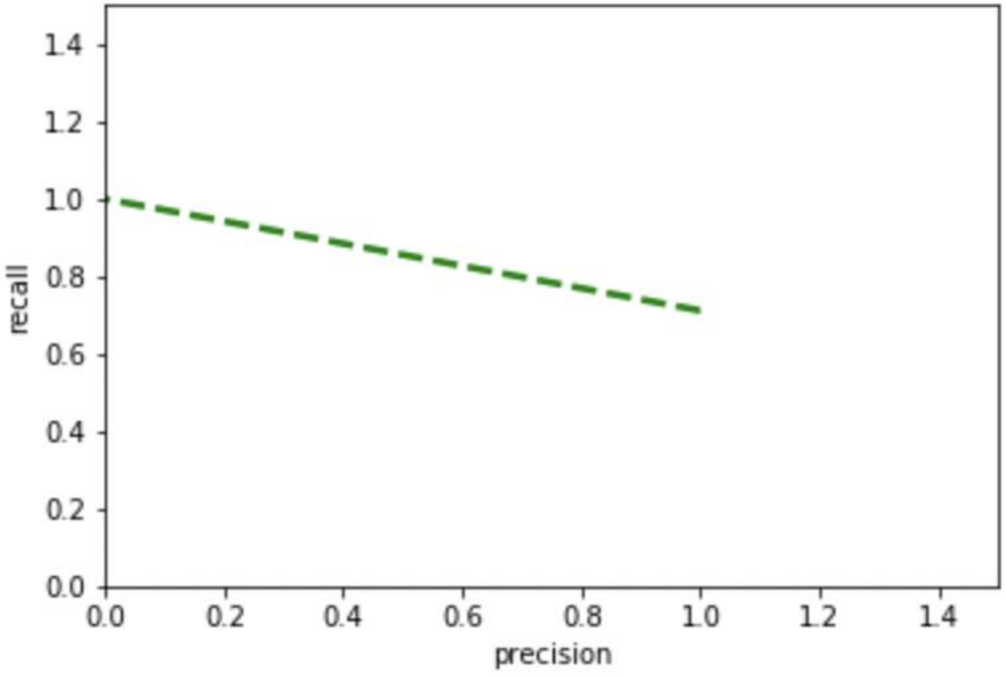
	precision	recall	f1-score	support
0	0.20	0.00	0.00	1741
1	0.71	1.00	0.83	4306
accuracy			0.71	6047
macro avg	0.46	0.50	0.42	6047
weighted avg	0.56	0.71	0.59	6047

```
True Positive : 4298
True Negative : 2
False Positive : 1739
False Negative : 8
Specificity : 0.0011487650775416428
Sensitivity : 0.9981421272642824
```

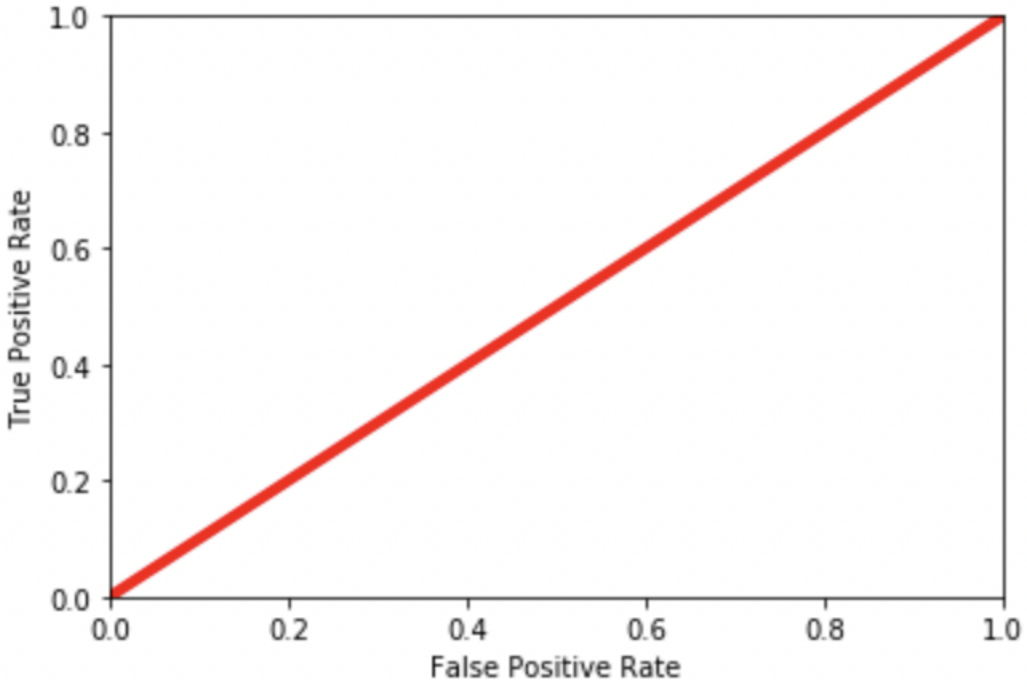
Threshold-



Precision vs Recall-



TP vs FP Rate-



Conclusion-

After using different machine learning models like KNN, Random Forest, Decision Tree, Naïve Bayes, Logistic Regression and Support Vector Machine we found **KNN has best accuracy rate with 73.12% for our given data**.

Notably, Decision tree has second best accuracy with 71.37%

Therefore, using K nearest Neighbor algorithm the company can predict which customer is more likely to claim the insurance.