**Design and Application of a Machine**

**Learning System for a Practical Problem**

**Pilot-Study Proposal**

**By**

**Smruti Das**

**Overview:**

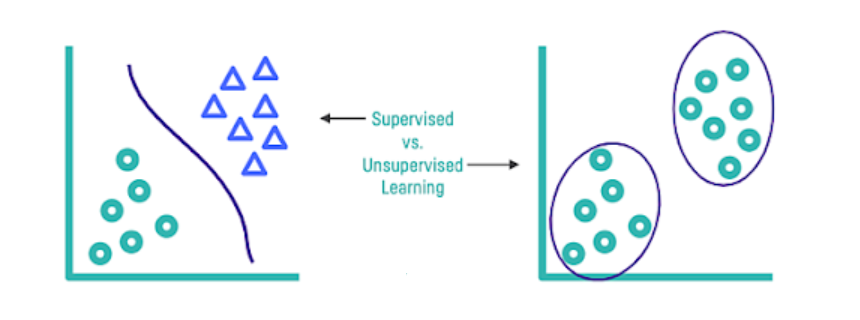
With increase in travel across and within countries, travel insurance claim has also increased in the past few decades. As there are many competitions in the market that has many deals for the customer, giving a premium discount to the customer could give us an edge. Based on the history data, we need to predict if a customer would claim in the future or not, based on which we can provide offers. This problem is an appropriate machine learning problem. It is clearly a binary classification problem.

**Proposal:**

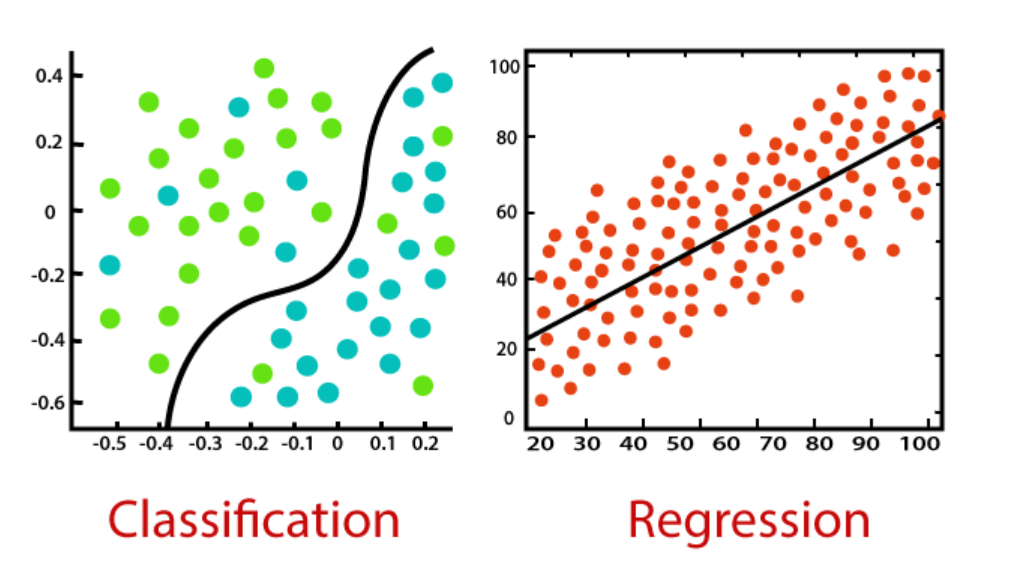
**1.1. Type of predictive task**

In Machine learning the type of predictive task that must be used depends on two aspects

* **Inputs:** If it is a labeled data, it’s a supervised learning problem. If it’s unlabelled data with the purpose of finding structure or clusters, it’s an unsupervised learning problem. If we want to optimize an objective function by interacting with an environment, it’s a reinforcement learning problem. As this is a labeled Data it is Supervised Learning



* **Outputs:** If the output of the model is a number, it’s a regression problem. If the output of the model is a class, it’s a classification problem. In this, we want to predict if the customer will claim insurance or not. Hence, it is a binary classification problem.



**1.2. Features:**

Selection and availability of correct features play a vital role in the performance of the machine learning algorithm. Below are a few features which I think are needed to make the prediction.

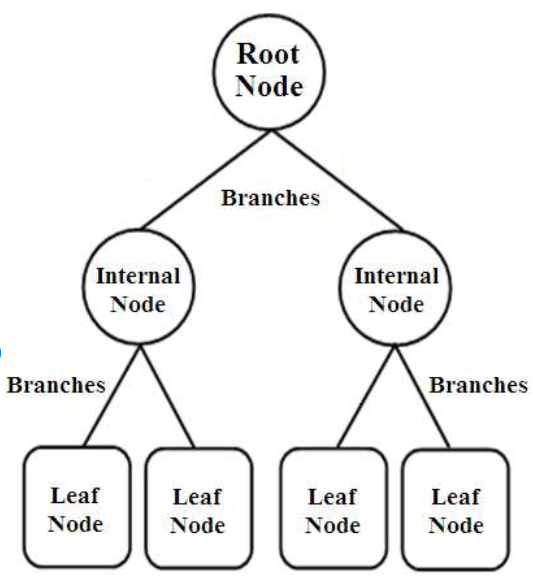
* Destination of travel
* Origin of travel
* Duration of travel
* Travel Date, Weekday
* If the travel was during any holiday
* Travel insurance type.
* Number of people insured
* Total cost of the insurance
* Insurance claim limit
* Gender of insured (Gender)
* Age of insured (Age)
* Origin country/region of customer

**1.3 Learning procedure**

As we have narrowed the problem to a **binary classification** problem, below Algorithms can be used.

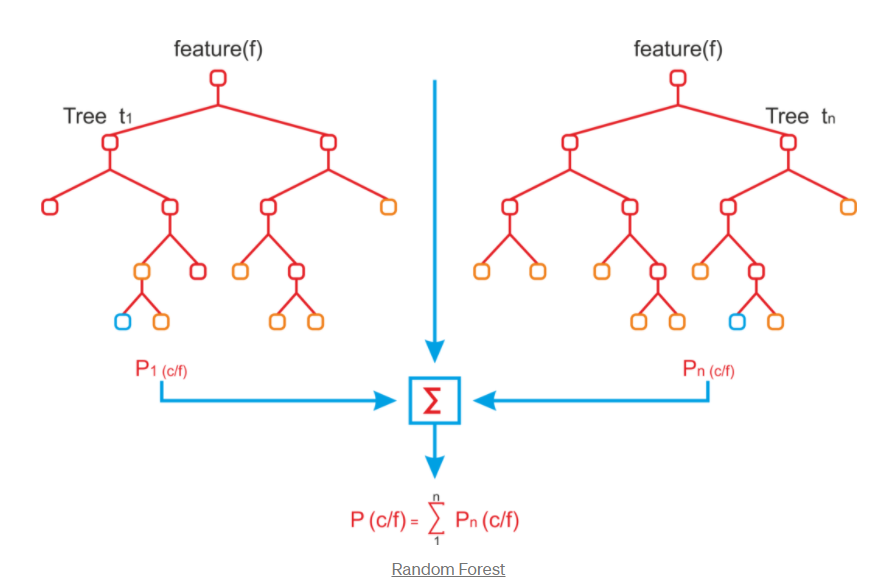
1. **Decision Tree:**

Decision tree is one of the easiest to implement and visualize Algorithms. It can be used both for classification as well as regression. However, as a Decision tree is a greedy algorithm it can easily overfit. Hence, we need to implement pruning in order to get the best results



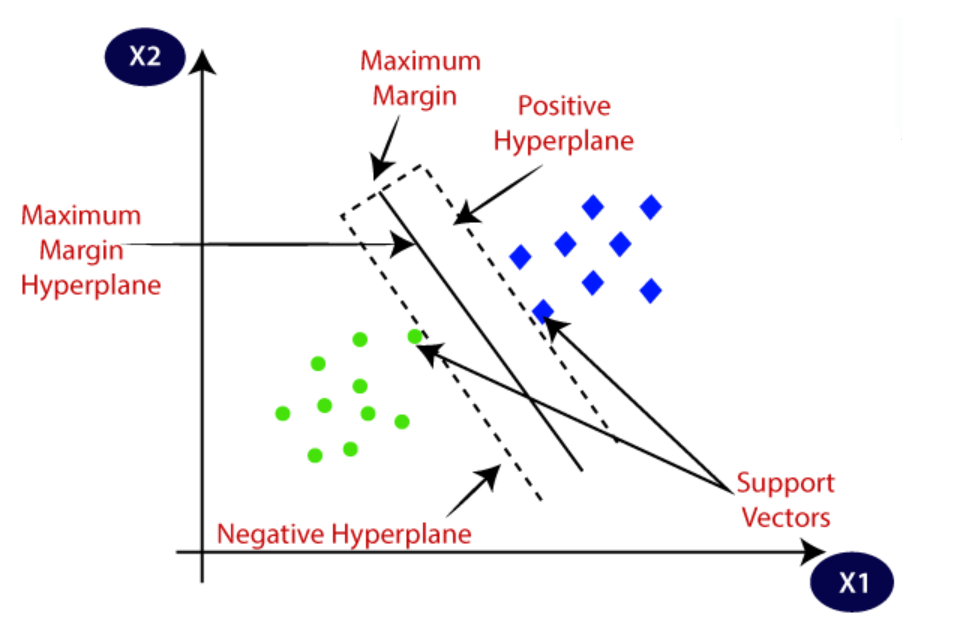
1. **Random Forest:**

Random forests or random decision forests are an ensemble learning method for classification & regression that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean/average prediction (regression) of the individual trees.



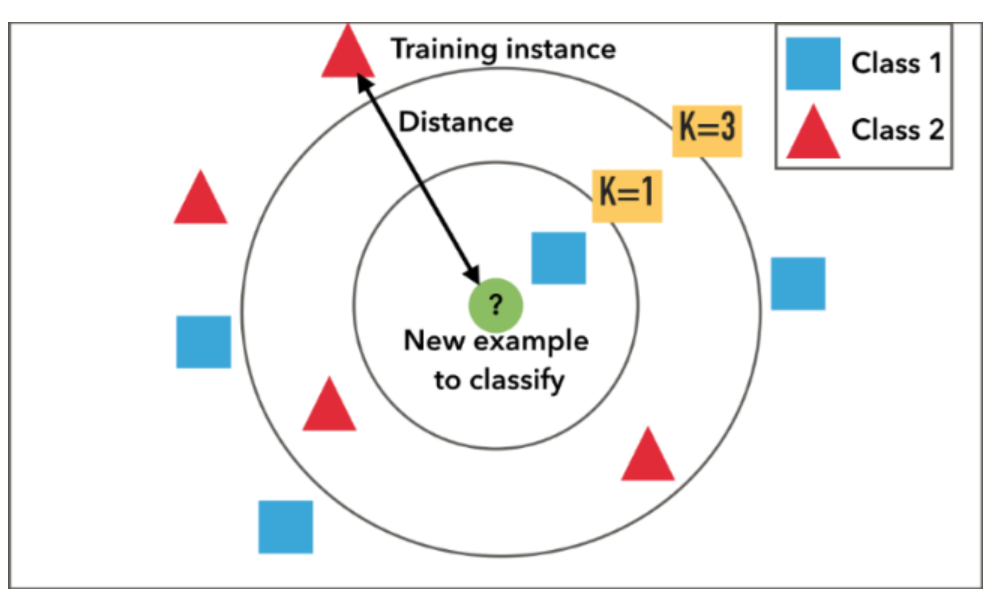
1. **Support Vector Machine**

SVM is efficient in high dimensional spaces. Support vectors are the data points that lie within or on the margins or mis-classified points. It is quiet memory efficient.



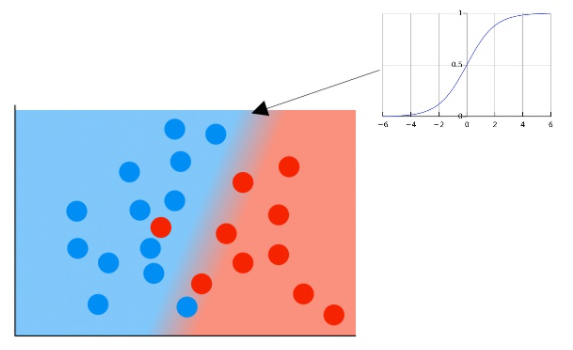
1. **K-NN**

*k*-NN is a type of instance-based learning which is easy to implement. Weighted K-NN can be used if the distance of the neighbours is to be taken into consideration. However, K-NN doesn’t give best results in higher dimension data.



1. **Logistic Regression**

Logistic regression works best for binary classification problem. It uses a logistic function like sigmoid to separate the classes

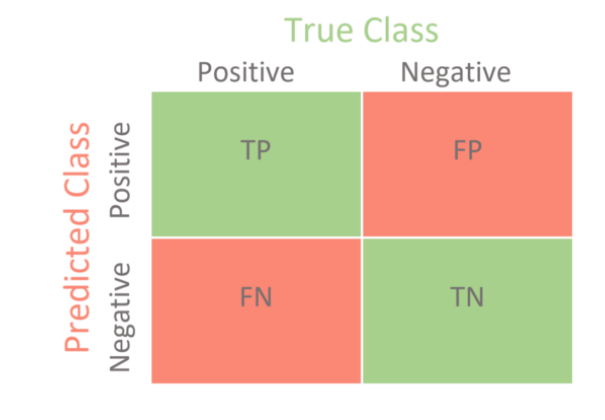


**1.4 Evaluate the performance**

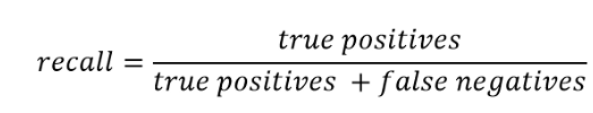
These algorithms (mentioned above) can be evaluated on various measures. Some are listed below

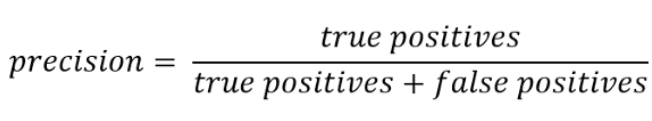
* **Confusion matrix:**

A confusion matrix of binary classification is a two by two table formed by counting of the number of the four outcomes of a binary classifier.



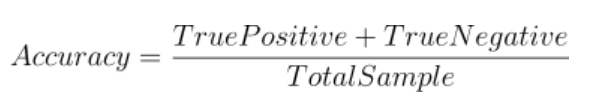
There are other matrices which can be derived from confusion matrix which are useful.





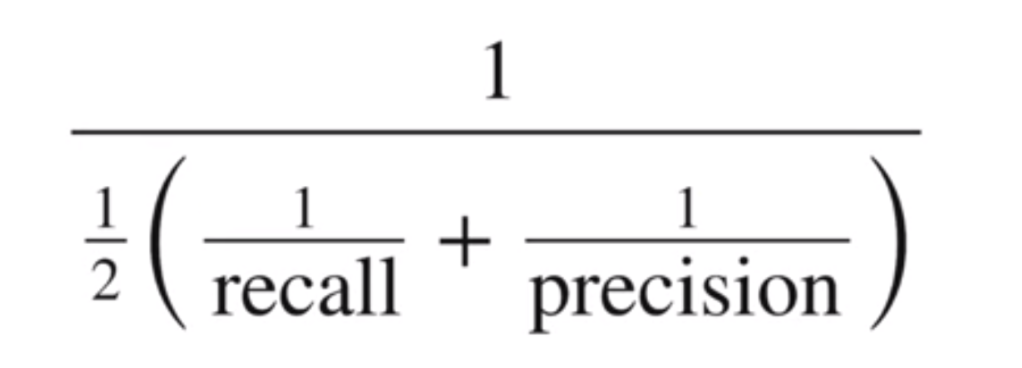
* **Accuracy:**

Accuracy (ACC) is calculated as the number of all correct predictions divided by the total number of the dataset. The best accuracy is 1.0, whereas the worst is 0.0.



* **F1 Score**

F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account.



* **Kappa statistics:**

[Cohen’s kappa](https://en.wikipedia.org/wiki/Cohen%27s_kappa) statistic is a very good measure that can handle very well both multi-class and imbalanced class problems.

**Conclusion:**

Using above Machine Learning Algorithms and evaluation techniques, we can predict if the customer would claim for not, given we have enough and much needed data to train the model.

**References:**

* Scikit-learn.org. 2021. *Sklearn.Metrics.Accuracy\_Score — Scikit-Learn 0.24.0 Documentation*. [online] Available at: <https://scikit-learn.org/stable/modules/generated/sklearn.metrics.accuracy_score.html>
* Scikit-learn.org. 2021. *Sklearn.Metrics.confusion\_matrix — Scikit-Learn 0.24.0 Documentation*. [online] Available at: <https://scikit-learn.org/stable/modules/generated/sklearn.metrics.confusion_matrix.html>
* Scikit-learn.org. 2021. *Sklearn.Metrics.DecisionTreeClassifier— Scikit-Learn 0.24.0 Documentation*. [online] Available at: <https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html>
* Scikit-learn.org. 2021. *Sklearn.Metrics.SVM — Scikit-Learn 0.24.0 Documentation*. [online] Available at: <https://scikit-learn.org/stable/modules/svm.html>
* Mitchell, T. M., (1997),“Machine Learning”,McGraw-Hill. Chapter 3. Tan, Steinbach & Kumar (2006) “Introduction to Data Mining”.
* Exsilio Blog. 2021. *Accuracy, Precision, Recall & F1 Score: Interpretation Of Performance Measures - Exsilio Blog*. [online] Available at: <https://blog.exsilio.com/all/accuracy-precision-recall-f1-score-interpretation-of-performance-measures/>
* Medium. 2021. *Do You Know How To Choose The Right Machine Learning Algorithm Among 7 Different Types?*. [online] Available at: <https://towardsdatascience.com/do-you-know-how-to-choose-the-right-machine-learning-algorithm-among-7-different-types-295d0b0c7f60>