ML PROJECT BASED ON CLASIIFICATION OF JOB POSTING

# 33131- Mohammed Husain

# 33150- Rajvardhan Mane

# 33153-Atharva Nagare

## PROBLEM STATEMENT OF ML PROJECT

1. Performing data cleaning on a data set, which classifies a job posting into fraudulent or not fraudulent, in a way such that accuracy of classification model increases.
2. Applying 2 classification algorithms namely, SVM (support vector machine) and Random Forest classifier. Comparing the accuracy of both models to decide which one works better

## INTRODUCTION

Data cleaning:

Data cleaning is the process of preparing data for analysis by removing or modifying data that is incorrect, incomplete, irrelevant, duplicated, or improperly formatted. It includes:

* Fixing spelling and syntax errors
* Standardizing data sets
* Correcting mistakes such as empty fields
* Identifying duplicate data points

Why is data cleaning important?

Businesses have a plethora of data. But not all of it is accurate or organized. When it comes to machine learning, if data is not cleaned thoroughly, the accuracy of your model stands on shaky grounds.

Preparing/cleaning data helps maintain quality and makes for more accurate analytics, which increases effective, intelligent decision-making.

Benefits of cleaning data:

* Better decision making
* Save time
* Increase productivity
* Better accuracy

Classification:

Classification is a process of categorizing a given set of data into classes, it can be performed on both structured or unstructured data. The process starts with predicting the class of given data points. The classes are often referred to as target, label or categories.

The classification predictive modelling is the task of approximating the mapping function from input variables to discrete output variables. The main goal is to identify which class/category the new data will fall into.

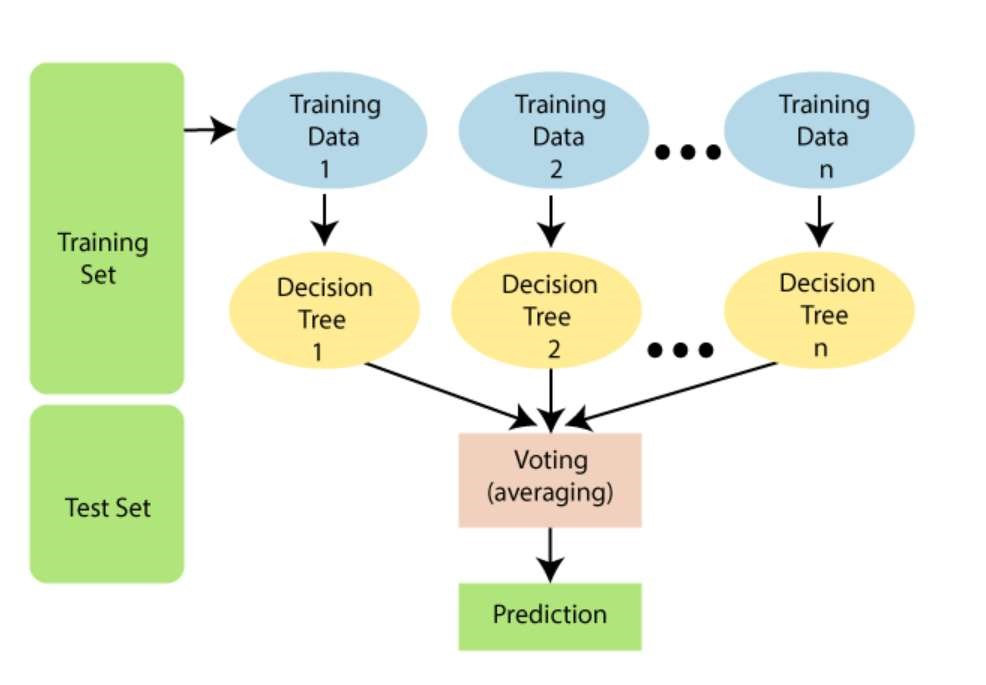
The 2 classification methods used are SVM and random forest.

# RANDOM FOREST

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.

Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset. Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.

The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

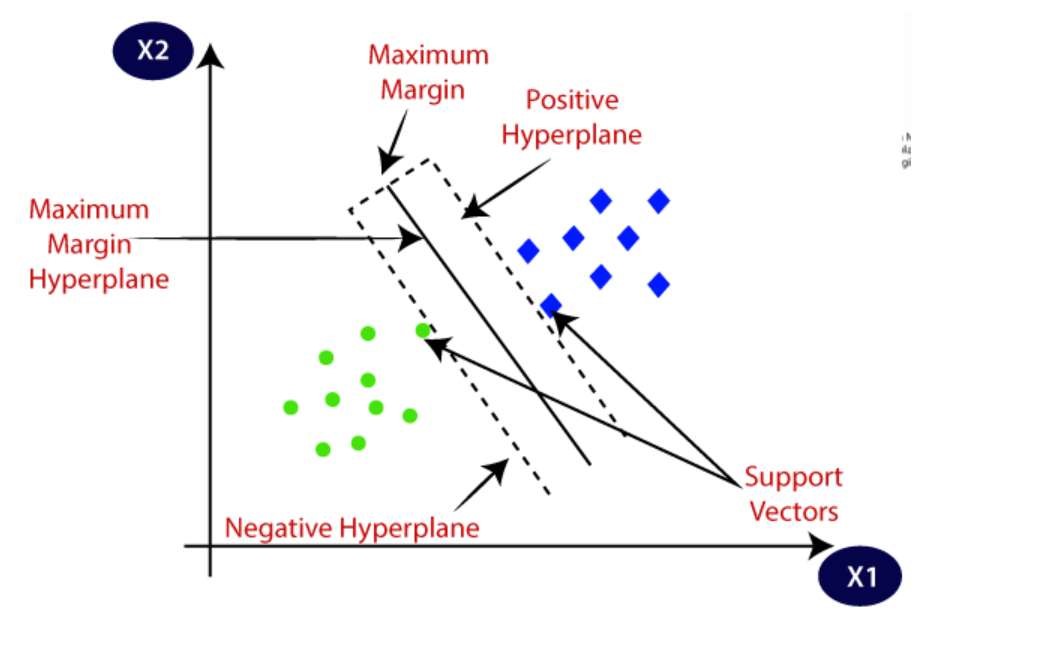


## SVM

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning.

The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.

SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine. Consider the below diagram in which there are two different categories that are classified using a decision boundary or hyperplane

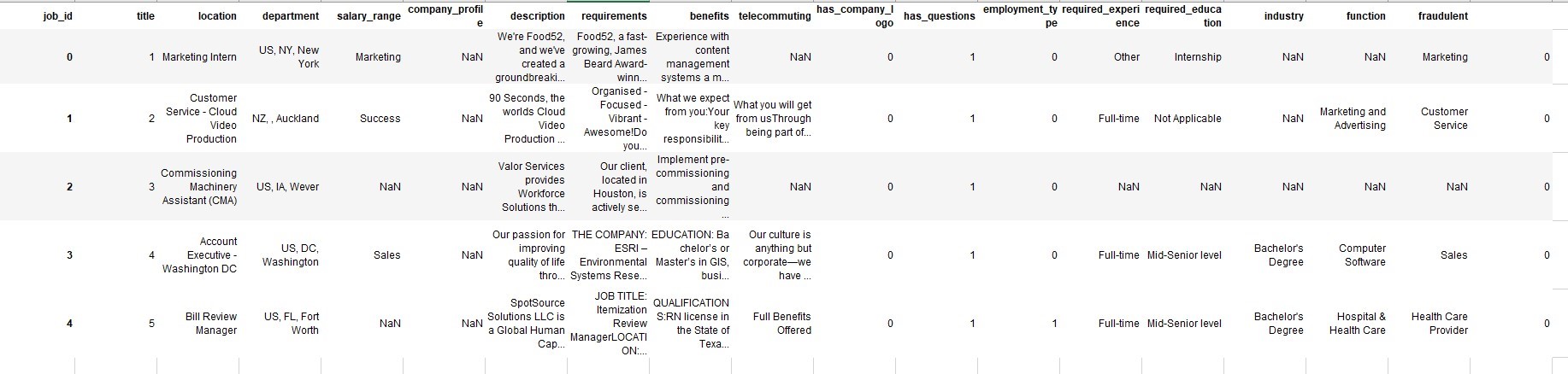


### DATASET DETAILS

The dataset includes various attributes/factors on the basis of which a job posting is classified into either fraudulent or non-fraudulent.

https://www.kaggle.com/datasets/shivamb/real-or-fake-fakejobposting-prediction?resource=download

Name of dataset - fake\_job\_postings.csv



CODE AND OUTPUT

Displayed in separate pdf file present in ZIP file

### CONCLUSION

We see that the accuracy of both models is very high after performing data cleaning.

We see that the accuracy of model using SVM is 91% and accuracy of model using random forest is 93% So we can conclude that random forest works better on our data-set

### REFERENCES

https://www.javatpoint.com/machine-learning-random-forestalgorithm

https://www.javatpoint.com/machine-learning-support-vectormachine-algorithm