STEVENS
INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY

Prediction of Termination by an Employee

CS-513 : Knowledge Discovery & Data Mining
Fall 2021

Professor - Khasha Dehnad

Presented by: Group 3

STEVENS
INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®

Team Members :

- Aditya Shivankar - *10472887*
- Devendra K Chauhan - *10473877*
- Devika Mhatre - *10467698*
- Sudarshana Sarma - *10469063*

STEVENS
INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®

Agenda :

- Introduction
- Roadmap
- Dataset Description
- Exploratory Data Analysis
- Data Preprocessing
- Models
- Comparison
- Conclusion

STEVENS
INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®

Introduction

Overview :

- Every year, many employees terminate a organization based on numerous factors. We strongly believe that what we have learn in KDD can be used to analyse the data and predict

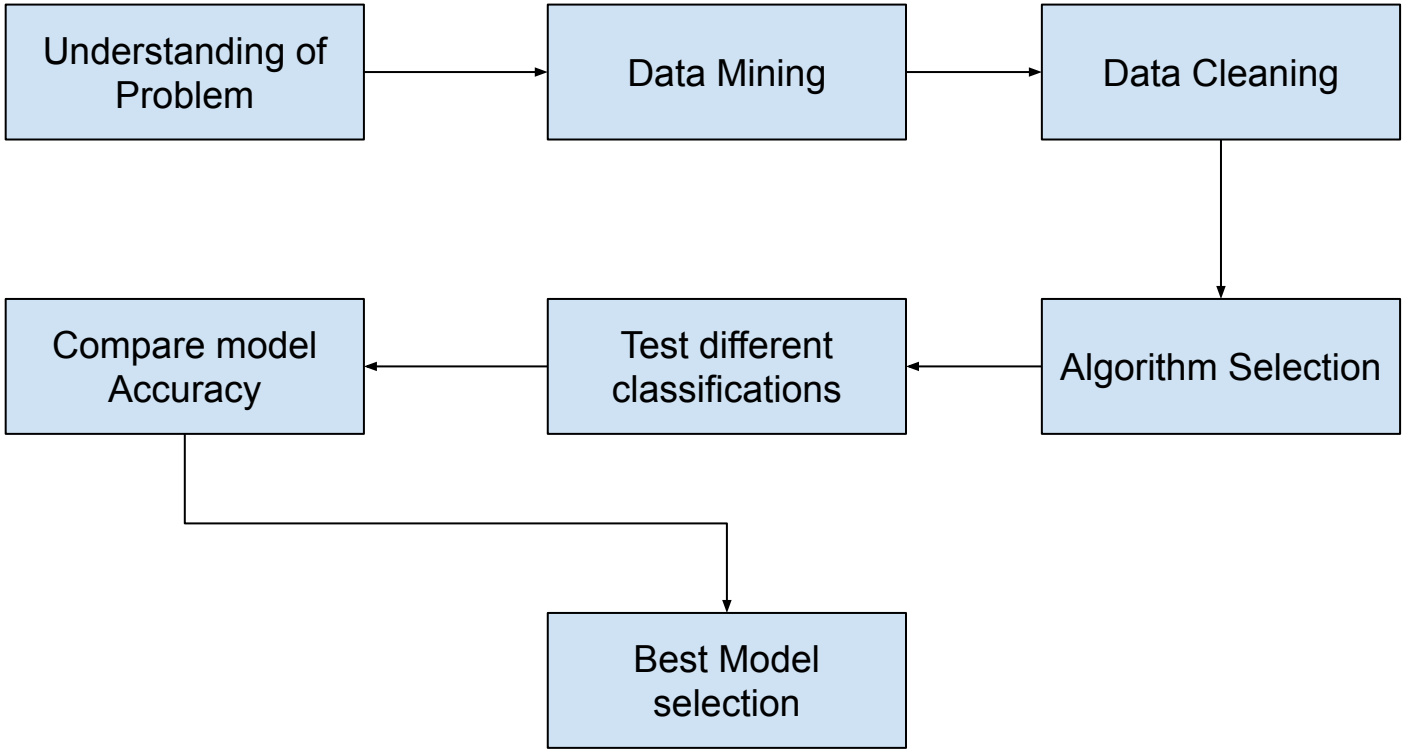
- Employee termination varies widely from employee to employee, and a deep understanding of various factors helps to predict the number of employees terminating

- Many factors contribute when employee wanted to terminate a company like including age, job satisfaction, hourly rate as well as performance rating

- The previous five years rating is an important factor in deciding the number of employees being terminated as rating is directly proportional to annual rate

Goal: This project predicts the number of an employee terminating a organization in a year based on various characteristics such as age, job satisfaction

Road Map :


Understanding of Problem
Data Mining
Data Cleaning
Algorithm Selection
Test different classifications
Compare model Accuracy
Best Model selection

Dataset Description :

The dataset contains almost 10,000 data entries of the employee till the year of 2017 and prior.

The dataset consists of 1 file as mentioned below which contain 27 column:

- attrition_data.csv

	EMP_ID	ANNUAL_RATE	HRLY_RATE	...	PREVYR_3	PREVYR_4	PREVYR_5
0	3285941608	33615	22	...	0	0	0
1	3687079832	70675	40	...	3	2	3
2	7209970080	34320	23	...	3	2	3
3	9084013977	103199	59	...	0	0	0
4	4566148978	141801	71	...	2	2	2

Exploratory Data Analysis

Exploratory Data Analysis :

Data Cleaning

EMP_ID	0
ANNUAL_RATE	0
HRLY_RATE	0
JOBCODE	0
ETHNICITY	0
SEX	0
MARITAL_STATUS	0
JOB_SATISFACTION	0
AGE	0
NUMBER_OF_TEAM_CHANGED	0
REFERRAL_SOURCE	445
HIRE_MONTH	0
REHIRE	0
TERMINATION_YEAR	5394
IS_FIRST_JOB	0
TRAVELLED_REQUIRED	0
PERFORMANCE_RATING	0
DISABLED_EMP	0
DISABLED_VET	0
EDUCATION_LEVEL	0
STATUS	0
JOB_GROUP	0
PREVYR_1	0
PREVYR_2	0
PREVYR_3	0
PREVYR_4	0
PREVYR_5	0

dtype: int64

Data Preprocessing :

- Initial data features columns including annual and hourly rates, ethnicity, age, sex, job group, first job, education level.
- Columns employee id, termination year, job code, and referral source are removed because they have missing data.
- Status, whether kept or terminated, is selected as the target column and factorized.

Table 1: Data before Processing

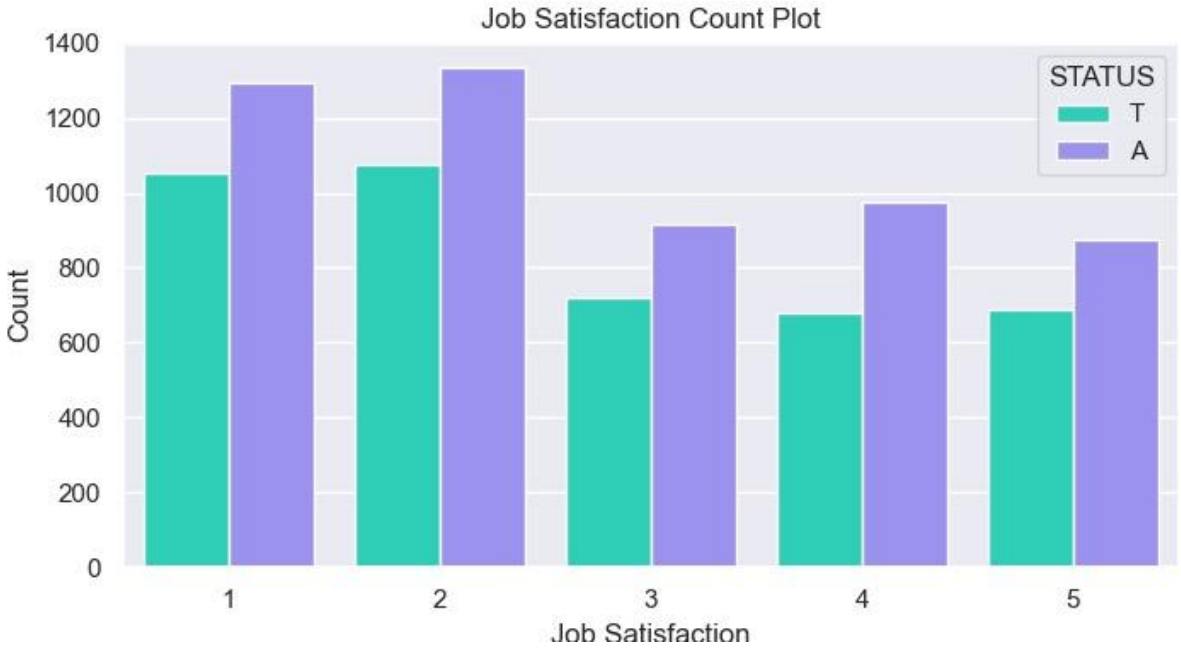
EMP_ID	ANNUAL_RATE	HRLY_RATE	JOBCODE	...	STATUS	JOB_GROUP	...
3285941608	33615	22	71850	...	T	Support	...
3687079832	70675	40	59806	...	A	Support	...
7209970080	34320	23	60311	...	A	Support	...
9084013977	103199	59	16233	...	T	Finance	...
4566148978	141801	71	64415	...	A	Marketing	...

- Some features were factored:
  - Annual rate is split based on $20,000, $50,000, $75,000, $100,000, and $2,000,000.
  - Hourly rate is split based on $25, $50, $75, $100, and $1000
  - Age was split based on 20, 30, 50, 60, 100.
  - Hire month was split based on months in Q1, Q2, Q3, and Q4
  - Ethnicity, sex, martial status, number of teams, first job, travel requirement, disabled, veteran, job group, and education were factorized

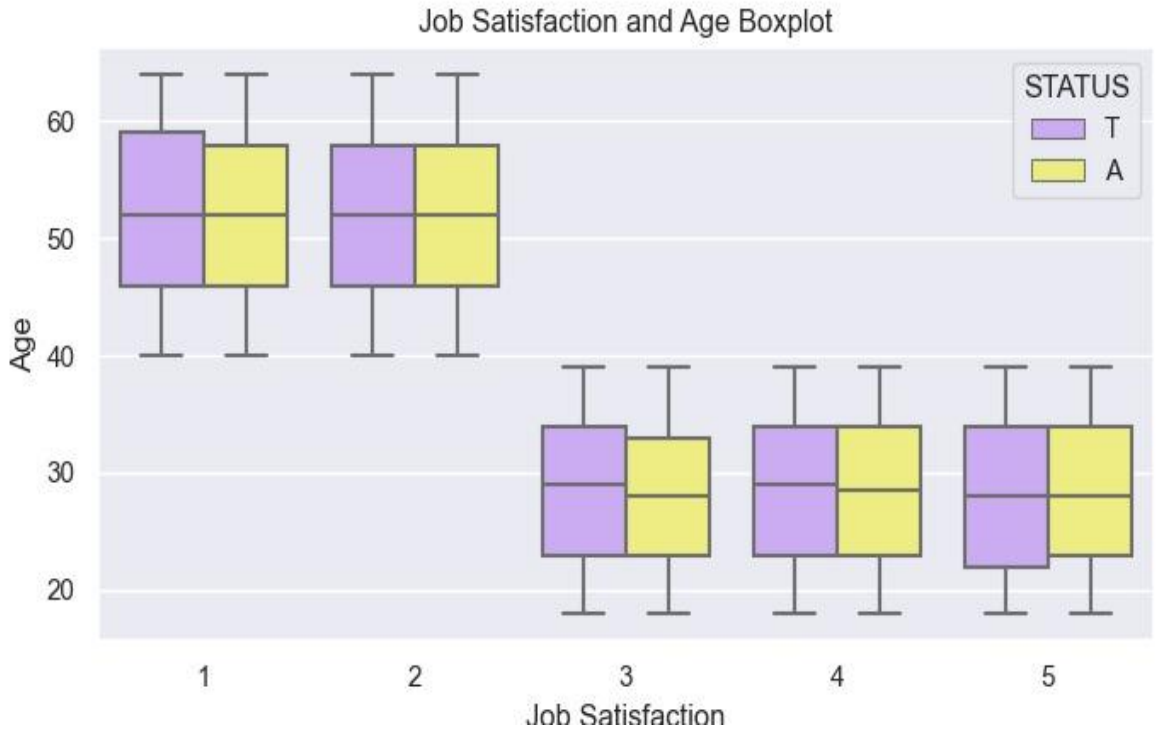
Table 2: Data after Processing

annual rate	hourly rate	ethnicity	sex	marital	satisfaction	...	education	...
2	1	0	0	0	4	...	0	...
3	2	1	1	1	3	...	1	...
2	1	2	0	1	5	...	1	...
5	3	1	0	1	2	...	1	...
5	3	1	0	1	4	...	1	...

Job Satisfaction Count Plot :

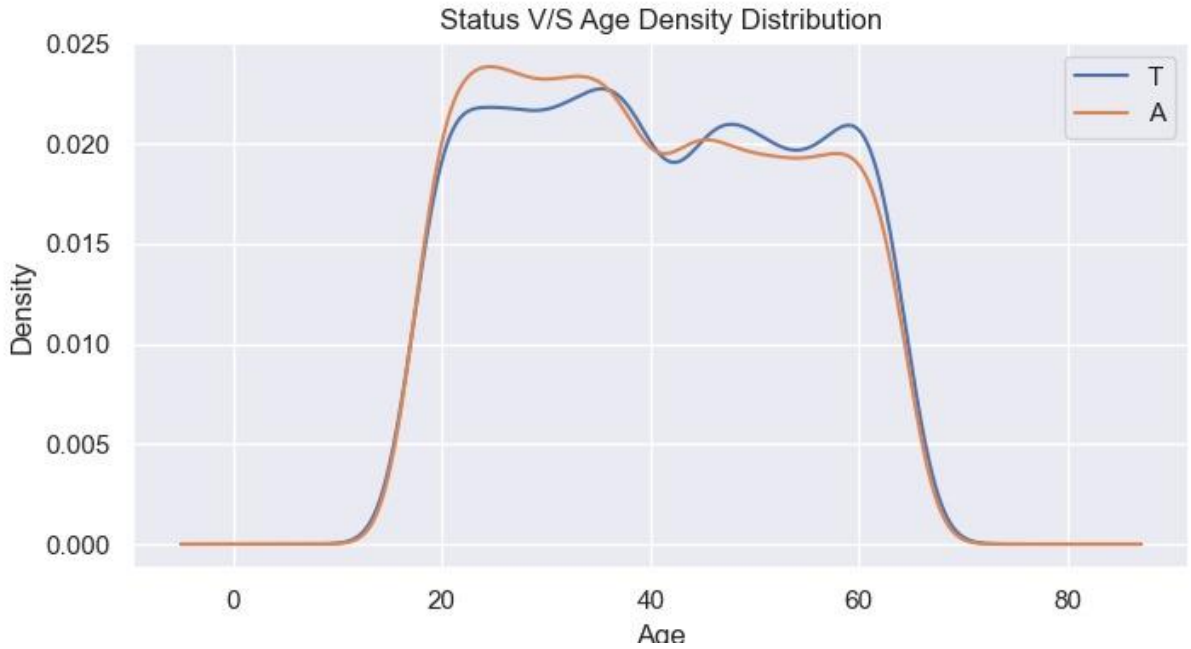


Job Satisfaction and Age Boxplot :

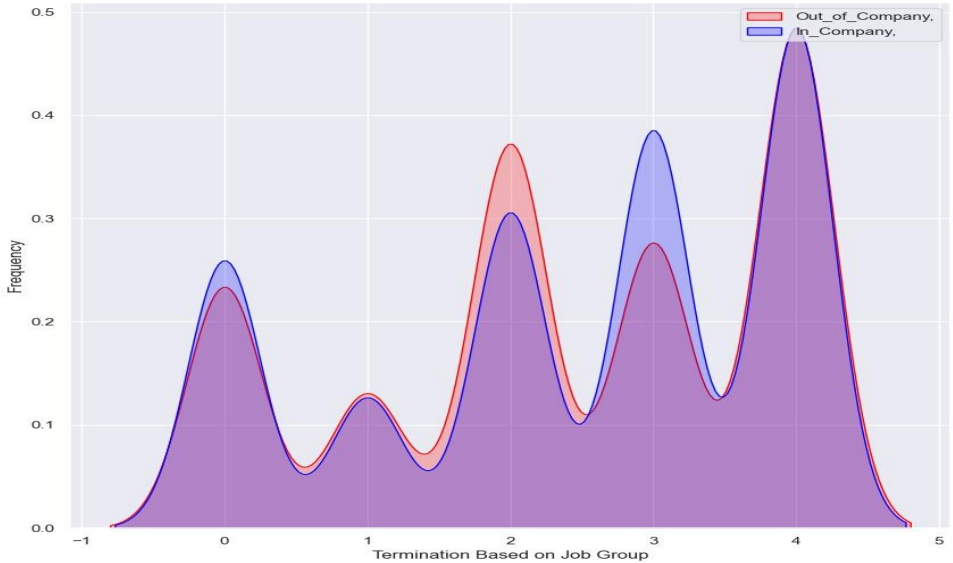


Performance Rating Count Plot :

Status V/S Age Density Distribution :



Termination based on job group :

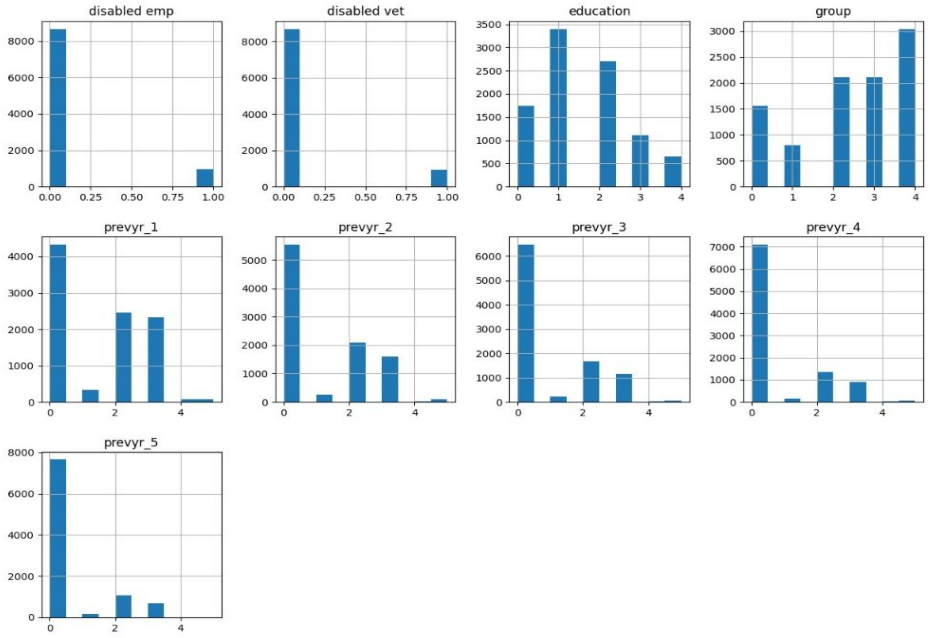


Exploration of Data :

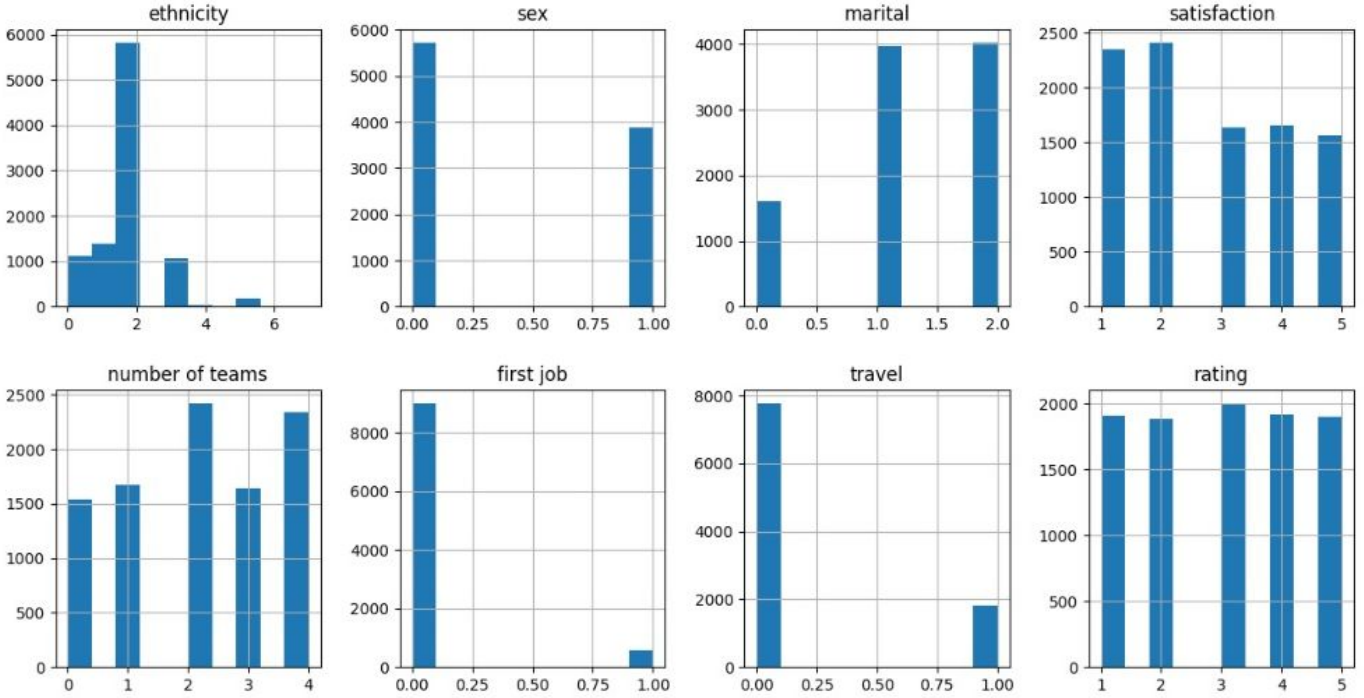
Table 3: Feature Correlation

Feature	Correlation
annual rate	0.178426
hourly rate	0.177944
ethnicity	-0.004023
sex	0.015205
marital	0.014455
satisfaction	0.015613
number of teams	-0.014027
hire month	0.001065
first job	0.005801
travel	-0.003475
rating	-0.001225
disabled emp	-0.008042
disabled vet	-0.003469
education	-0.018809
group	0.032679
prevyr_1	0.148430
prevyr_2	0.163136
prevyr_3	0.177668
prevyr_4	0.213362
prevyr_5	0.220317

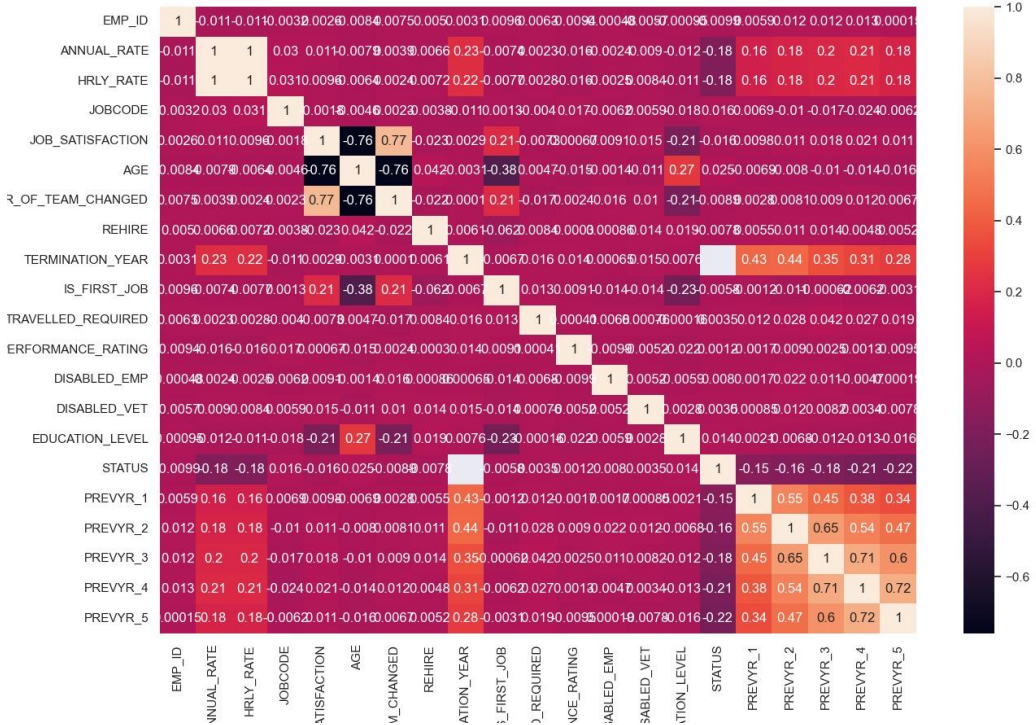
Examining the correlation between the different features and the status (target) of the employee.



Exploration of Data :



Feature engineering - Correlation heatmap :



Preparing for Modeling :

The data was split into 70% training and 30% test subsects.

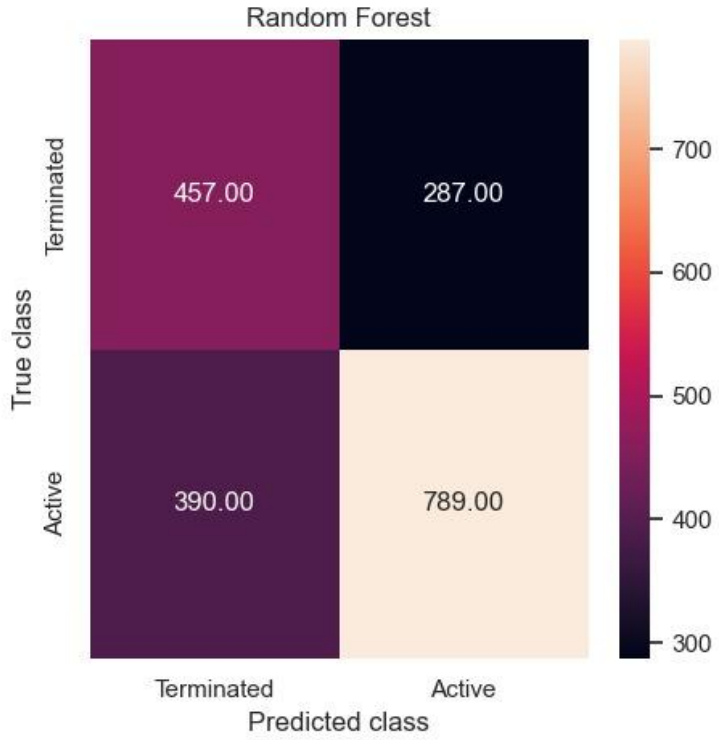
Table 4: Feature Selection

annual rate	hourly rate	sex	marital	age	disabled emp	group	prevyr_1	prevyr_4	prevyr_5
2	1	0	0	3	0	0	0	0	0
3	2	1	1	1	0	0	3	2	3
2	1	0	1	1	0	0	3	2	3
5	3	0	1	3	0	1	0	0	0
5	3	0	1	3	0	2	2	2	2

Random Forest :

- Random Forest classification was performed using the training data
- The test data was used to measure accuracy of the model.
- Random forest algorithm is a supervised classification and regression algorithm. This algorithm randomly creates a forest with several trees.
- Generally, the more trees in the forest the more robust the forest looks like. Similarly, in the random forest classifier, the higher the number of trees in the forest, greater is the accuracy of the results.

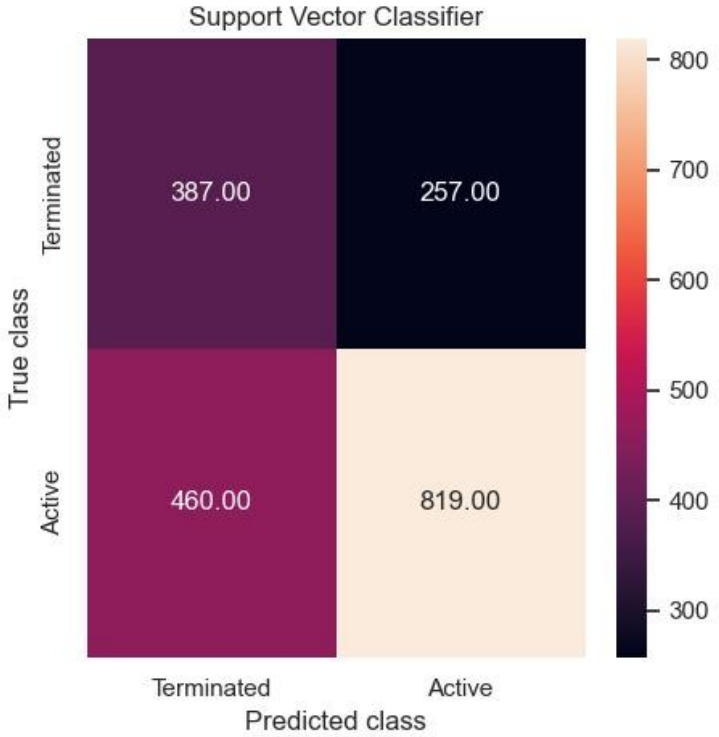
0	0.61	0.54	0.57	847
1	0.67	0.73	0.70	1076
accuracy			0.65	1923
macro avg	0.64	0.64	0.64	1923
weighted avg	0.65	0.65	0.64	1923



Support Vector Classifier :

- SVM is a supervised machine learning algorithm which is mainly used to classify data into different classes
- SVM can be used to generate multiple separating hyperplanes such that the data is divided into segments and each segment contains only one kind of data

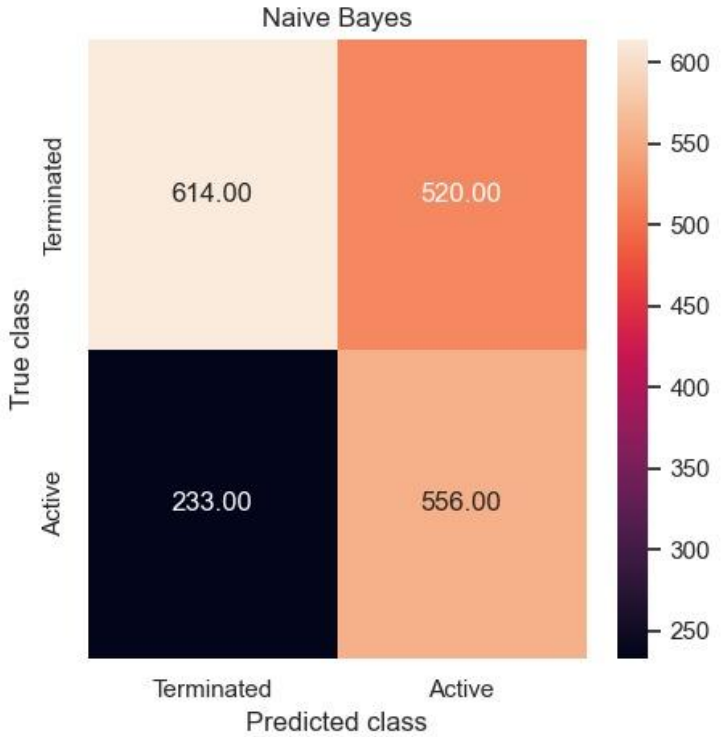
	precision	recall	f1-score	support
0	0.60	0.46	0.52	847
1	0.64	0.76	0.70	1076
accuracy			0.63	1923
macro avg	0.62	0.61	0.61	1923
weighted avg	0.62	0.63	0.62	1923



Naive Bayes :

It's a classification technique based on Bayes' theorem with an assumption of independence among predictors. In simple terms, it assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.

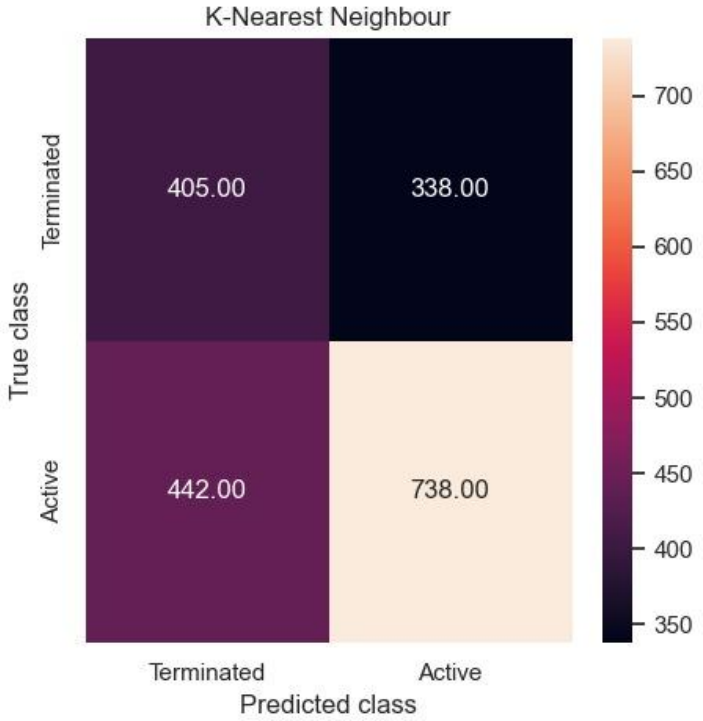
	precision	recall	f1-score	support
0	0.54	0.72	0.62	847
1	0.70	0.52	0.60	1076
accuracy			0.61	1923
macro avg	0.62	0.62	0.61	1923
weighted avg	0.63	0.61	0.61	1923



K-Nearest Neighbour :

The k-nearest neighbors (KNN) algorithm is a simple, supervised machine learning algorithm that can be used to solve both classification and regression problems.

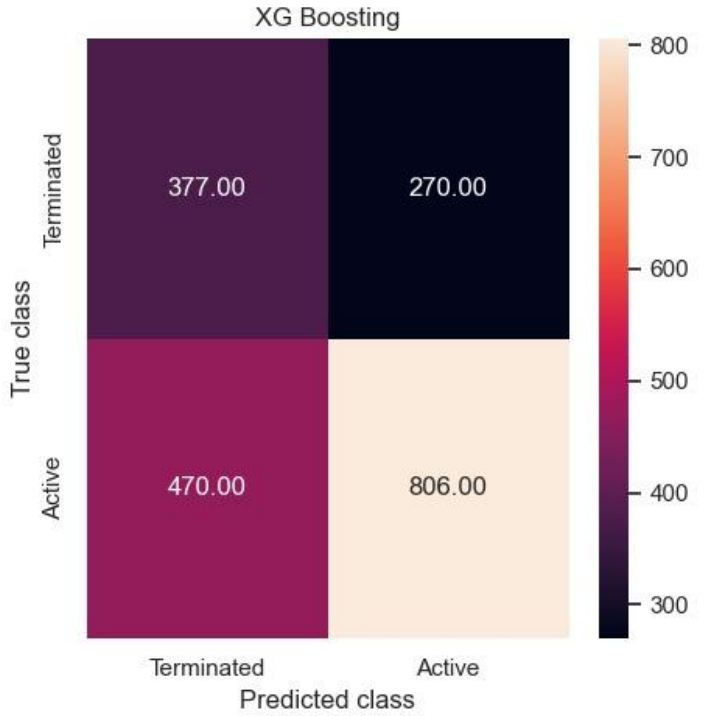
	precision	recall	f1-score	support
0	0.55	0.48	0.51	847
1	0.63	0.69	0.65	1076
accuracy			0.59	1923
macro avg	0.59	0.58	0.58	1923
weighted avg	0.59	0.59	0.59	1923



XG Boosting :

- XGBoost stands for eXtreme Gradient Boosting
- XGBoost is a decision-tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework

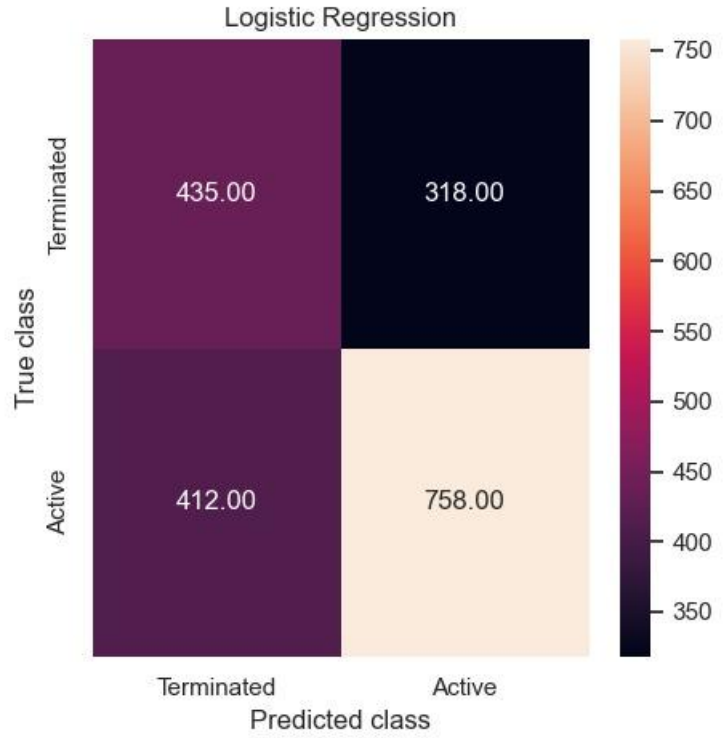
	precision	recall	f1-score	support
0	0.58	0.45	0.50	847
1	0.63	0.75	0.69	1076
accuracy			0.62	1923
macro avg	0.61	0.60	0.60	1923
weighted avg	0.61	0.62	0.61	1923



Logistic Regression :

- Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable
- Logistic Regression classification was performed using the training data

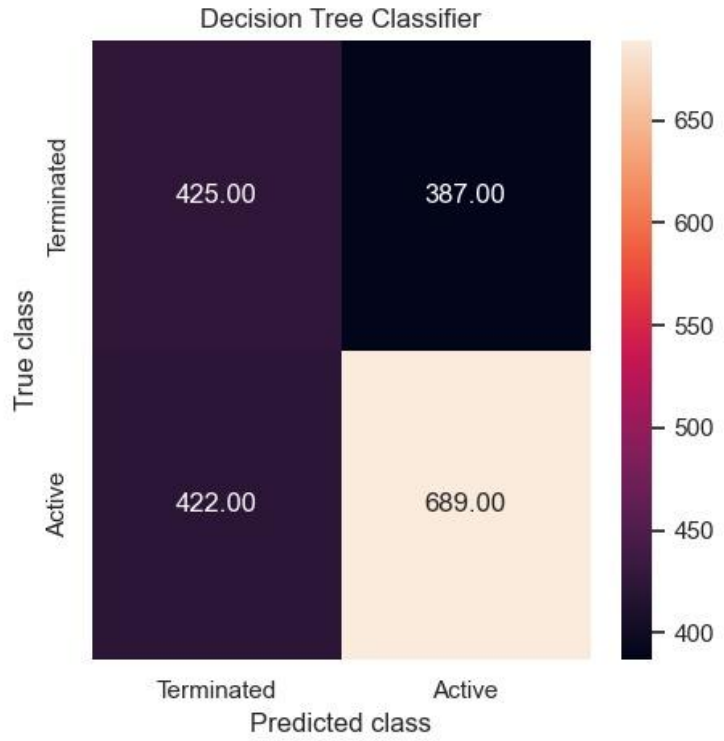
	precision	recall	f1-score	support
0	0.58	0.51	0.54	847
1	0.65	0.70	0.67	1076
accuracy			0.62	1923
macro avg	0.61	0.61	0.61	1923
weighted avg	0.62	0.62	0.62	1923



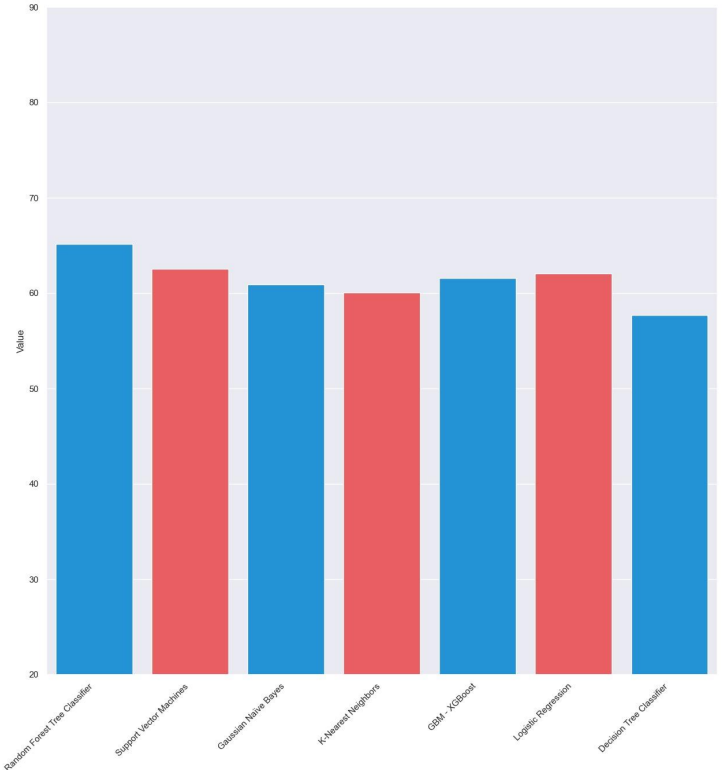
Decision Tree Classifier :

- Decision Trees (DTs) are a non-parametric supervised learning method used for classification and regression.
- It is a graphical representation of all possible solutions to a decision based on certain conditions. On each step or node of a decision tree, used for classification, we try to form a condition on the features to separate all the labels or classes contained in the dataset to the fullest purity

	precision	recall	f1-score	support
0	0.52	0.50	0.51	847
1	0.62	0.64	0.63	1076
accuracy			0.58	1923
macro avg	0.57	0.57	0.57	1923
weighted avg	0.58	0.58	0.58	1923



Comparison of Model :



Conclusion:

The Random Forest Tree Classifier model performed better than the other classification models.

	Model	Value
0	Random Forest Tree Classifier	64.58658346333853
1	Support Vector Machines	62.558502340093604
2	Gaussian Naïve Bayes	60.8944357774311
3	K-Nearest Neighbors	60.06240249609984
4	GBM - XGBoost	61.57046281851274
5	Logistic Regression	62.03848153926157
6	Decision Tree Classifier	57.46229849193968

Thank you!