FUNDAMENTALS OF MACHINE LEARNING FINAL EXAM

STUDENT NAME: LAYA SREE GANGULA

STUDENT ID: 811231669

ATTRITION OF EMPLOYEES IN AN ORGANIZATION

TABLE OF CONTENTS

LIST OF FIGURES

INTRODUCTION

The percentage of the amount of people leaving the organization and being replaced by the new employees is known as attrition. In any organization, retainment of hardworking, and loyal employees is important, as it helps the company survive in the market place. This is because if the percentage of attrition is high in a company, the cost of hiring and training of recruitment increases, which puts a strain on the companies' finances. In addition, competent and hardworking loyal employees are hard to find and recruit. Therefore, it is very important for an organization to know which employees would most probably leave the company or not. This will help the company in many ways, i.e., reduce recruitment and training cost, reduce pressure on HR department, retain good employees, and of course foresee the attrition of employee to better be prepared beforehand, so that mismanagement doesn't happen again. Therefore, I would like to study attrition of employees in an organization to analyze important features that result in increased attrition, and also to develop a Machine Learning model that predicts attrition – whether an employee will leave the company or not. This will make things simpler and the model would help companies or organizations to know which employee would most likely to leave to help them plan better strategies for a better future of their organization and help retainment of employees.

PROBLEM STATEMENT

Develop a model using supervised machine learning techniques to predict the attrition of employees by measuring certain factors involved in the organization.

PROJECT OBJECTIVES

Following are the research objectives of the study:

1. Perform Exploratory Data Analysis (EDA) on dataset for initial investigation on the dataset to find missing values, outliers, relationships, correlations between various features of the dataset.

2. Perform Data Pre-Processing & Cleaning to clean and prepare the dataset for the implementation of Machine learning models.
3. Build a Machine Learning model that predicts the attrition of employee in organizations – whether an employee will leave the organization or not, with an accuracy higher than 85%.

RESEARCH METHODOLOGY/APPROACH

In Machine learning, there are four types of categories, i.e., Supervised Learning, Unsupervised Learning, Semi-supervised Learning, and Reinforced learning, and in this dataset, I am studying a labelled dataset, therefore, I would implement supervised machine learning techniques to complete the project.

I will use following machine learning methodology to complete the project:

I. Exploratory Data Analysis

Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations [1].

II. Data Preprocessing and Cleaning

In this phase, the pre-processing and cleaning will be performed to get the dataset ready for the implementation of Machine Learning models. The data is organized in a standard format, null or missing values are removed, white or empty spaces are removed, categorical features conversion to numerical, etc., in the data cleaning and pre-processing phase [2].

III. Implementing Models

Machine learning allows computers to learn from the data directly just like humans and animal do – learn from experience. In this study, as I am dealing with a labelled dataset, therefore, I will use supervised machine learning techniques to have the models trained for the purpose of predicting.

- Gaussian Naïve Bayes
- K-nearest Neighbor
- Support Vector Machine

IV. Results & Findings

In this phase, I will discuss the results and findings of my analysis on the employee's attrition dataset in detail.

V. Conclusion

In this phase, I will conclude the project provide my remarks for future work.

DATASET

A real-world dataset with title “IBM HR Analytics Employee Attrition & Performance” which has been downloaded from Kaggle.com containing information regarding attrition of employees in an organization (IBM). It has 35 features including Age, BusinessTravel, DailyRate, Department, Education, JobLevel, JobRole, MonthlyIncome, OverTime, WorkLifeBalance, YearsAtCompnay, and others including the target feature Attrition that determines whether an employee will leave the organization or not. This dataset contains almost 1500 rows which is good because it is enough to train the models well and it will increase the performance of the models. In fig. 1, a snapshot of employee attrition dataset from Kaggle has been displayed.

It is a real-world dataset because it contains all the important variables measuring education, environment satisfaction, job involvement, job satisfaction, performance rating, relationship satisfaction, and work life balance, which are the determinant of employee attrition in the organization. It has a record of almost 1500 employees which is suitable enough to develop a model that will be able to determine employee attrition in organization. After a model has been developed, this can be employed in organizations to determine the employee attrition so that the organizations can run efficient and effectively by retaining employees and sorting issues with employees who would most likely to leave. Therefore, it is important to conduct an analysis on employee attrition.

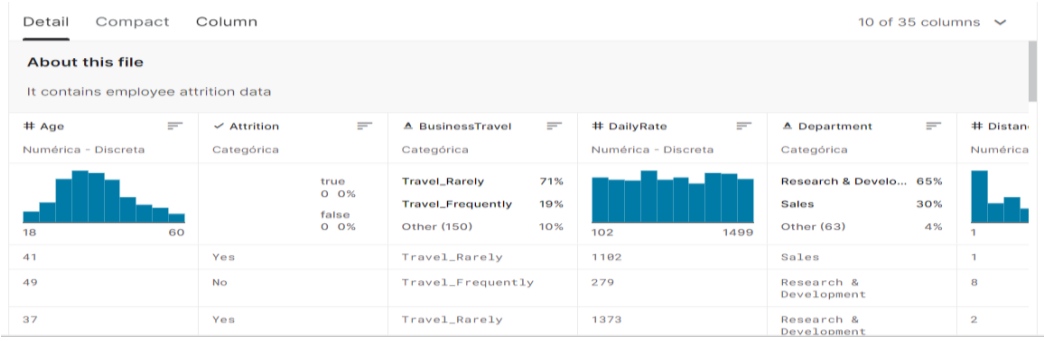


Figure 1 Employee Attrition Dataset - Kaggle

IMPLEMENTATION

EXPLORATORY DATA ANALYSIS

In fig. 2, the dataset being read into the Jupyter's notebook using read_csv() function, and also displaying the first five rows of the dataset using head() function.

```
df=pd.read_csv('WA_Fn-UseC_-HR-Employee-Attrition.csv')
df.head()
```

	Age	Attrition	BusinessTravel	DailyRate	Department	DistanceFromHome	Education	EducationField	EmployeeCount	EmployeeNumber	...	RelationshipSa
0	41	Yes	Travel_Rarely	1102	Sales	1	2	Life Sciences	1	1	...	
1	49	No	Travel_Frequently	279	Research & Development	8	1	Life Sciences	1	2	...	
2	37	Yes	Travel_Rarely	1373	Research & Development	2	2	Other	1	4	...	
3	33	No	Travel_Frequently	1392	Research & Development	3	4	Life Sciences	1	5	...	
4	27	No	Travel_Rarely	591	Research & Development	2	1	Medical	1	7	...	

5 rows × 35 columns

Figure 2 Reading the dataset

In fig. 3, it can be seen that the total number of classes (Attrition) available in the dataset. 'Yes' data point is close to 230, while 'No' data point is close to 1200. This means that around 230 people left the organization, and around 1200 didn't.

```
ax = sns.countplot(x="Attrition", data=df)
```

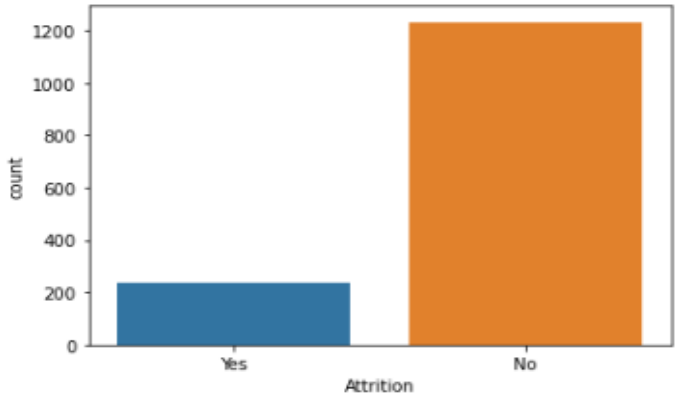


Figure 3 Points available in the dataset

In fig. 4, it can be seen that the years in current role column against ‘Attrition’ attributes. It can be said that the people who spend more than 5 years in an organization, didn’t leave it.

```
sns.scatterplot(df['YearsInCurrentRole'],df['YearsInCurrentRole'],hue=df['Attrition'])
plt.show()
```

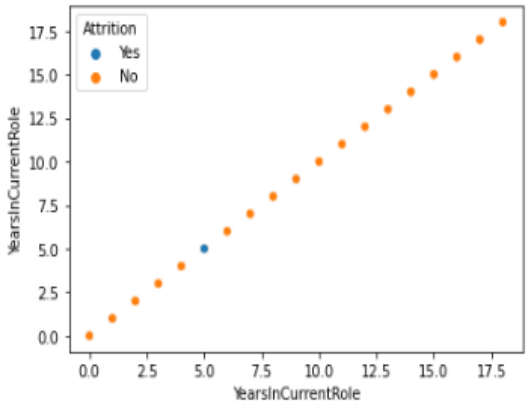


Figure 4 Years in Current Role

In fig. 5, It can be said that people either leave organization before 5 years of service or after 20 years of services.

```
sns.scatterplot(df['YearsAtCompany'],df['YearsAtCompany'],hue=df['Attrition'])
plt.show()
```

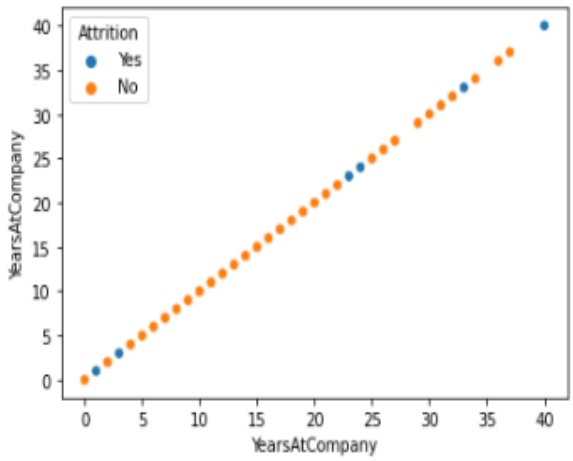


Figure 5 Years at Company

In fig. 6, It can be said that the attrition of employees with respect to department, and Research & Development Department is at the top having maximum number of employees leaving the organization, followed by Sales, and Human Resource department.

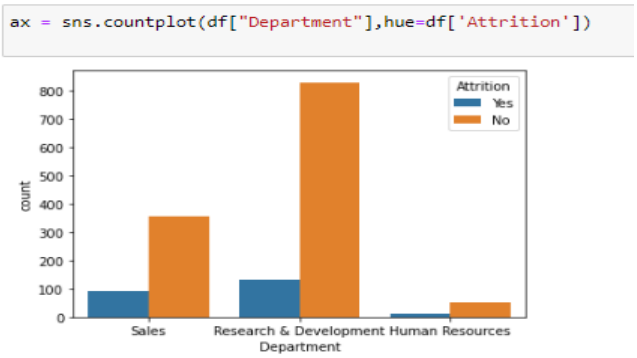


Figure 6 Department wise attrition of employees

In fig. 7, It can be said that the attrition of employees according to gender, and I found that a greater number of male employees leaving as compared to female employees. In terms of their numbers 150 male employees left out of 882, and 87 females left out of 588 from the organization.

Figure 7 Gender: Attrition

In fig. 8, it can be said that the attrition of employees according to their job role. Research directors don't leave the organization too often as there are only 2 number of employees who left the organization. Laboratory technicians leave the organization mostly, followed by Sales Executive, Research Scientists, and Sales Representatives.

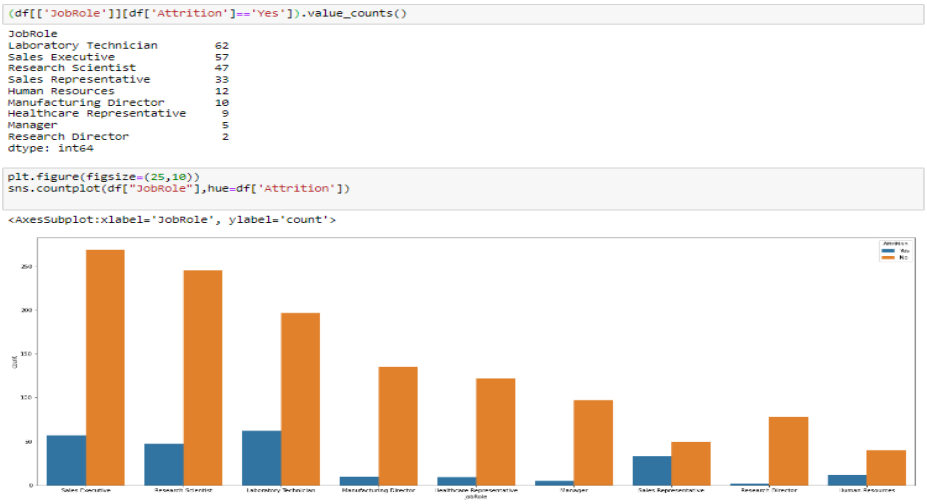


Figure 8 Job role: attrition

DATA PRE-PROCESSING & CLEANING

I performed various data cleaning and pre-processing steps; however, some major steps have been displayed in the following points:

1. In dataset, features having object data type were converted into numerical ones using Label Encoder. This is because computers only understand numerical values, and so is the case in Machine Learning. In fig. 9, encoding of categorical features can be seen.

```
from sklearn.preprocessing import LabelEncoder
labelencoder_X = LabelEncoder()
obj=df[['Attrition', 'BusinessTravel', 'Department', 'EducationField', 'Gender',
       'JobRole', 'MaritalStatus', 'Over18', 'OverTime']]
for i in obj.columns:
    df[i]=labelencoder_X.fit_transform(df[i])
```

Figure 9 object type features conversion

2. I used *imblearn* library to balance the dataset using oversampling for 'Yes' class, and under sampling for 'No' in the attrition feature, which is the classification data frame. This is because there were a greater number of 'No' classes, and a smaller number of 'Yes' classes, and this is problematic in model's prediction accuracy. In fig. 10, over sampling and under sampling using imblearn pipeline can be seen.

```
# define resampling
from imblearn.pipeline import Pipeline
from imblearn.over_sampling import RandomOverSampler
from imblearn.under_sampling import RandomUnderSampler
over = RandomOverSampler(sampling_strategy=0.5)
under = RandomUnderSampler(sampling_strategy=0.7)
# define pipeline
pipeline = Pipeline(steps=[('o', over), ('u', under)])
# transform the dataset
X, y = pipeline.fit_resample(X, y)
```

Figure 10 Over and Under sampling

MACHINE LEARNING MODELS

Gaussian Naïve Bayes

Gaussian Naïve Bayes is a probabilistic classification algorithm in Machine learning that uses Bayes theorem to draw conclusions. This theorem helps it to use conditional probability to predict the unknown classes. It can handle large amount of data with ease, and performs fast classification in prediction [3].

In fig. 11, it can be seen that I implemented Gaussian Naïve Bayes using Sklearn library.

```
from sklearn.naive_bayes import GaussianNB

from sklearn.metrics import confusion_matrix
dt =GaussianNB()
dt.fit(X_train, y_train)
y_predict_dt = dt.predict(X_test)
# confusion_matrix
plt.figure(figsize=(10,10))
cm = confusion_matrix(y_test, y_predict_dt)
sns.heatmap(cm, annot=True, fmt="d")
```

Figure 11 Code: Gaussian Naive Bayes

In fig. 12, the confusion matrix graph shows the prediction results of Gaussian naïve bayes. It predicted attrition of employees with 'No' data point 120 time right, and 45 times wrong. Similarly, it predicted 'Yes' 101 time right, and 34 times wrong. Moreover, the accuracy of Gaussian Naïve Bayes is 73%, while precision, and recall is 69%, and 74% respectively.

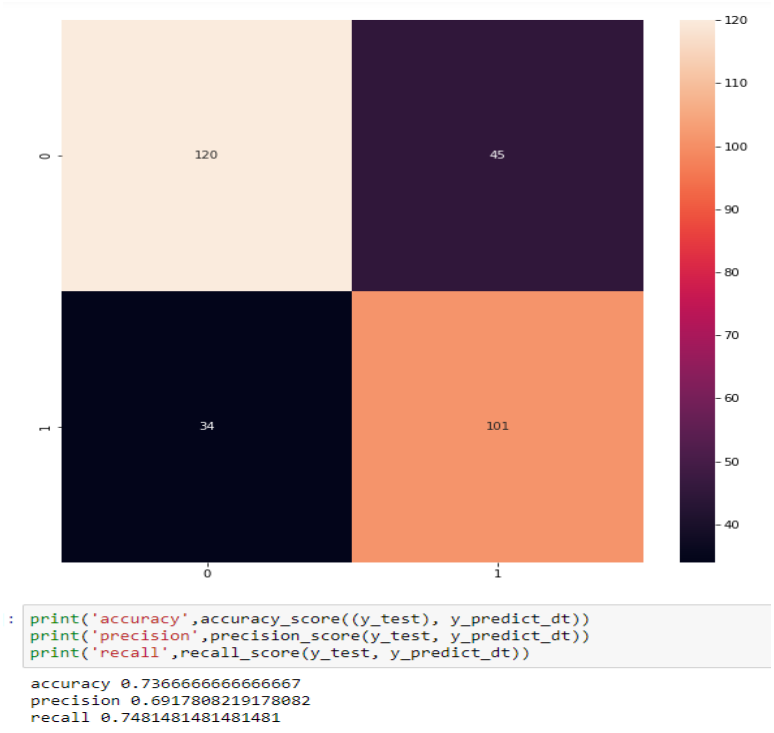


Figure 12 Confusion matrix, Accuracy, Precision, and Recall

K-Nearest Neighbor

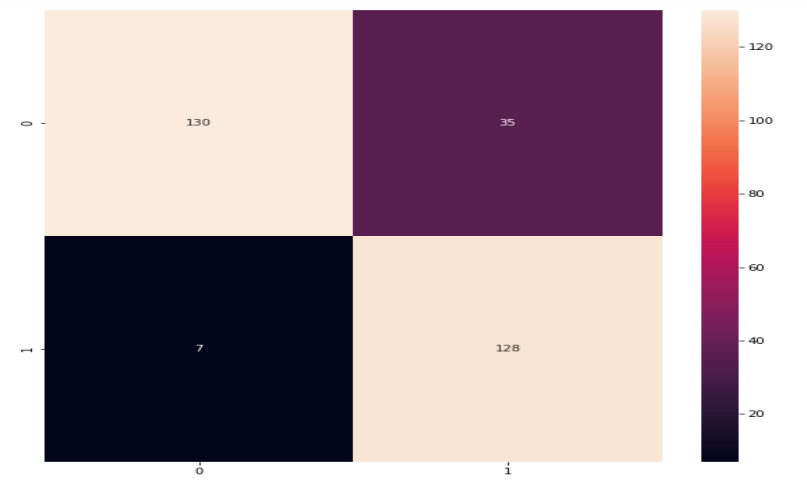
It can solve classification, and regression problems very effectively. It is also a non-parametric algorithm, because few hyper-parameters are required for the functioning of the algorithm – k-value, and distance. It adapts easily to new data, however, is very lazy because it doesn't train on the data itself, rather upon classification, it keeps new cases with the most similar available ones [6].

In fig. 13, it can be seen that I implemented K-nearest Neighbor using Sklearn library.

```
from sklearn.neighbors import KNeighborsClassifier
import warnings
from sklearn.metrics import confusion_matrix
knn = KNeighborsClassifier(algorithm='brute', metric='manhattan', n_neighbors=1)
knn.fit(X_train, y_train)
y_predict_knn = knn.predict(X_test)
# confusion_matrix
plt.figure(figsize=(10,10))
cm = confusion_matrix(y_test, y_predict_knn)
sns.heatmap(cm, annot=True, fmt="d")
```

Figure 13 Code: KNN

In fig. 14, the confusion matrix graph shows the prediction results of K-nearest neighbour. It predicted attrition of employees with 'No' data point 130 time right, and 35 times wrong. Similarly, it predicted 'Yes' 120 time right, and 7 times wrong. Moreover, the accuracy of K-nearest neighbour is 86%, while precision, and recall is 78%, and 94% respectively.




```
print('accuracy',accuracy_score((y_test), y_predict_knn))
print('precision',precision_score(y_test, y_predict_knn))
print('recall',recall_score(y_test, y_predict_knn))
```

```
accuracy 0.86
precision 0.7852760736196319
recall 0.9481481481481482
```

Figure 14 Confusion matrix, Accuracy, Precision, and Recall

Support Vector Machines

In N-dimensional space, where N is the total number of features, it finds the hyperplane to classify the datapoints in a distinguish way. Many numbers of hyperplanes can be chosen to classify the different

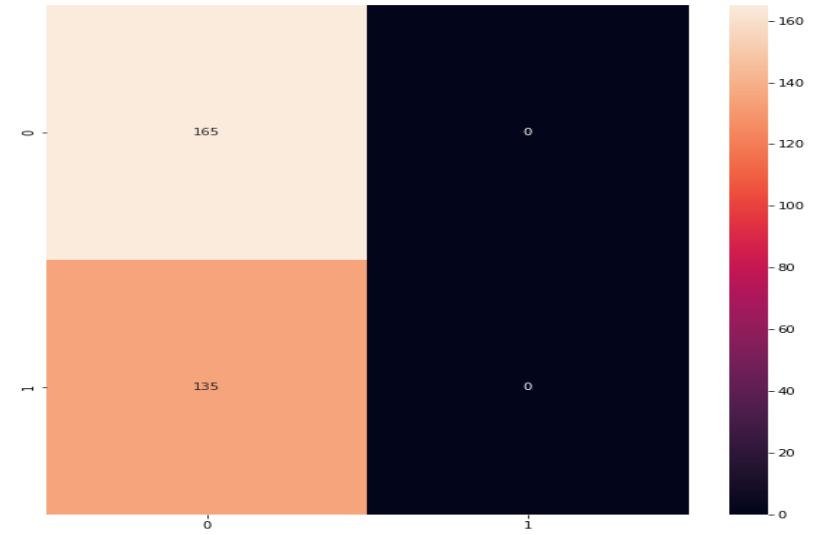
hyperplanes. It classifies extremely well with a small dataset, and is dependable for classification and regression problems [5].

In fig. 15, In fig. 16, it can be seen that I implemented Support Vector Machine using Sklearn library.

```
from sklearn.svm import SVC
rf =SVC()
rf.fit(X_train, y_train)
y_predict_rf = rf.predict(X_test)
# confusion_matrix
plt.figure(figsize=(10,10))
cm = confusion_matrix(y_test, y_predict_rf)
sns.heatmap(cm, annot=True, fmt="d")
```

Figure 15 Code: Support Vector Machine

In fig. 16, the confusion matrix graph shows the prediction results of Support Vector Machine. It predicted attrition of employees with 'No' data point 165 time right, and 0 time wrong. Similarly, it predicted 'Yes' 0 time right, and 135 times wrong. Moreover, the accuracy of Support Vector Machine is 55%, while precision, and recall is 0%, and 0% respectively.




```
print('accuracy',accuracy_score((y_test), y_predict_rf))
print('precision',precision_score(y_test, y_predict_rf))
print('recall',recall_score(y_test, y_predict_rf))
```

```
accuracy 0.55
precision 0.0
recall 0.0
```

Figure 16 Confusion matrix, Accuracy, Precision, and Recall

RESULTS & FINDINGS

Following points highlight the findings of Exploratory Data Analysis (EDA):

1. People who spend more than 5 years in an organization, didn't leave it.
2. People either leave organization before 5 years of service or after 20 years of services.
3. The attrition of employees with respect to department, and Research & Development Department is at the top having maximum number of employees leaving the organization, followed by Sales, and Human Resource department.
4. I found that a greater number of male employees leaving as compared to female employees. In terms of their numbers 150 male employees left out of 882, and 87 females left out of 588 from the organization.
5. Research directors don't leave the organization too often as there are only 2 number of employees who left the organization. Laboratory technicians leave the organization mostly, followed by Sales Executive, Research Scientists, and Sales Representatives.

During modelling phase, I implemented Gaussian naïve bayes, K-nearest Neighbor, and Support Vector Machine. I found K-nearest neighbor with an excellent prediction accuracy of 86%, while the precision and recall were 78%, and 94% respectively. In addition, K-nearest neighbor predicted attrition of employees with 'No' data point 130-time righty, and 35 times wrong. Similarly, it predicted 'Yes' 120 time rightly, and 7 times wrong. In contrast, Gaussian naïve bayes and Support vector machine followed K-nearest neighbor with an accuracy of 73%, and 55% accordingly.

In table 1, accuracy, precision, and recall of three implemented models including, Gaussian Naïve Bayes, K-nearest Neighbors, and Support Vector Machine can be seen.

Table 1 Accuracy, Precision, and Recall

Serial no.	Machine Learning Models	Accuracy	Precision	Recall
1.	Gaussian Naïve Bayes	73%	69%	74%
2.	***K-nearest Neighbor***	86%	78%	94%
3.	Support Vector Machine	55%	0%	0%

CONCLUSIONS

In this study, I analyzed the attrition of employees in the organization using a real-world dataset from Kaggle. After performing Exploratory Data Analysis (EDA) I was able to draw some important conclusions: first, there are a greater number of chances for an employee to leave the company within first five years. Secondly, Males have high tendency to leave than females. Thirdly, Research & Development Department is at the top for having maximum number of employees leaving the organization, followed by Sales, and Human Resource department. Finally in job roles, Research directors don't leave the organization too often as there are only 2 number of employees who left the organization. Laboratory technicians leave the organization mostly, followed by Sales Executive, Research Scientists, and Sales Representatives. Then, I performed the cleaning and pre-processing of the dataset, where I balanced the dataset using over sampling and under sampling technique of imblearn library. Finally, I was able to implement Machine Learning models including Gaussian Naïve Bayes, K-nearest neighbour, and Support Vector Machine, and found K-nearest neighbours outperforming all others with an accuracy of 86%. In future, a greater number of models can be tested such as Random Forest, Decision Tree, etc., and Forward Feature Selection technique can be applied to enhance the predictive accuracy of the models.

REFERENCES

[1] Patil, P., 2018, What is exploratory data analysis? Towards data science. Available at: https://towardsdatascience.com/exploratory-data-analysis-8fc1cb20fd15

[2] Goyal, K., (2021), "Data Pre-processing in Machine Learning: 7 Easy Steps to Follow", Upgrad, https://www.upgrad.com/blog/data-preprocessing-in-machine-learning/

[3] Sharma, P., (2021), "Implementation of Gaussian Naive Bayes in Python Sklearn", Analytics Vidhya, https://www.analyticsvidhya.com/blog/2021/11/implementation-of-gaussian-naive-bayes-in-python-sklearn/

[4] Gandhi, R., (2018), "Support Vector Machine — Introduction to Machine Learning Algorithms", Towards Data Science, https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47

[5] Stecanella, B., (2017), "Support Vector Machines (SVM) Algorithm Explained", Monkey Learn, https://monkeylearn.com/blog/introduction-to-support-vector-machines-svm/

[6] Anonymous, (n.d.), "What is the k-nearest neighbors' algorithm?", IBM, https://www.ibm.com/topics/knn?mhsrc=ibmsearch_a&mhq=what%20is%20k-nearest%20neighbo

[7] Riggio, C., "What's the deal with Accuracy, Precision, Recall and F1?", Towardsdatascience, https://towardsdatascience.com/whats-the-deal-with-accuracy-precision-recall-and-f1-f5d8b4db1021

[8] Zubair, M.D, "Find the Patterns of a Dataset by Visualizing Frequency Distribution", Towardsdatascience, https://towardsdatascience.com/find-the-patterns-of-a-dataset-by-visualizing-frequency-distribution-c5718ab1f2c2 , 2021

[9] Gad, A.F.," Evaluating Deep Learning Models: The Confusion Matrix, Accuracy, Precision, and Recall", blog.paperspace.com, https://blog.paperspace.com/deep-learning-metrics-precision-recall-accuracy/

[10] Kumar, A., "Correlation Concepts, Matrix & Heatmap using Seaborn", Vitalflux, https://vitalflux.com/correlationheatmap-with-seaborn-pandas/

[11] Pavansubhash, 2016. IBM HR Analytics Employee Attrition & Performance. Kaggle. Available at: https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset