

Corporate Employee Attrition Analysis

A PROJECT REPORT

Submitted by

Team ID: PNT2022TMID53479

Team Leader: KESAV S J (2127190801038)

Team member: KOUSHIK K (2127190801040)

Team member: DEEKSHITHA M (2127190801013)

Team member: GIRIDHAR PRASHANTTH S (2127190801021)

DEPARTMENT OF INFORMATION TECHNOLOGY

SRI VENKATESWARA COLLEGE OF ENGINEERING

CHENNAI 602117

TABLE OF CONTENTS

CHAPTER NO.	TITLE	PAGE NO.
1	ACKNOWLEDGEMENTS	3
2	OBJECTIVE	4
3	DESCRIPTION OF PROJECT	5
4	METHODOLOGY	5
5	ASSUMPTIONS	8
6	VISUALIZATIONS	9
7	REFLECTION ON THE PROJECT	13
8	CONCLUSION	14
9	LINK TO CODE AND EXECUTABLE FILE	15

ACKNOWLEDGEMENTS

I had a fantastic opportunity for learning and career development with the Nalaiya Thiran opportunity I had with IBM in the Data Analytics field. I, therefore, consider myself extremely fortunate to be a part of it. I also want to express my gratitude to all the experts who guided me during my project.

I would like to take this opportunity to express my sincere gratitude to my principal, who helped me despite being incredibly busy with her duties and enabled me to complete my internship at the prestigious company.

I would like to extend my sincere gratitude to my industry mentor, Mr. Shanawaz Anwar, for his participation in helpful decisions, advice, and guidance, as well as for setting up all of the necessary facilities to facilitate the internship. I've decided to express my gratitude for his contribution right now.

I would like to express my sincere appreciation to Dr. T. Sukumar, an associate professor and the assistant department head for information technology at our college, for his thoughtful and invaluable guidance, which was extremely helpful for my study both theoretically and practically.

I see this opportunity as a significant turning point in my professional development. In order to achieve my desired career goals, I will make every effort to utilise newly acquired skills and knowledge to the fullest extent possible and to keep working to improve them.

Sincerely,

Team Leader: KESAV S J

Team member: KOUSHIK K

Team member: DEEKSHITHA M

Team member: GIRIDHAR PRASHANTH S

OBJECTIVE

- This project's goal is to predict each employee's attrition rate in order to identify those who are most likely to leave the company.
- Finding ways to stop attrition or scheduling the hiring of a new candidate in advance will help the organisation.
- The organisation finds attrition to be a time- and money-consuming problem that also results in the loss of probability.
- The project's scope includes businesses in all sectors.

DESCRIPTION OF PROJECT

- The dataset in this project needs to be cleaned. Then, train the dataset to forecast the organization's employee attrition rate.

METHODOLOGY

1. Business Understanding:

Before attempting to solve a problem in the business domain, it must first be properly understood. A solid foundation created by business knowledge makes it easier to answer questions. We need to be clear about the precise issue we intend to address.

2. Analytic Understanding:

One should choose the analytical strategy to use based on the business understanding discussed above. There are four different types of approaches: descriptive (current status and information provided), diagnostic (A.K. A statistical analysis, what is occurring and why), a predictive strategy (which projects trends or the likelihood of future events), and a prescriptive strategy (how the problem should be solved actually).

3. Data Requirements:

The above-mentioned analytical approach identifies the pertinent data's required content, formats, and sources. Finding the

answers to the following questions during the data requirements process is necessary: "What," "Where," "When," "Why," "How," and "Who".

4.Data Collection:

Any random format can be used to obtain the collected data. Therefore, the data gathered should be validated in accordance with the methodology picked and the results expected. Therefore, if more information is needed, it can be collected or it can be discarded.

5.Data Understanding:

The question "Is the data collected representative of the problem to be solved?" is answered by data understanding. The measures that are applied to the data in descriptive statistics are calculated in order to assess the matter's quality and content. This step could result in going back to the previous step to make adjustments.

6.Data Preparation:

Let's connect this idea with two analogies to better understand it. Washing just-picked vegetables and only taking what you want from the buffet to put on your plate are the first two things to remember. Vegetable washing represents the removal of impurities, or unwanted materials, from the data. Noise cancellation is done here.

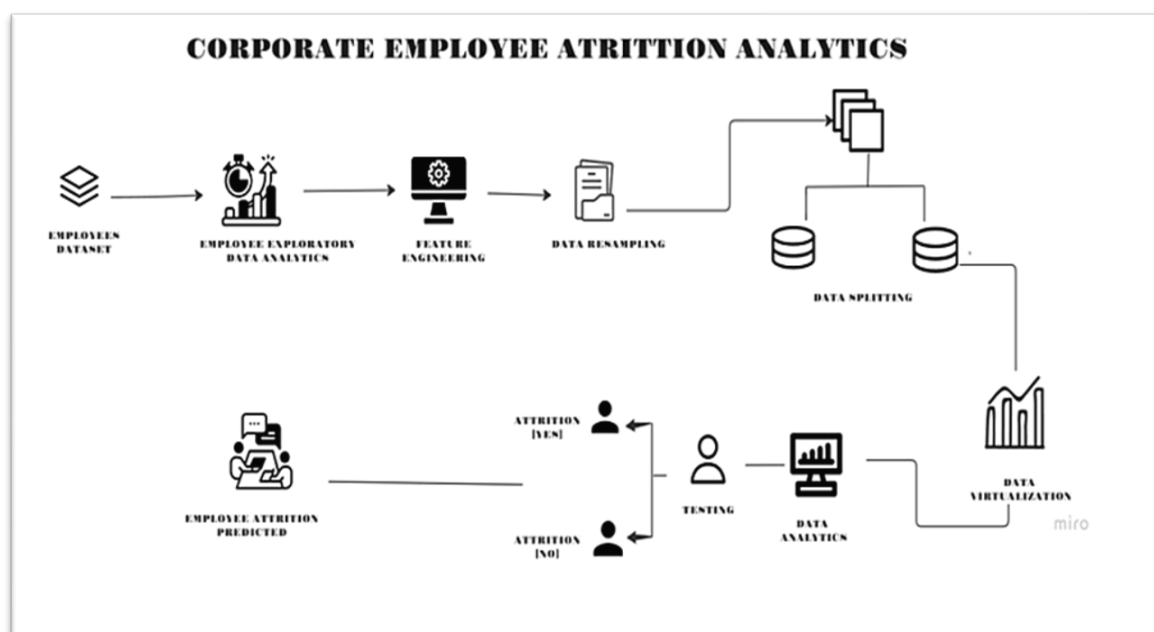
If we are only considering edible items on the plate and we don't require specific information, we shouldn't proceed with the process. Included in this entire process are transformation, normalisation, etc.

7. Modelling:

Modeling determines whether the data that has been prepared for processing is suitable or needs more seasoning and finishing. The development of predictive and descriptive models is the main goal of this stage.

8. Evaluation:

Model development includes model evaluation. It examines the model's quality and determines whether it satisfies the business requirements. It goes through a diagnostic measure phase (which determines whether the model functions as intended and where adjustments are needed) and a statistical significance testing phase.



9. Deployment:

The model is prepared for deployment in the business market as it is successfully evaluated. The deployment phase determines how well the model performs in comparison to competitors and how much

external stress it can withstand.

10. Feedback:

In order to improve the model and assess its performance and impact, feedback is a crucial goal. The steps in providing feedback include defining the review process, maintaining a record, assessing effectiveness, and reviewing with improvement. tiveness and review with refining.

ASSUMPTIONS

1. Each possible sample has assigned to it a known probability of selection.
2. We select one of the samples by a random process in which each sample receives its appropriate probability of being selected.
3. The method for computing the estimate must lead to a unique estimate for any specific sample.

4. Homoscedasticity: The variance of residual is the same for any value of X.
5. Independence: Observations are independent of each other.
6. Normality: For any fixed value of X, Y is normally distributed.
7. Multicollinearity: There should be no or little multicollinearity.
8. No auto-correlation

VISUALIZATIONS

This figure shows the data frame (which isn't cleaned and sanitized as it has lots of null values)

The screenshot shows a Jupyter Notebook interface with the following code and output:

```

In [5]: # machine learning
from sklearn import model_selection, tree, preprocessing, metrics, linear_model
from sklearn.metrics import confusion_matrix, classification_report
from sklearn.svm import SVC, LinearSVC
from sklearn.ensemble import RandomForestClassifier, GradientBoostingClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn.naive_bayes import GaussianNB
from sklearn.linear_model import Perceptron, SGDClassifier, LogisticRegression
from sklearn.tree import DecisionTreeClassifier
from sklearn.model_selection import train_test_split, StratifiedFold, GridSearchCV, learning_curve, cross_val_score
from catboost import CatBoostClassifier, Pool, cv

In [6]: # ignore warnings
import warnings
warnings.filterwarnings('ignore')

In [7]: import os
os.chdir("C:/Users/DELL/OneDrive/Desktop/Dataset")

In [8]: df = pd.read_csv('Employee-Attrition.csv')

In [9]: df

Out[9]:

```

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

The screenshot shows the continuation of the Jupyter Notebook with data cleaning and transformation steps:

```

In [10]: df.head()

Out[10]:

```

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

```

5 rows x 35 columns

In [11]: df.shape

Out[11]: (1470, 35)

Exploratory Data Analysis

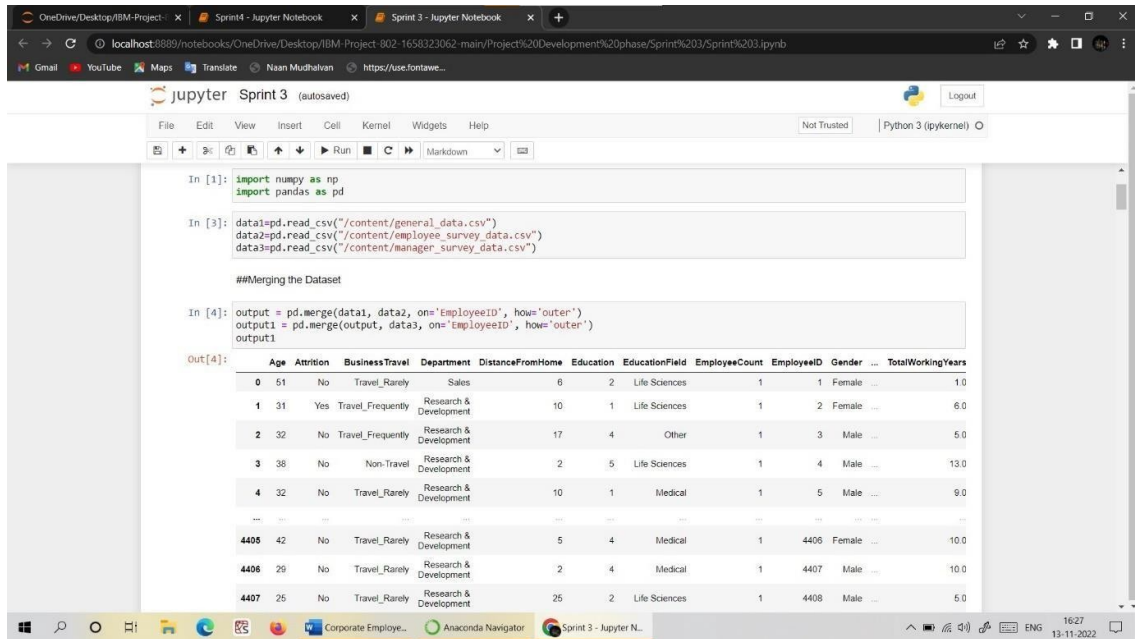
In [12]: # drop the unnecessary columns
df.drop(['EmployeeNumber', 'Over18', 'StandardHours', 'EmployeeCount'], axis=1, inplace=True)

In [13]: df['Attrition'] = df['Attrition'].apply(lambda x: 1 if x == "Yes" else 0)
df['OverTime'] = df['OverTime'].apply(lambda x: 1 if x == "Yes" else 0)

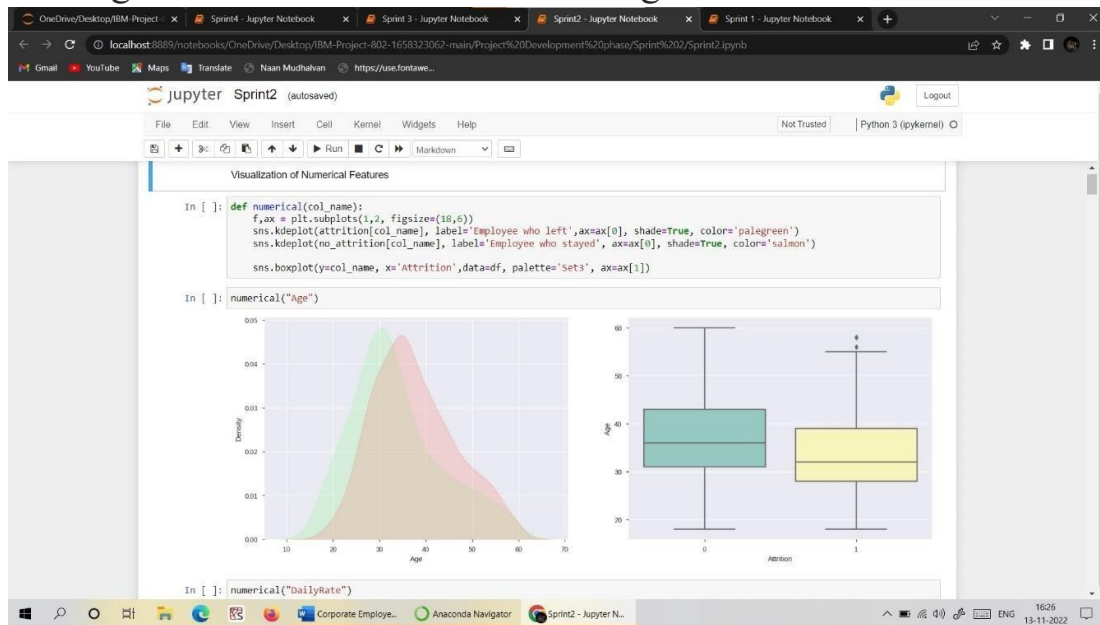
In [14]: attrition = df[df['Attrition'] == 1]
no_attrition = df[df['Attrition'] == 0]

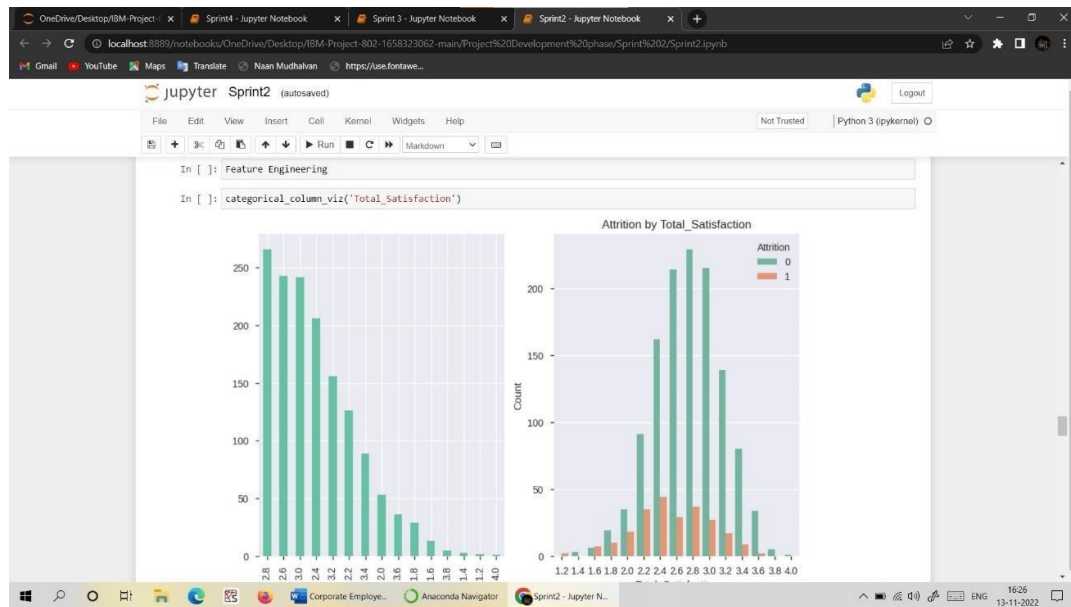
```

The below figure shows the dataset after removing unnecessary columns and the rows containing missing values and reordering the same.

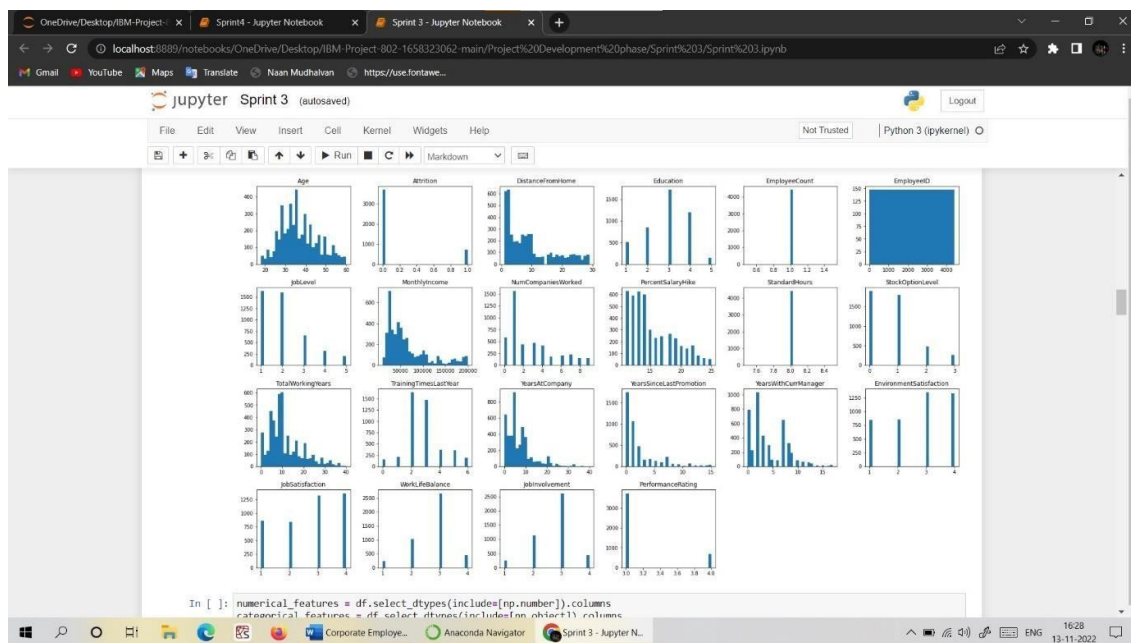


This figure shows the distribution of target variables

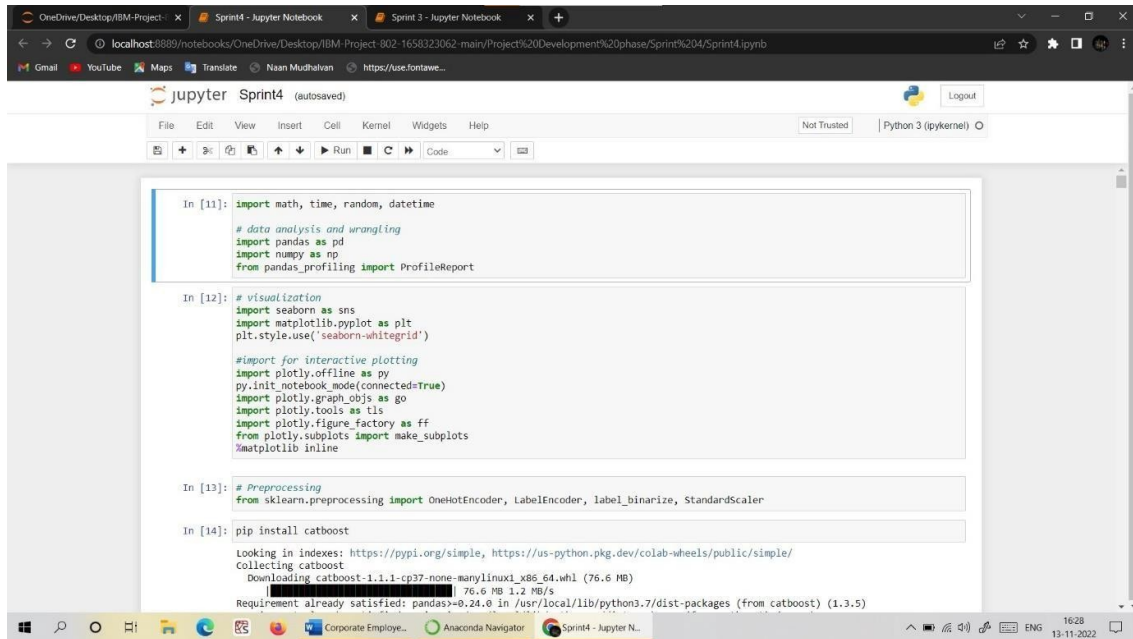




The visualisations of the numerical variables is



Similarly of the non numerical variables



```

In [11]: import math, time, random, datetime

# data analysis and wrangling
import pandas as pd
import numpy as np
from pandas_profiling import ProfileReport

In [12]: # visualization
import seaborn as sns
import matplotlib.pyplot as plt
plt.style.use('seaborn-whitegrid')

# import for interactive plotting
import plotly.offline as py
py.init_notebook_mode(connected=True)
import plotly.graph_objs as go
import plotly.tools as tls
import plotly.figure_factory as ff
from plotly.subplots import make_subplots
%matplotlib inline

In [13]: # Preprocessing
from sklearn.preprocessing import OneHotEncoder, LabelEncoder, Label_Binarize, StandardScaler

In [14]: pip install catboost

```

Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-wheels/public/simple/
Collecting catboost
 Downloading catboost-1.1.1-cp37-none-manylinux1_x86_64.whl (76.6 MB)
 76.6 MB 1.2 MB/s
Requirement already satisfied: pandas>=0.24.0 in /usr/local/lib/python3.7/dist-packages (from catboost) (1.3.5)

And then, we train the model and test for unknown data to predict the attrition rate. And the screenshot is attached herewith.

```

# Logistic Regression
start_time = time.time()
train_pred_log, acc_log, acc_cv_log = fit_ml_algo(LogisticRegression(), X_train, y_train, 10)
log_time = (time.time() - start_time)
print("Accuracy: %s" % acc_log)
print("Accuracy CV 10-Fold: %s" % acc_cv_log)
print("Running Time: %s" % datetime.timedelta(seconds=log_time))

```

```

Accuracy: 89.8
Accuracy CV 10-Fold: 88.63
Running Time: 0:00:01.753627

```

```
# SVC
start_time = time.time()
train_pred_svc, acc_svc, acc_cv_svc = fit_ml_algo(SVC(),X_train,y_train,10)
svc_time = (time.time() - start_time)
print("Accuracy: %s" % acc_svc)
print("Accuracy CV 10-Fold: %s" % acc_cv_svc)
print("Running Time: %s" % datetime.timedelta(seconds=svc_time))
```

Accuracy: 88.53
Accuracy CV 10-Fold: 85.91
Running Time: 0:00:00.497278

```
# Linear SVC
start_time = time.time()
train_pred_svc, acc_linear_svc, acc_cv_linear_svc = fit_ml_algo(LinearSVC(),X_train, y_train,10)
linear_svc_time = (time.time() - start_time)
print("Accuracy: %s" % acc_linear_svc)
print("Accuracy CV 10-Fold: %s" % acc_cv_linear_svc)
print("Running Time: %s" % datetime.timedelta(seconds=linear_svc_time))
```

Accuracy: 89.89
Accuracy CV 10-Fold: 88.73
Running Time: 0:00:01.055932

REFLECTIONS ON THE PROJECT

Our work on the Nalaiya Thiren project with IBM and the Tamil Nadu government has been the most rewarding and instructive. I was able to finish my project thanks to the supportive, sympathetic, and empathetic mentors I had during this experience. because of the techniques, we learned from books and the internet in addition to my mentors and professors.

We have faith that both my professional and personal endeavours will continue to advance. The community involvement in discussion forums and self-learning within my internship stand out to me as the two distinct learning experiences that have had the biggest impacts on my development this semester.

Through the support and guidance of our mentors, we were able to create and nurture a learning and implementation environment that was genuinely positive and compassionate throughout my project experience.

Time management, organisation, discipline, and regular practice helped us become much better at self-exploration and learning. My progress with the project we were given, as well as the planning and implementation of the same, had a direct impact on our academic progress.

We have faith in our ability to advance and grow. If not for our project experience with the industry mentor, college mentors, and other interns, we would not have the knowledge or skills we do today.

CONCLUSION

Overall, working on this project was beneficial. We have learned new things, improved our skills, and met several of our learning objectives. We gained knowledge about how professionals operate. We gained knowledge of the various aspects of working.

We discovered that, as in many organisations, self-examination is crucial to a project's success. We gained more knowledge about predicting employee attrition rate and the different methods and algorithms used to do so in relation to our study.

There is still much to learn and develop. Current practices are still not standardized, and a consistent approach is being developed.

Additionally, we have learned how important it is to understand how each strategy compares to the current algorithm and in which applications. We discovered that the internship is not one-sided, but rather a way of exchanging information, suggestions, and ideas while putting them into practise to produce desired outcomes.

Additionally, the internship helped us identify our strengths and weaknesses. This assisted me in defining the knowledge and skills I possess.

We think that the time we spent reading about and researching different algorithms and the mathematics underlying them was well spent, and it helped us come up with a workable solution to create a model and predict the attrition rate of employees. We have mainly learned the value of self-motivation and effective time management. Finally, this project has provided us with fresh perspectives and inspiration to pursue a career in the field of machine learning.

Link to code and executable file

<https://github.com/IBM-EPBL/IBM-Project-17384-1659662477>