Analysis of Employee Retention rate

Internship in eSparsh Technologies Pvt Ltd(Bangalore)

Table of Contents

Abstract

We live in a time period where the past saw a sudden increase in employees in the IT sector due to the covid pandemic and the future where recession is unavoidable due to over employment. Still, the attrition rates have not affected as much as predicted due to recession. Employee attrition is a major problem faced by companies in the present years. In the perspective of the employer, despite employees having a history of multiple companies, employers are ready to recruit them for the experience of the specific employee, even if their loyalty is doubted. This research study aims to do a study on the employee attrition rates in the recent past using machine learning algorithms. The reason for employees leaving is analyzed considering factors like Designation , age, work environment, work-life balance, etc. The proposed work yields a significant performance in predicting the employee attrition. By using feature selection and exploring the different ensemble classifiers accuracy of 80% was achieved using 6 features.

Introduction

- Voluntary or involuntary departure of a working employee from a company.
- The majority of people eventually move on, or the corporation forcibly terminates them, forcing them to do so.
- When your company is struggling, it may be simpler to keep employees who leave on their own rather than to eliminate employment.
- The attrition rate has changed through time as a result of growing competition and stricter requirements for employee proficiency.
- Employee recruitment and training are very expensive processes. Companies must look for, hire, and train new staff.
- Losing experienced employees, especially high performers, is challenging to manage and has a detrimental impact on an organization's performance and success.
- The study focuses on the elements that could influence the employee attrition rate.

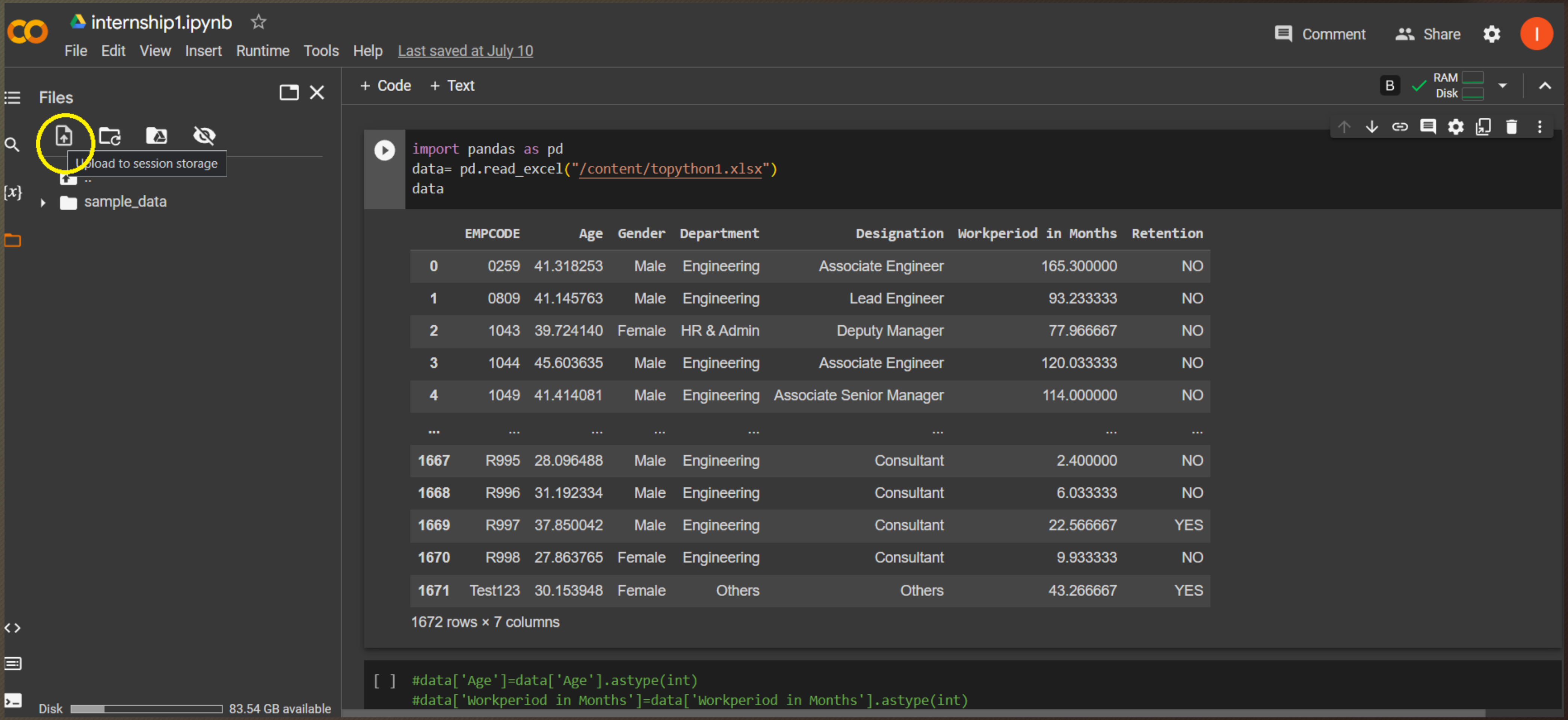
The Dataset

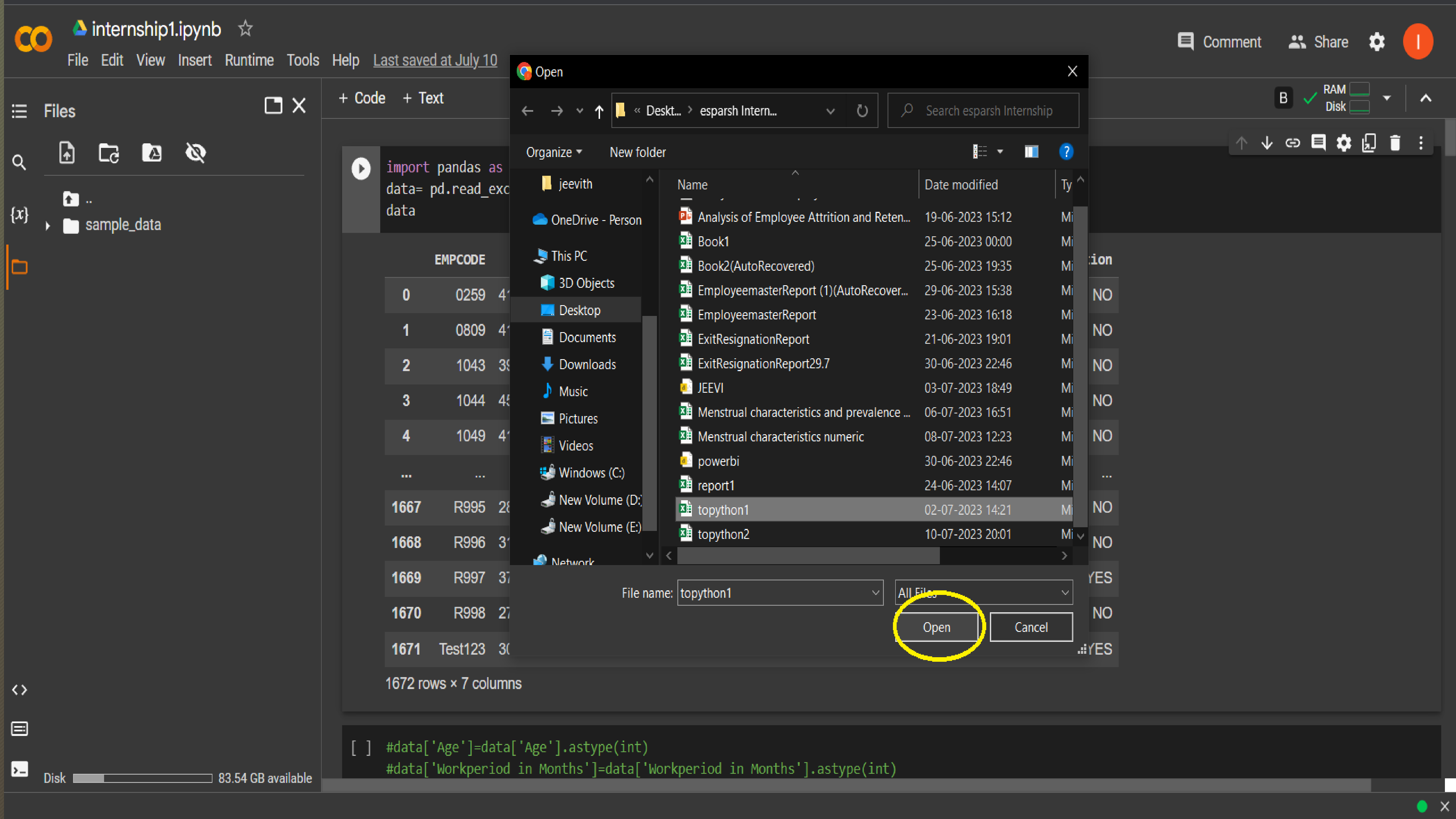
EMPCODE	Nam	BirthDat	Age	Gende	Departme	Designation	JoiningDa	LeavingDate	WORKPERIOD IN MONTHS	WORKPERIOD IN DAYS	Work Period	Retention	Reason	PermanentState
0259	Manikantan P C	07-Mar-82	41.345632	Male	Engineering	Associate Engineer	14-Nov-05	13-Jun-19	165	4959.00	13 years-6 months	NO	Higher Education	
0809	Karunakara Reddy M	09-May-82	41.173142	Male	Engineering	Lead Engineer	14-Feb-11	12-Oct-18	93	2797.00	7 years-7 months	NO	Career Growth	Karnataka
1043	NAVYA THADATIIL	10-Oct-83	39.751519	Female	HR & Admin	Deputy Manager	28-Nov-11	24-Apr-18	78	2339.00	6 years-4 months	NO	Career Growth	Karnataka
1044	SUBBAREDDY K	23-Nov-77	45.631015	Male	Engineering	Associate Engineer	05-Dec-11	14-Oct-21	120	3601.00	9 years-10 months	NO	Personal Reason	Karnataka
1049	LIJIN SHAJI	31-Jan-82	41.44146	Male	Engineering	Associate Senior Manager	09-Jan-12	21-May-21	114	3420.00	9 years-4 months	NO	Personal Reason	Kerala
1063	YOGESHA Y S	19-Oct-85	37.726832	Male	Engineering	Associate Senior Manager	27-Feb-12	06-Feb-23	133	3997.00	10 years-11 months	NO	Better Opportunity	Karnataka
1086	BASAVANNEVVA CHOUDHARI	22-Jun-86	37.05267	Female	Engineering	Lead Engineer	11-Jun-12	01-Jun-20	97	2912.00	7 years-11 months	NO	Personal Reason	
1090	GURURAJA N R	31-Dec-85	37.526959	Male	Engineering	Associate Engineer	02-Jul-12	10-Jan-19	79	2383.00	6 years-6 months	NO	Better Opportunity	
1100	PRADEEP KUMAR B	01-Jan-85	38.523589	Male	Engineering	Lead Engineer	16-Jul-12	16-Jul-18	73	2191.00	6 years-0 months	NO	Career Growth	Karnataka
1118	VIKAS G HEGDE	07-Oct-82	40.759713	Male	Engineering	Senior Manager	17-Sep-12			-41169.00		YES		Karnataka
1172	MADHUSUDAN SAHOO	20-Jun-87	36.058162	Male	Engineering	Associate Architect	11-Mar-13	30-Jul-21	102	3063.00	8 years-4 months	NO	Career Growth	Odisha
1180	VIJAYALAXMI S ULVEKAR	10-Nov-85	37.666596	Female	Engineering	Architect	01-Apr-13			-41365.00		YES		
1182	VANGARA NAGAMANI	10-Jun-81	42.085514	Female	Engineering	Architect	01-Apr-13	06-Dec-21	106	3171.00	8 years-8 months	NO	Better Opportunity	Andhra Pradesh
1209	N HARIKRISHNARAJAN	09-Oct-74	48.754244	Male	Engineering	Associate Engineer	01-Jul-13	14-Jun-19	72	2174.00	5 years-11 months	NO	Career Growth	Tamil Nadu
1217	ANANTH K L	07-Jun-88	35.091034	Male	Engineering	Associate Engineer	22-Jul-13	22-Nov-18	65	1949.00	5 years-4 months	NO	Career Growth	Karnataka
1228	ASHWINI K.G	09-Mar-91	32.340164	Female	Engineering	Senior Software Engineer	19-Aug-13	31-Dec-18	65	1960.00	5 years-4 months	NO	Career Growth	
1230	PAVAN KUNTE A	13-Mar-91	32.329213	Male	Engineering	Senior Software Engineer	19-Aug-13	23-Nov-18	64	1922.00	5 years-3 months	NO	Career Growth	Karnataka
1233	KAVYA.R	07-May-92	31.175907	Female	Engineering	Senior Software Engineer	19-Aug-13	14-Aug-18	61	1821.00	4 years-11 months	NO	Career Growth	Karnataka
	DHANARAJ.													

Attribute Description

Attribute	Description	Type
Age	Age of Employees	Numerical Discrete
Gender	Gender of the person	Categorical
Department	1-Engineer,2-IT,3-Finance,...............	Categorical
Designation	1- Associate Engineer,2-Architect,3-Assistant manger,................	Categorical
Work period in Months	Numeric	Discrete
Joining Date	Date	dd//mm//yyyy
Leaving Date	Date	dd//mm//yyyy
Retention	1- Yes, 0 - No	Categorical

Steps to upload Excel file in Google collab




internship1.ipynb
File Edit View Insert Runtime Tools Help Last saved at July 10
Comment
Share
Files
+ Code + Text
sample_data
import pandas as
data= pd.read_exc
data
Open
Deskt... › esparsh Intern...
Search esparsh Internship
Organize ▾
New folder
jeevith
OneDrive - Person
This PC
3D Objects
Desktop
Documents
Downloads
Music
Pictures
Videos
Windows (C:)
New Volume (D:)
New Volume (E:)
Network
Name
Date modified
Analysis of Employee Attrition and Reten...
19-06-2023 15:12
Book1
25-06-2023 00:00
Book2(AutoRecovered)
25-06-2023 19:35
EmployeemasterReport (1)(AutoRecover...
29-06-2023 15:38
EmployeemasterReport
23-06-2023 16:18
ExitResignationReport
21-06-2023 19:01
ExitResignationReport29.7
30-06-2023 22:46
JEEVI
03-07-2023 18:49
Menstrual characteristics and prevalence ...
06-07-2023 16:51
Menstrual characteristics numeric
08-07-2023 12:23
powerbi
30-06-2023 22:46
report1
24-06-2023 14:07
topython1
02-07-2023 14:21
topython2
10-07-2023 20:01
File name: topython1
Open
Cancel
EMPCODE
1672 rows × 7 columns
#data['Age']=data['Age'].astype(int)
#data['Workperiod in Months']=data['Workperiod in Months'].astype(int)
Disk 83.54 GB available

internship1.ipynb

File Edit View Insert Runtime Tools Help Last saved at July 10

Comment Share

Files

..
sample_data
topyth

Download
Rename file
Delete file
Copy path
Refresh

+ Code + Text

RAM
Disk

```
import pandas as pd
data= pd.read_excel("/content/topython1.xlsx")
data
```

	EMPCODE	Age	Gender	Department	Designation	Workperiod in Months	Retention
0	0259	41.318253	Male	Engineering	Associate Engineer	165.300000	NO
1	0809	41.145763	Male	Engineering	Lead Engineer	93.233333	NO
2	1043	39.724140	Female	HR & Admin	Deputy Manager	77.966667	NO
3	1044	45.603635	Male	Engineering	Associate Engineer	120.033333	NO
4	1049	41.414081	Male	Engineering	Associate Senior Manager	114.000000	NO
...	...	...	...	...	...	...	...
1667	R995	28.096488	Male	Engineering	Consultant	2.400000	NO
1668	R996	31.192334	Male	Engineering	Consultant	6.033333	NO
1669	R997	37.850042	Male	Engineering	Consultant	22.566667	YES
1670	R998	27.863765	Female	Engineering	Consultant	9.933333	NO
1671	Test123	30.153948	Female	Others	Others	43.266667	YES

1672 rows × 7 columns

```
[ ] #data['Age']=data['Age'].astype(int)
    #data['Workperiod in Months']=data['Workperiod in Months'].astype(int)
```

Disk 83.54 GB available

Data Preprocessing

- Removing the unwanted columns/fields in the dataset.
- Checking for Null values.
- Fill the Null values.
  - Data Type – Numerical – Use Mean / Median
  - Data Type – Text – Mode
- Change the categorical values to numerical.
  - Label encoder
  - Word to vector

Training and testing the Data into Model

```
[ ] from sklearn.model_selection import train_test_split
    from sklearn.ensemble import GradientBoostingClassifier
    from sklearn.metrics import accuracy_score
    from sklearn.preprocessing import LabelEncoder
```

```
[ ] # Splitting the dataset into input features (X) and target variable (y)
    X = data[['Age', 'Workperiod in Months', 'Gender', 'Department', 'Designation']]
    y = data['Retention']
```

```
[ ] X_encoded = pd.get_dummies(X)
```

```
    # Convert the target variable to numerical labels
    label_encoder = LabelEncoder()
    y = label_encoder.fit_transform(y)
    print(y)
```

```
[0 0 0 ... 1 0 1]
```

```
[ ] print(X_encoded)
```

```
[ ] # Splitting the data into training and testing sets
    X_train, X_test, y_train, y_test = train_test_split(X_encoded, y, test_size=0.2,random_state=42)#,stratify = y
```

Choose the Machine Learning Model

For the above features and the target variable we can use the following model,

- GradientBoostingClassifier – 0.794
- LogisticRegression – 0.686
- XGBClassifier – 0.8

Hence here we have high accuracy in XGB Classifier.

GradientBoostingClassifier

```
# Create an instance of GradientBoostingClassifier
model = GradientBoostingClassifier()

# Train the model
model.fit(X_train, y_train)

# Make predictions on the test set
y_pred = model.predict(X_test)

# Calculate accuracy
accuracy = accuracy_score(y_test, y_pred)
print('Accuracy:', accuracy)
```

```
Accuracy: 0.7940298507462686
```

LogisticRegression

```
from sklearn.linear_model import LogisticRegression
# Create an instance of LogisticRegression
model = LogisticRegression()

# Train the logistic regression model
model.fit(X_train, y_train)

# Make predictions on the test set using logistic regression
y_pred = model.predict(X_test)

# Calculate accuracy for logistic regression
accuracy = accuracy_score(y_test, y_pred)
print('Accuracy (Logistic Regression):', accuracy)
```

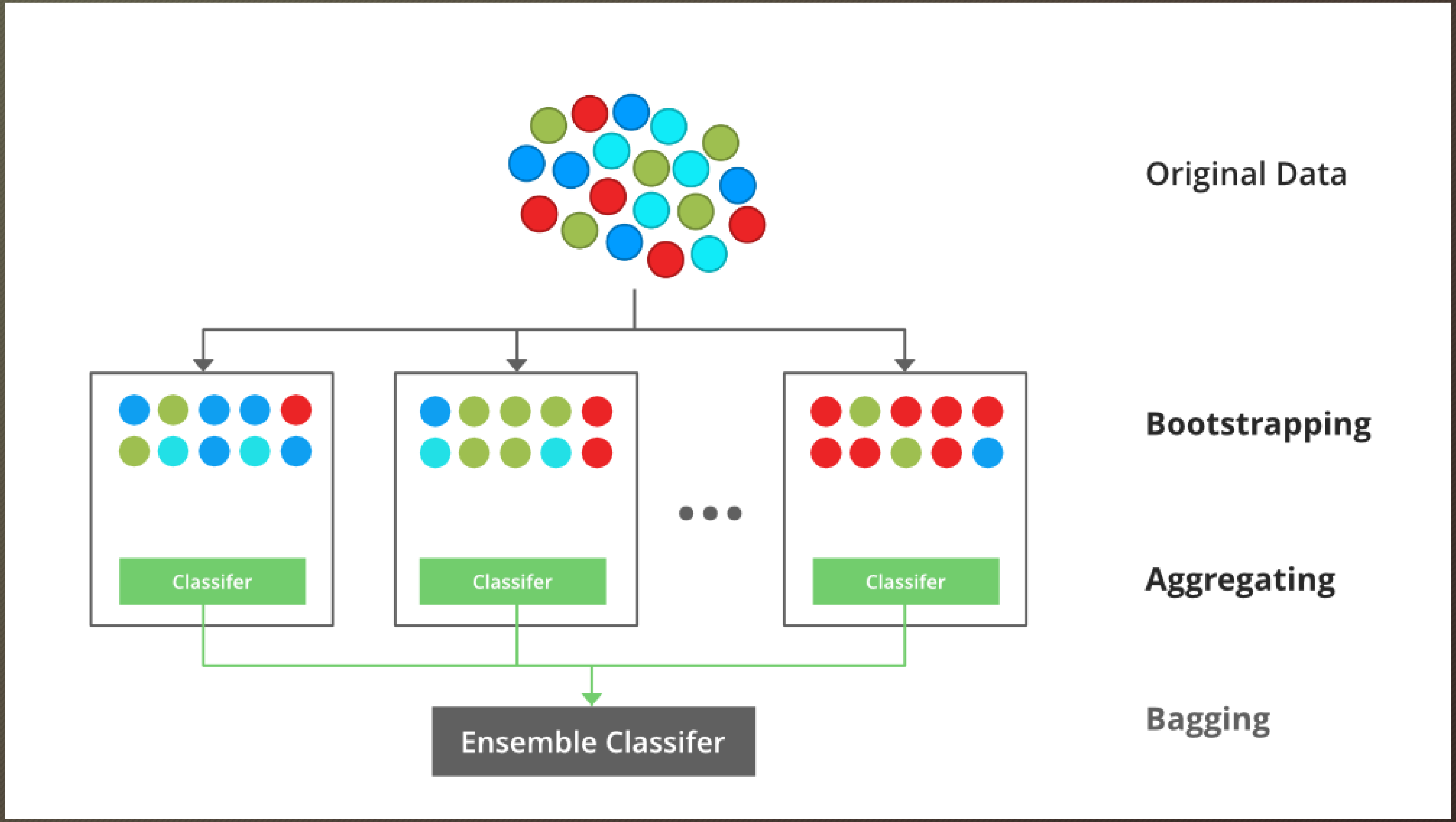
```
Accuracy (Logistic Regression): 0.6865671641791045
```

XGB Classifier

```
# Create an instance of XGBClassifier
model5 = xgb.XGBClassifier()

# Train the model
model5.fit(X_train, y_train)

# Make predictions on the test set
y_pred = model5.predict(X_test)

# Calculate accuracy
accuracy = accuracy_score(y_test, y_pred)
print('Accuracy:', accuracy)
```

```
Accuracy: 0.8
```

About the M.L Model(XGB Classifier)



- XGBClassifier uses an ensemble of weak prediction models called decision trees. It builds trees sequentially, where each subsequent tree tries to correct the mistakes made by the previous trees. This process is known as gradient boosting.
- XGBClassifier optimizes an objective function that quantifies the model's performance. The objective function measures the difference between the predicted and actual values and guides the algorithm to minimize this difference.

Deploy the Model

To deploy the model first we need to load the model as a file(.jkl) for which we are suing the joblib library function.

Save the model for further use

```
import joblib


# Save the trained model to a file
joblib.dump(model5, 'xgb_model.pkl')
```

```
['xgb_model.pkl']
```

```
[21] # Load the saved Random Forest Classifier
     model5 = joblib.load('xgb_model.pkl')
```

- After saving the model we need to get the user input.
- Then deploy the model with the user input data.

Deploying the Machine Learning Model

Where the user input is collected in the getvalue_.

```python
# Create a DataFrame from user input
#input_data = pd.DataFrame({'Age': [31],'Gender': ['Female'],'Department': ['Engineering'],
#'Designation': ['Senior Engineer'],'Workperiod in Months': [36]})

# Using the user input
input_data = pd.DataFrame({'Age': [getvalue3],
                           'Gender': [getvalue2],
                           'Department': [getvalue1],
                           'Designation': [getvalue],
                           'Workperiod in Months': [getvalue4]})

# Encoding categorical features using one-hot encoding
input_data_encoded = pd.get_dummies(input_data)
input_data_encoded = input_data_encoded.reindex(columns=X_train.columns, fill_value=0)


# Make the prediction using the trained model
retention_prediction = model5.predict(input_data_encoded)
print(retention_prediction)

# Display the predicted retention status
if retention_prediction == [1]:
    print("The employee is predicted to stay.")
else:
    print("The employee is predicted to leave.")
```

```
[0]
The employee is predicted to leave.
```

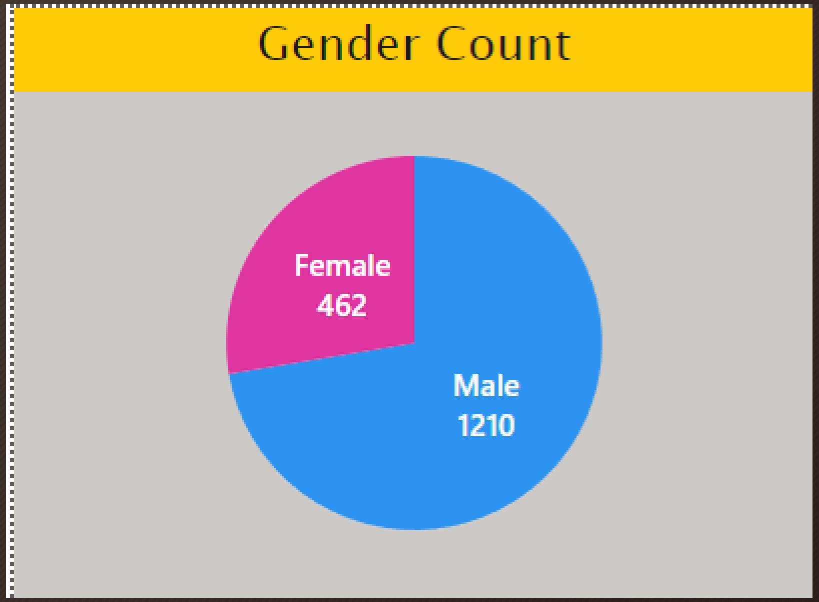
```
# Create a DataFrame from user input
#input_data = pd.DataFrame({'Age': [34],'Gender': ['Male'],'Department': ['Engineering'],
#'Designation': ['Associate Architect'],'Workperiod in Months': [114]})

# Using the user input
input_data = pd.DataFrame({'Age': [getvalue3],
                           'Gender': [getvalue2],
                           'Department': [getvalue1],
                           'Designation': [getvalue],
                           'Workperiod in Months': [getvalue4]})

# Encoding categorical features using one-hot encoding
input_data_encoded = pd.get_dummies(input_data)
input_data_encoded = input_data_encoded.reindex(columns=X_train.columns, fill_value=0)


# Make the prediction using the trained model
retention_prediction = model5.predict(input_data_encoded)
print(retention_prediction)

# Display the predicted retention status
if retention_prediction == [1]:
    print("The employee is predicted to stay.")
else:
    print("The employee is predicted to leave.")
```

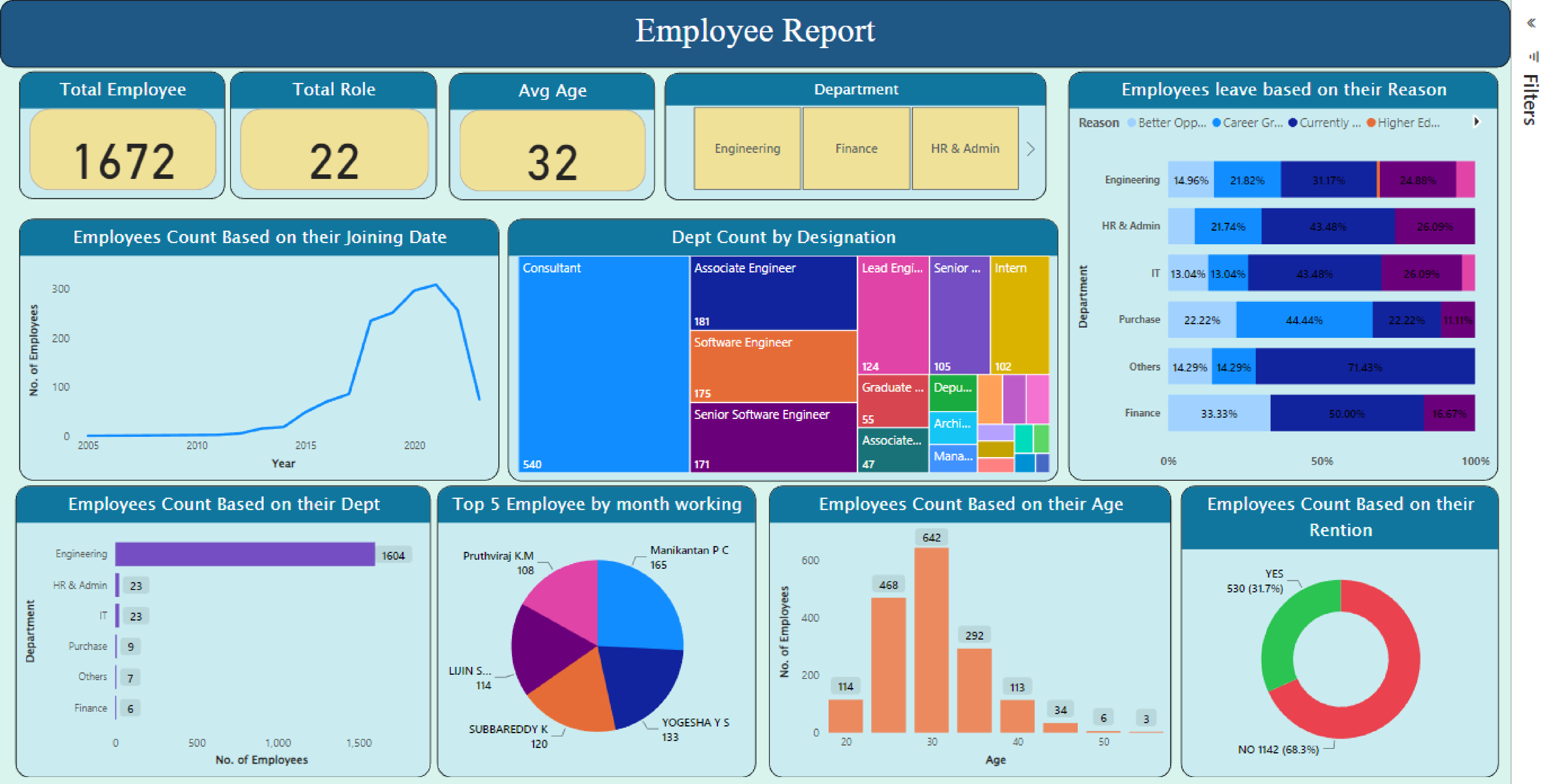
```
[1]
The employee is predicted to stay.
```

Visualization

- Data visualization is the most important step in the data analysis and prediction process, because the visuals would be easily captured by our brain then the text.
- We have used the Power BI software to visual the dataset.

- And this report is linked with the Power BI visual so we could directly have the interaction.





JEEVI, Page 1

Data updated on 11/7/23, 5:46 pm

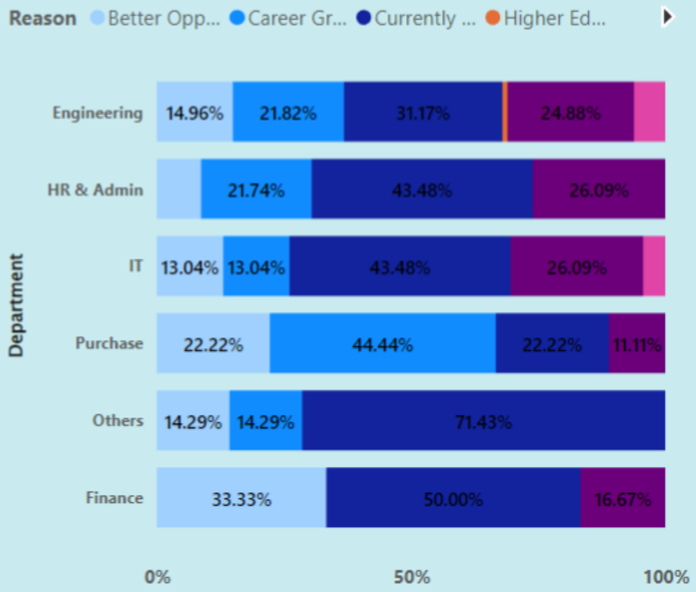
Employee Report

Total Employee	Total Role	Avg Age
1672	22	32

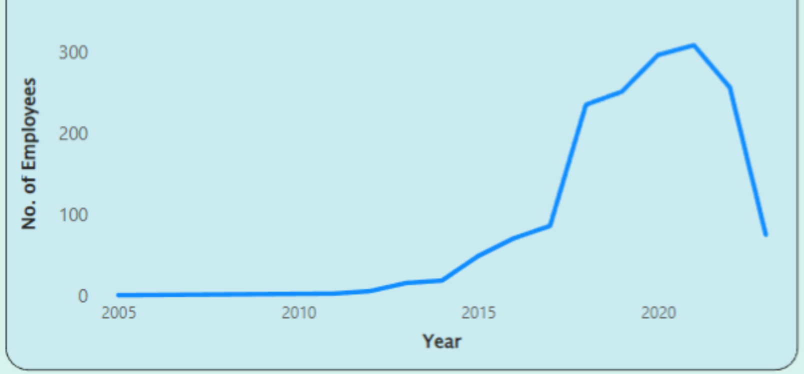
Department

Engineering | Finance | HR & Admin

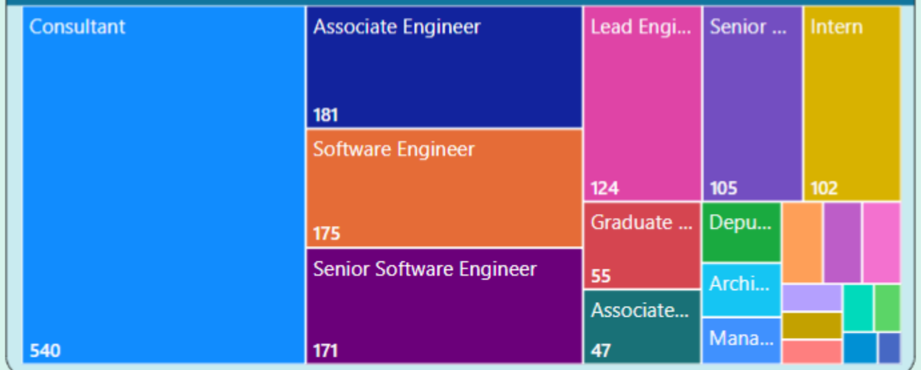
Employees leave based on their Reason



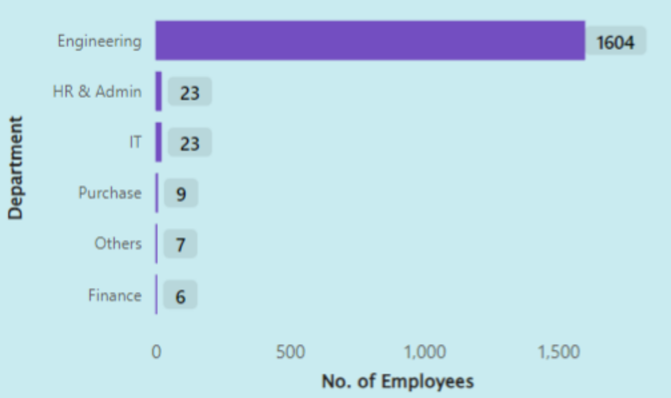
Employees Count Based on their Joining Date



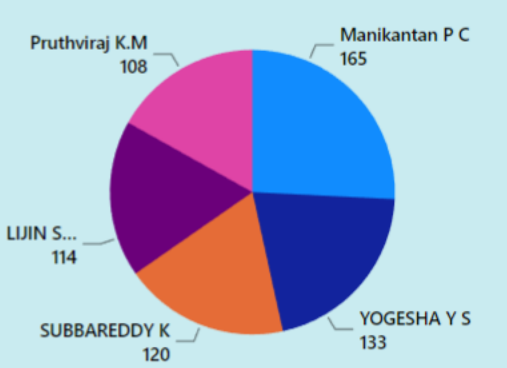
Dept Count by Designation



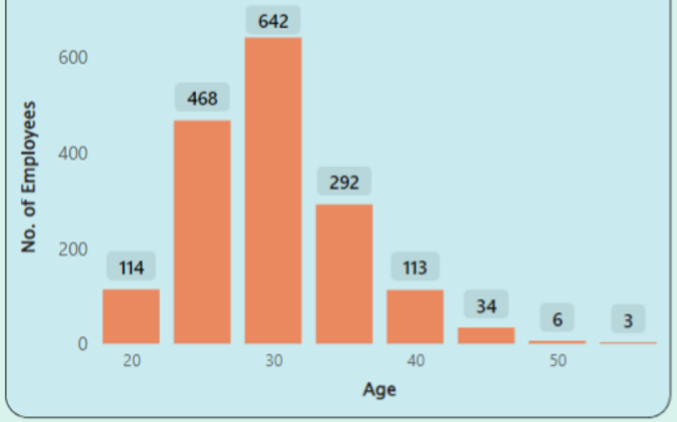
Employees Count Based on their Dept



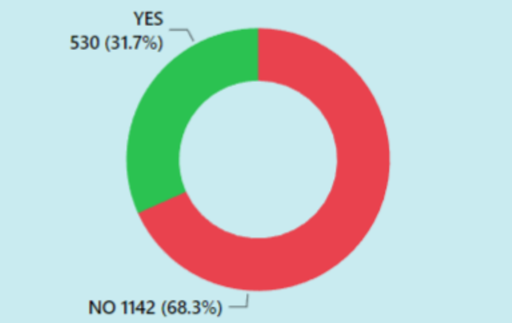
Top 5 Employee by month working



Employees Count Based on their Age

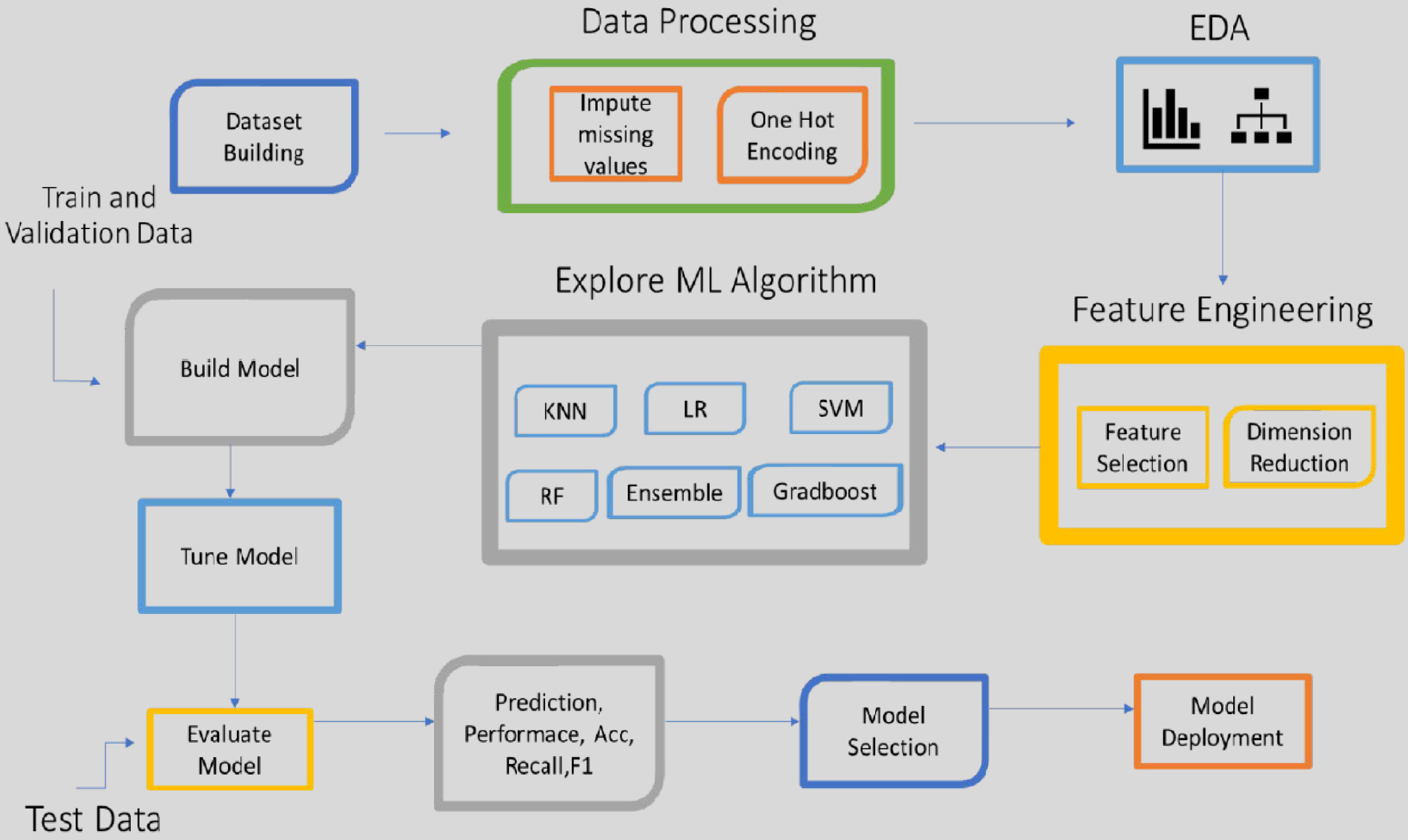


Employees Count Based on their Rention



Filters

Proposed Work- PipeLine



Conclusion

- Thus a complete end to end ML pipeline was explored for predicting the employee retention rate.
- The dataset is a good representative of the general workforce in today’s organizations. The good
- results from multiple classifiers justify that the features chosen are causes that contribute to voluntary attrition.
- The XGBoost classifier performed well than other ML algorithms with a validation accuracy of 80%
- The reason for attrition of employees can’t be exactly predicted, because each person would have different ideas for their future goals.
- Future work might include more number of attributes pertaining to the employee and a Sentiment Analysis can be made by collecting data from employees.

Thank You

- S.Jeevith Kumar