Assignment: AI & ML Role - Python Programming

Objective:

This assignment aims to evaluate your proficiency in Python programming and your understanding of key concepts in artificial intelligence (AI) and machine learning (ML).

Dataset: IBM HR Analytics Employee Attrition & Performance

1. Dataset Analysis and Preprocessing:

The first 5 rows of the dataset are:

```
#reading the data

data = pd.read_csv("WA_Fn-UseC_-HR-Employee-Attrition.csv")

data.head()
```

✓ 0.2s Python

	Age	Attrition	BusinessTravel	DailyRate	Department	DistanceFromHome	Education	EducationField	EmployeeCount	EmployeeNumber	...	RelationshipSatisfaction	Stan
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	...	4	
2	37	Yes	Travel_Rarely	1373	Research & Development	2	2	Other	1	4	...	2	
3	33	No	Travel_Frequently	1392	Research & Development	3	4	Life Sciences	1	5	...	3	
4	27	No	Travel_Rarely	591	Research & Development	2	1	Medical	1	7	...	4	

5 rows × 35 columns

The dataset contains 1470 rows and 35 columns.

The columns are:

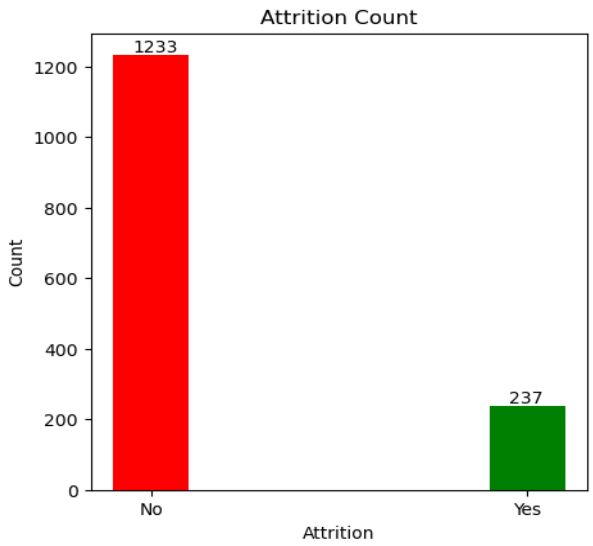
```
# column names
data.columns
```

✓ 0.5s

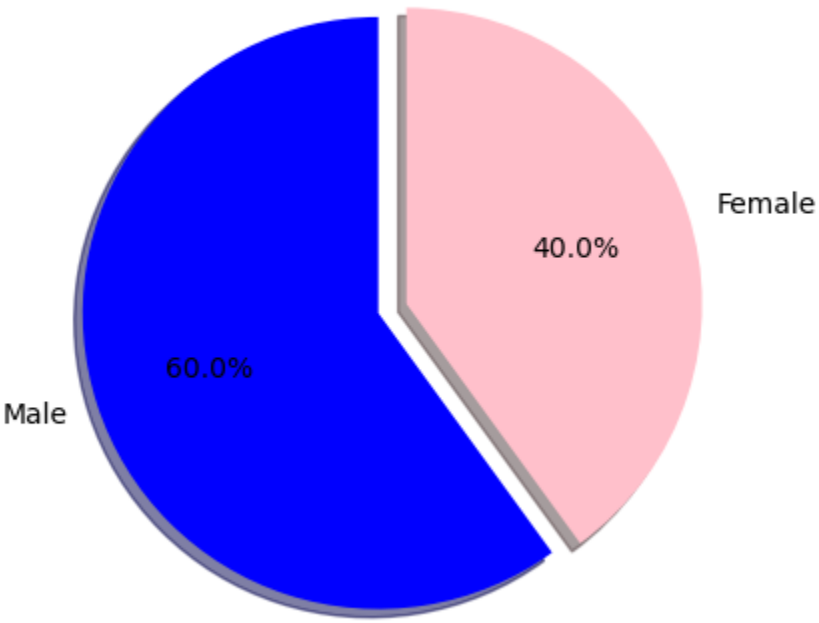
```
Index(['Age', 'Attrition', 'BusinessTravel', 'DailyRate', 'Department',
       'DistanceFromHome', 'Education', 'EducationField', 'EmployeeCount',
       'EmployeeNumber', 'EnvironmentSatisfaction', 'Gender', 'HourlyRate',
       'JobInvolvement', 'JobLevel', 'JobRole', 'JobSatisfaction',
       'MaritalStatus', 'MonthlyIncome', 'MonthlyRate', 'NumCompaniesWorked',
       'Over18', 'OverTime', 'PercentSalaryHike', 'PerformanceRating',
       'RelationshipSatisfaction', 'StandardHours', 'StockOptionLevel',
       'TotalWorkingYears', 'TrainingTimesLastYear', 'WorkLifeBalance',
       'YearsAtCompany', 'YearsInCurrentRole', 'YearsSinceLastPromotion',
       'YearsWithCurrManager'],
      dtype='object')
```

There are no missing values in the data.

The dataset is not a balanced dataset. The target variable Attrition has 1233 ”No” class labels and 237 ”Yes” class labels.

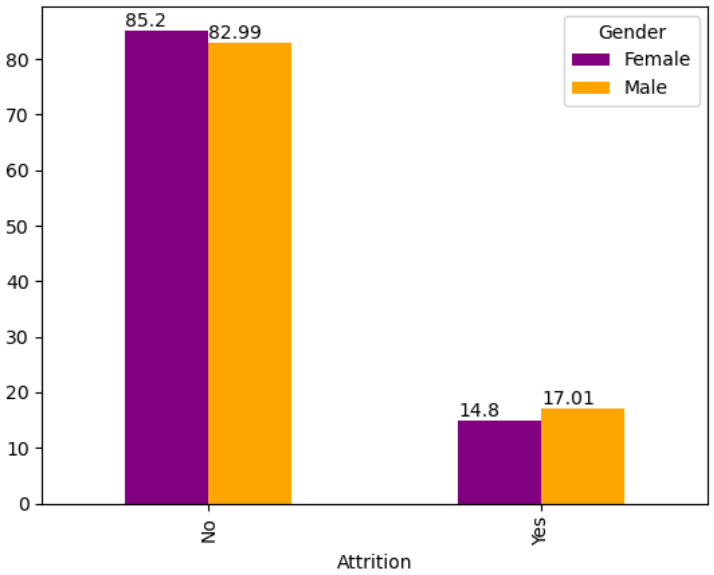


The number of males in the data is greater than the number of females. The data contains 60% of male employees and 40% of female employees.

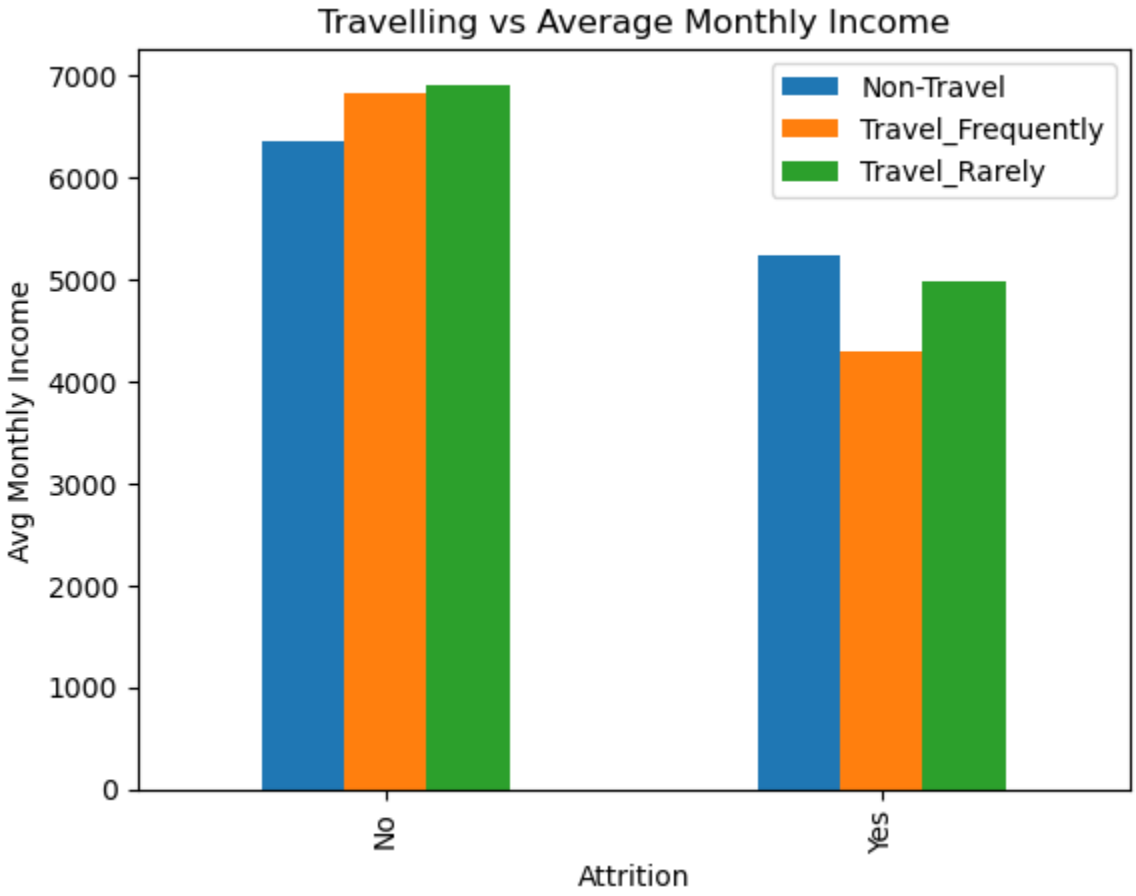


Attrition vs Gender:

Out of 100% of male and female employees, 14.8 percentage of female employees and 17.01 percentage of male employees left the company.

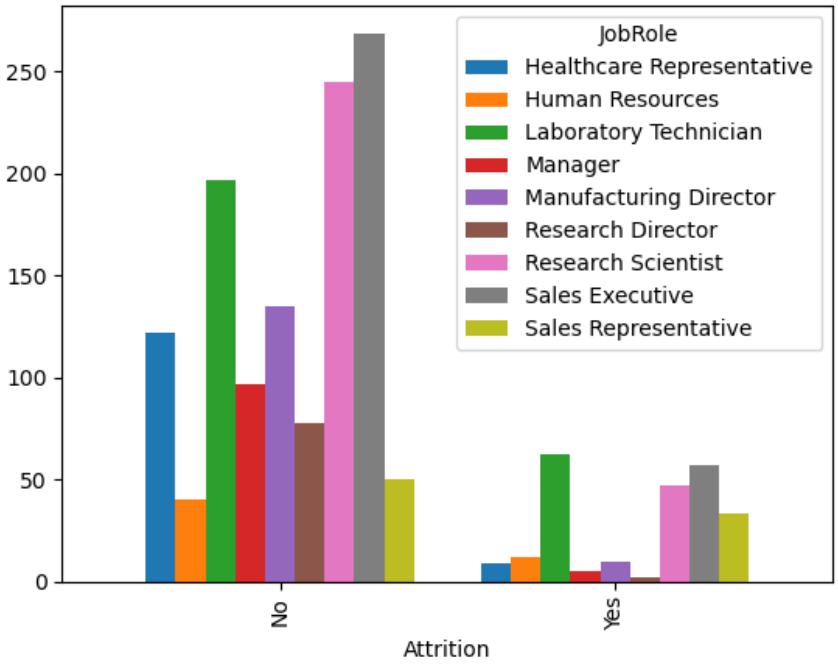


Travelling vs Average Monthly Income:



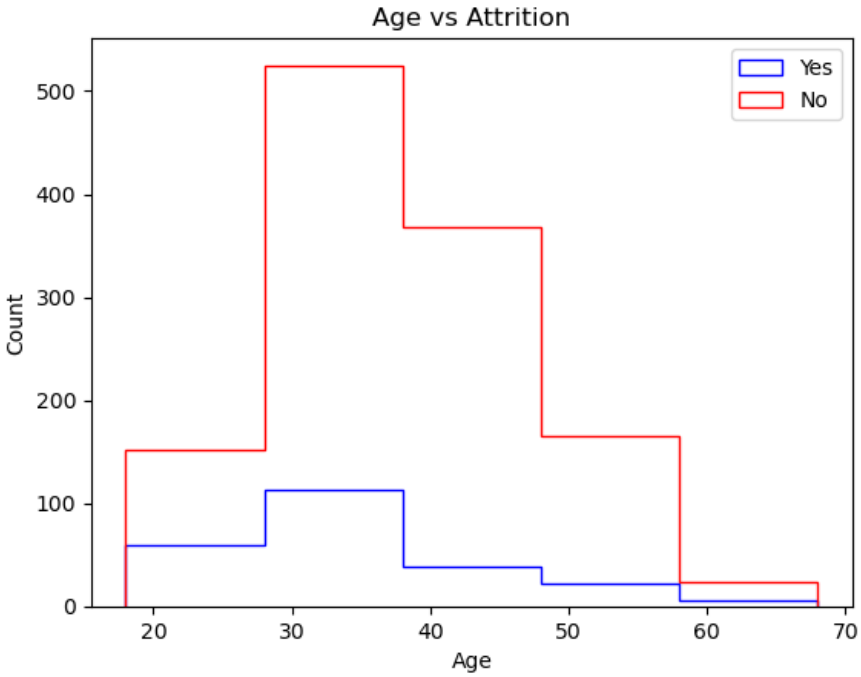
Most of the employees, who have less income and travel will likely to leave the company.

Department vs Attrition:



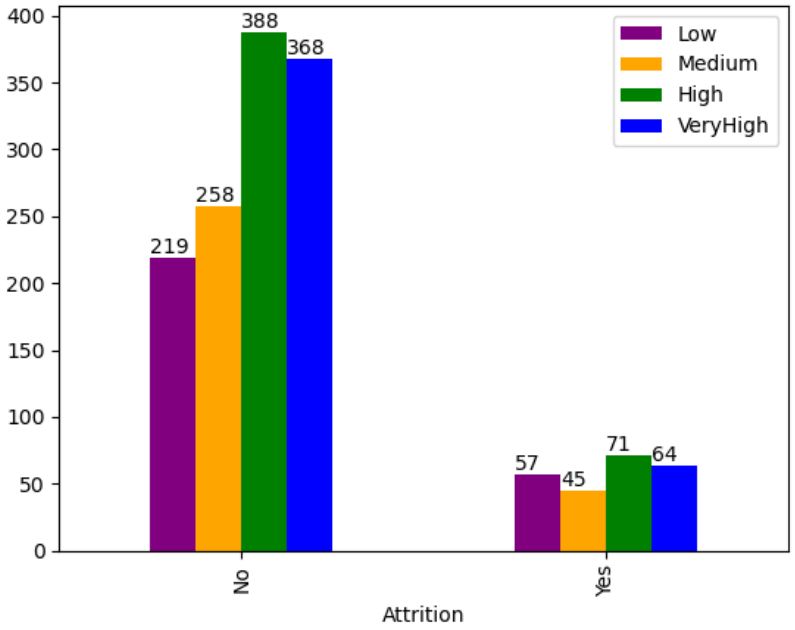
Laboratory Technician and Sales Executive Department Employees attrition rate is more.

Age vs Attrition:

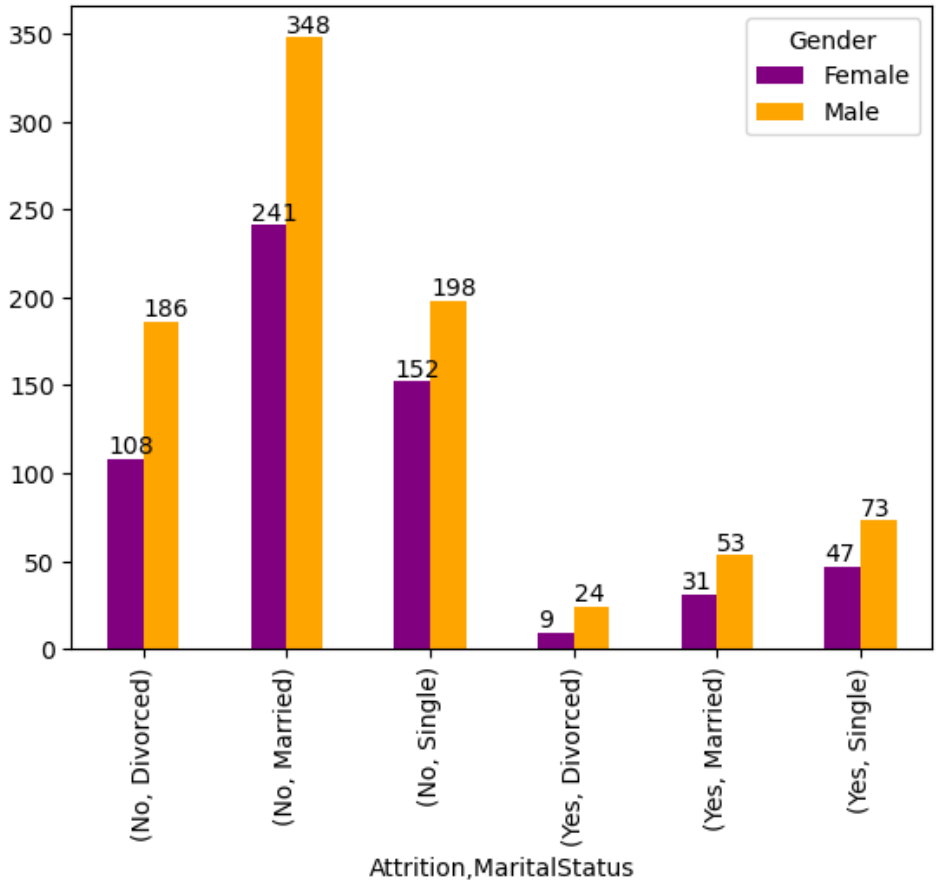


The employees in the age of 30-40 have more attrition rates. The employees in the age of 60-70 have very less attrition rates.

RelationshipSatisfaction vs Attrition



Employees with most relationship satisfaction remain in the company. This show how relationship affects the mental stability of the employees.



Unmarried employees are likely to leave the company. Married employees have less attrition rates.

2. Model Development:

The dataset is an imbalanced dataset. So first we make it balanced. We used imblearn module to achieve this. We upsampled the number of “Yes” labelled attrition rows using SMOTE(Synthetic Minority Over-sampling Tecnique) . The final number of rows we have is 1233+1233=2466. Now we split the data into training and test sets using train_test_split. The train set contains 1972 and the test set contains 494 rows.

Here, we used four classifiers RandomForestClassifier, LogisticRegression, Decision Tree Classifier, Support Vector Machine Classifier.
The results we got after training the base models are:

```
Model : Random Forest Classifier
Train Error : {'accuracy': 1.0, 'precision': 1.0, 'recall': 1.0, 'f1-score': 1.0}
Test Error : {'accuracy': 0.9291497975708503, 'precision': 0.9646017699115044, 'recall': 0.889795918367347, 'f1-score': 0.9256900212314225}


Model : Decision Tree Classifier
Train Error : {'accuracy': 1.0, 'precision': 1.0, 'recall': 1.0, 'f1-score': 1.0}
Test Error : {'accuracy': 0.8259109311740891, 'precision': 0.8142292490118577, 'recall': 0.8408163265306122, 'f1-score': 0.8273092369477911}


Model : Logistic Regression
Train Error : {'accuracy': 0.8620689655172413, 'precision': 0.8841201716738197, 'recall': 0.8340080971659919, 'f1-score': 0.8583333333333333}
Test Error : {'accuracy': 0.8481781376518218, 'precision': 0.8695652173913043, 'recall': 0.8163265306122449, 'f1-score': 0.8421052631578947}


Model : Support Vector Classifier
Train Error : {'accuracy': 0.8296146044624746, 'precision': 0.8438818565400844, 'recall': 0.8097165991902834, 'f1-score': 0.8264462809917356}
Test Error : {'accuracy': 0.8279352226720648, 'precision': 0.8539823008849557, 'recall': 0.7877551020408163, 'f1-score': 0.8195329087048833}
```

Since the base model RandomForestClassifier performed well on test data, we chose that model as our final model.

3. Model Evaluation and Optimization:

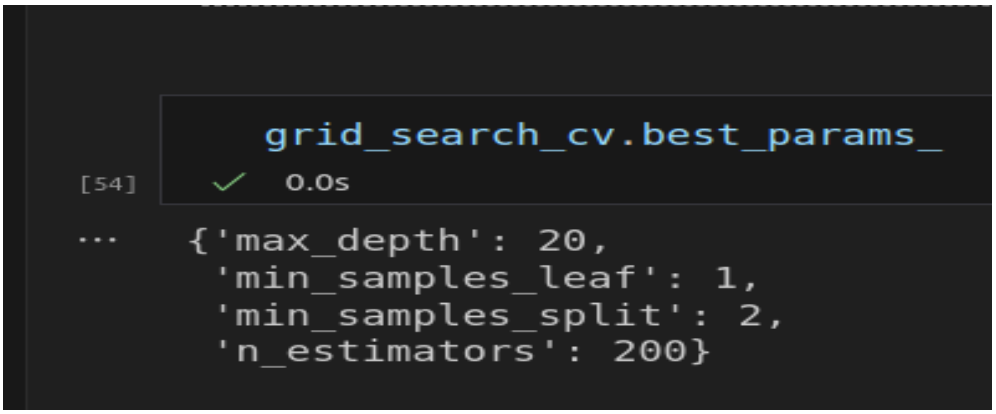
Here, we used GridSearchCV for tuning our RandomForestClassifier.

```
from sklearn.model_selection import GridSearchCV
param_grid = {
'n_estimators': [50, 100, 200],
'max_depth': [10, 20, 30],
'min_samples_split': [2, 5, 10],
'min_samples_leaf': [1, 2, 4]
}

grid_search_cv = GridSearchCV(rc, param_grid, cv=5)
```

```
grid_search_cv.fit(x_train, y_train)
```

The optimal model we got is



The final accuracy of the model is: 91.90%.

We also tried with neural networks and used keras-tuner for hyperparameter tuning. We got 88.25 % with this neural network. So we stick with RandomForestClassifier model with 91.90 % test accuracy.

The reasons for most of the employees leaving the company are Environment conditions, Travelling Conditions, Relationship Pressure. If the salary increment is less the employee is likely to leave the company. The managers under which the employees work are also one the reason for attrition. Over hours working and less salary this also leads to attrition. Pay more for overtime employees. On the other hand Departments also should look at their attrition rates, Reasearch departement has high attrition rates.