attribution-data-hr

May 28, 2024

[1]:
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import classification_report, confusion_matrix,␣
↪accuracy_score
```

[2]:
```
# Load the dataset
data = pd.read_csv(r'C:\Users\pc\Downloads\Attrition data.csv')
```

[3]:
```
# Data Cleaning
data.dropna(inplace=True)
```

[4]:
```
# Exploratory Data Analysis
# Descriptive statistics
print(data.describe())

# Distribution of age
plt.figure(figsize=(10, 6))
sns.histplot(data['Age'], bins=30, kde=True)
plt.title('Age Distribution')
plt.xlabel('Age')
plt.ylabel('Frequency')
plt.show()
```

```
         EmployeeID          Age  DistanceFromHome    Education  EmployeeCount  \
count   4300.000000  4300.000000       4300.000000  4300.000000         4300.0
mean    2211.695116    36.926977          9.197907     2.913256            1.0
std     1272.117692     9.146517          8.097059     1.024774            0.0
min        1.000000    18.000000          1.000000     1.000000            1.0
25%     1110.750000    30.000000          2.000000     2.000000            1.0
50%     2215.500000    36.000000          7.000000     3.000000            1.0
75%     3314.250000    43.000000         14.000000     4.000000            1.0
max     4409.000000    60.000000         29.000000     5.000000            1.0

          JobLevel  MonthlyIncome  NumCompaniesWorked  PercentSalaryHike  \
count   4300.000000    4300.000000         4300.000000        4300.000000
```

mean	2.066977	65059.844186	2.690000	15.210698
std	1.106633	47045.398914	2.495764	3.662777
min	1.000000	10090.000000	0.000000	11.000000
25%	1.000000	29260.000000	1.000000	12.000000
50%	2.000000	49360.000000	2.000000	14.000000
75%	3.000000	83802.500000	4.000000	18.000000
max	5.000000	199990.000000	9.000000	25.000000

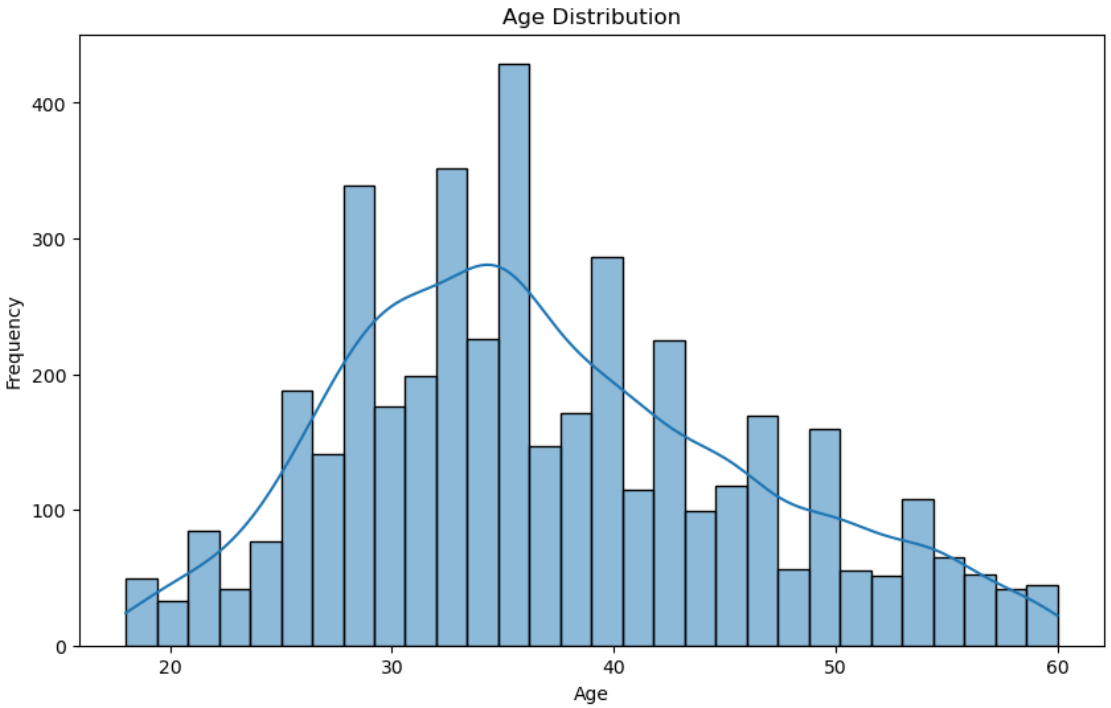
	StandardHours	…	TotalWorkingYears	TrainingTimesLastYear	\
count	4300.0	…	4300.000000	4300.000000	
mean	8.0	…	11.285116	2.796279	
std	0.0	…	7.790052	1.290142	
min	8.0	…	0.000000	0.000000	
25%	8.0	…	6.000000	2.000000	
50%	8.0	…	10.000000	3.000000	
75%	8.0	…	15.000000	3.000000	
max	8.0	…	40.000000	6.000000	

	YearsAtCompany	YearsSinceLastPromotion	YearsWithCurrManager	\
count	4300.000000	4300.000000	4300.000000	
mean	7.026047	2.190000	4.132558	
std	6.148036	3.230818	3.565831	
min	0.000000	0.000000	0.000000	
25%	3.000000	0.000000	2.000000	
50%	5.000000	1.000000	3.000000	
75%	9.250000	3.000000	7.000000	
max	40.000000	15.000000	17.000000	

	EnvironmentSatisfaction	JobSatisfaction	WorkLifeBalance	\
count	4300.000000	4300.000000	4300.000000	
mean	2.723953	2.724884	2.761163	
std	1.093802	1.101875	0.707800	
min	1.000000	1.000000	1.000000	
25%	2.000000	2.000000	2.000000	
50%	3.000000	3.000000	3.000000	
75%	4.000000	4.000000	3.000000	
max	4.000000	4.000000	4.000000	

	JobInvolvement	PerformanceRating
count	4300.000000	4300.000000
mean	2.728837	3.153953
std	0.710769	0.360946
min	1.000000	3.000000
25%	2.000000	3.000000
50%	3.000000	3.000000
75%	3.000000	3.000000
max	4.000000	4.000000

```
[8 rows x 21 columns]
```

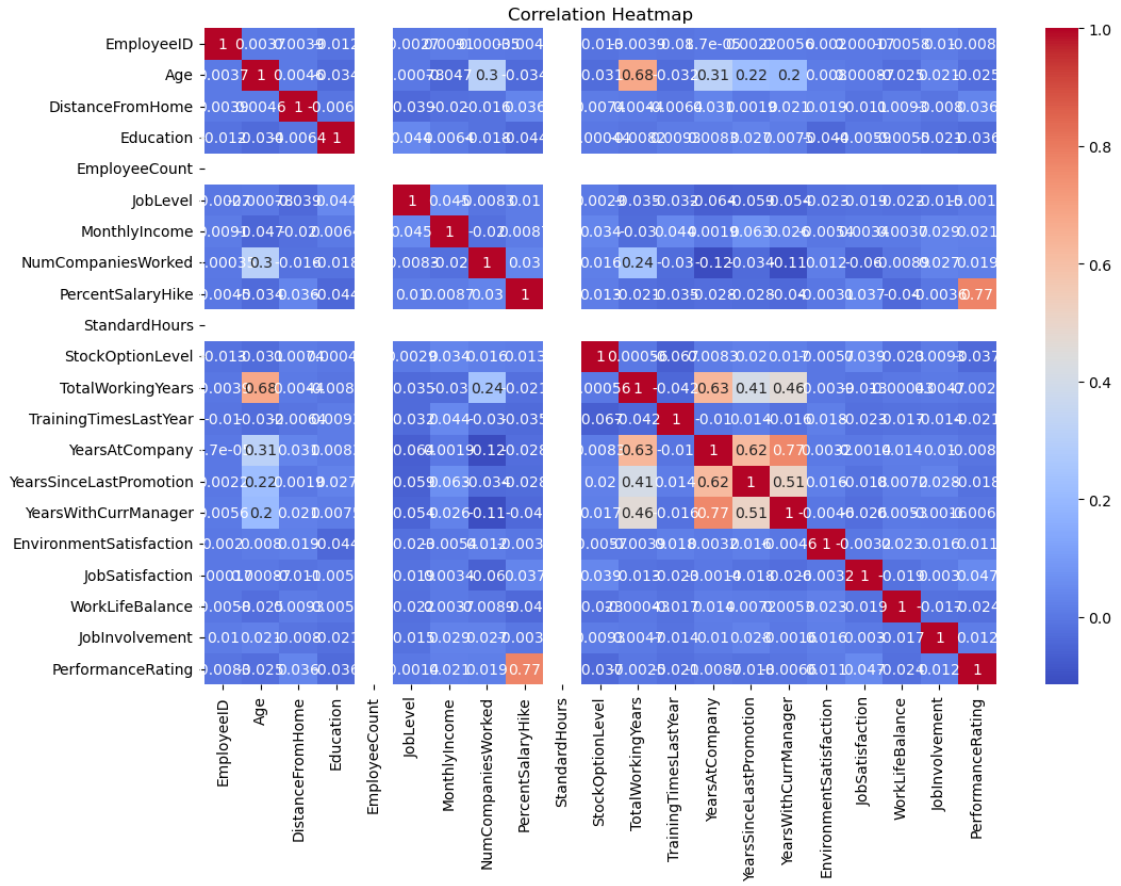


[5]:
```
# Attrition rate over time
attrition_rate = data['Attrition'].value_counts(normalize=True) * 100
print(f'Attrition Rate: {attrition_rate}')
```

```
Attrition Rate: No     83.837209
Yes    16.162791
Name: Attrition, dtype: float64
```

[6]:
```
# Correlation heatmap
plt.figure(figsize=(12, 8))
sns.heatmap(data.corr(), annot=True, cmap='coolwarm')
plt.title('Correlation Heatmap')
plt.show()
```

```
C:\Users\pc\AppData\Local\Temp\ipykernel_12452\653016383.py:3: FutureWarning:
The default value of numeric_only in DataFrame.corr is deprecated. In a future
version, it will default to False. Select only valid columns or specify the
value of numeric_only to silence this warning.
  sns.heatmap(data.corr(), annot=True, cmap='coolwarm')
```



[7]:
```
# Key Factors Influencing Attrition
# Convert categorical variables to numeric
data=pd.read_csv(r'C:\Users\pc\Downloads\Attrition data.csv')
```

[8]:
```
data
```

[8]:

	EmployeeID	Age	Attrition	BusinessTravel	Department	\
0	1	51	No	Travel_Rarely	Sales	
1	2	31	Yes	Travel_Frequently	Research & Development	
2	3	32	No	Travel_Frequently	Research & Development	
3	4	38	No	Non-Travel	Research & Development	
4	5	32	No	Travel_Rarely	Research & Development	
…	…	…	…	…	…	
4405	4406	42	No	Travel_Rarely	Research & Development	
4406	4407	29	No	Travel_Rarely	Research & Development	
4407	4408	25	No	Travel_Rarely	Research & Development	
4408	4409	42	No	Travel_Rarely	Sales	
4409	4410	40	No	Travel_Rarely	Research & Development	

```
      DistanceFromHome  Education EducationField  EmployeeCount  Gender  …  \
0                    6          2  Life Sciences              1  Female  …
1                   10          1  Life Sciences              1  Female  …
2                   17          4          Other              1    Male  …
3                    2          5  Life Sciences              1    Male  …
4                   10          1        Medical              1    Male  …
…                  …          …              …              …      …  …
4405                 5          4        Medical              1  Female  …
4406                 2          4        Medical              1    Male  …
4407                25          2  Life Sciences              1    Male  …
4408                18          2        Medical              1    Male  …
4409                28          3        Medical              1    Male  …

      TotalWorkingYears TrainingTimesLastYear YearsAtCompany  \
0                   1.0                     6              1
1                   6.0                     3              5
2                   5.0                     2              5
3                  13.0                     5              8
4                   9.0                     2              6
…                  …                       …              …
4405               10.0                     5              3
4406               10.0                     2              3
4407                5.0                     4              4
4408               10.0                     2              9
4409                NaN                     6             21

      YearsSinceLastPromotion  YearsWithCurrManager EnvironmentSatisfaction  \
0                           0                     0                      3.0
1                           1                     4                      3.0
2                           0                     3                      2.0
3                           7                     5                      4.0
4                           0                     4                      4.0
…                          …                     …                      …
4405                        0                     2                      4.0
4406                        0                     2                      4.0
4407                        1                     2                      1.0
4408                        7                     8                      4.0
4409                        3                     9                      1.0

      JobSatisfaction  WorkLifeBalance  JobInvolvement  PerformanceRating
0                 4.0              2.0               3                  3
1                 2.0              4.0               2                  4
2                 2.0              1.0               3                  3
3                 4.0              3.0               2                  3
4                 1.0              3.0               3                  3
…                …                …                 …                  …
4405              1.0              3.0               3                  3
```

```
4406              4.0              3.0              2                3
4407              3.0              3.0              3                4
4408              1.0              3.0              2                3
4409              3.0              NaN              4                3

[4410 rows x 29 columns]
```

[9]:
```
X = data.drop('Attrition',axis=1)
y = data['Attrition']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3,␣
 ↪random_state=42)
```

[10]:
```
# Ensure the 'Attrition' column is present
if 'Attrition' not in data.columns:
    raise KeyError("The 'Attrition' column is not found in the dataset")

# Data Cleaning
data.dropna(inplace=True)

# Convert 'Attrition' column to numeric (assuming 'Yes' = 1, 'No' = 0)
data['Attrition'] = data['Attrition'].apply(lambda x: 1 if x == 'Yes' else 0)

# Identify categorical columns
categorical_cols = data.select_dtypes(include=['object']).columns

# Convert categorical variables to numeric using pd.get_dummies()
data = pd.get_dummies(data, columns=categorical_cols, drop_first=True)

# Split the data into training and testing sets
X = data.drop('Attrition', axis=1)
y = data['Attrition']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3,␣
 ↪random_state=42)

# Train a Random Forest model
model = RandomForestClassifier(n_estimators=100, random_state=42)
model.fit(X_train, y_train)

# Model evaluation
y_pred = model.predict(X_test)
print('Accuracy:', accuracy_score(y_test, y_pred))
print('Confusion Matrix:\n', confusion_matrix(y_test, y_pred))
print('Classification Report:\n', classification_report(y_test, y_pred))
```

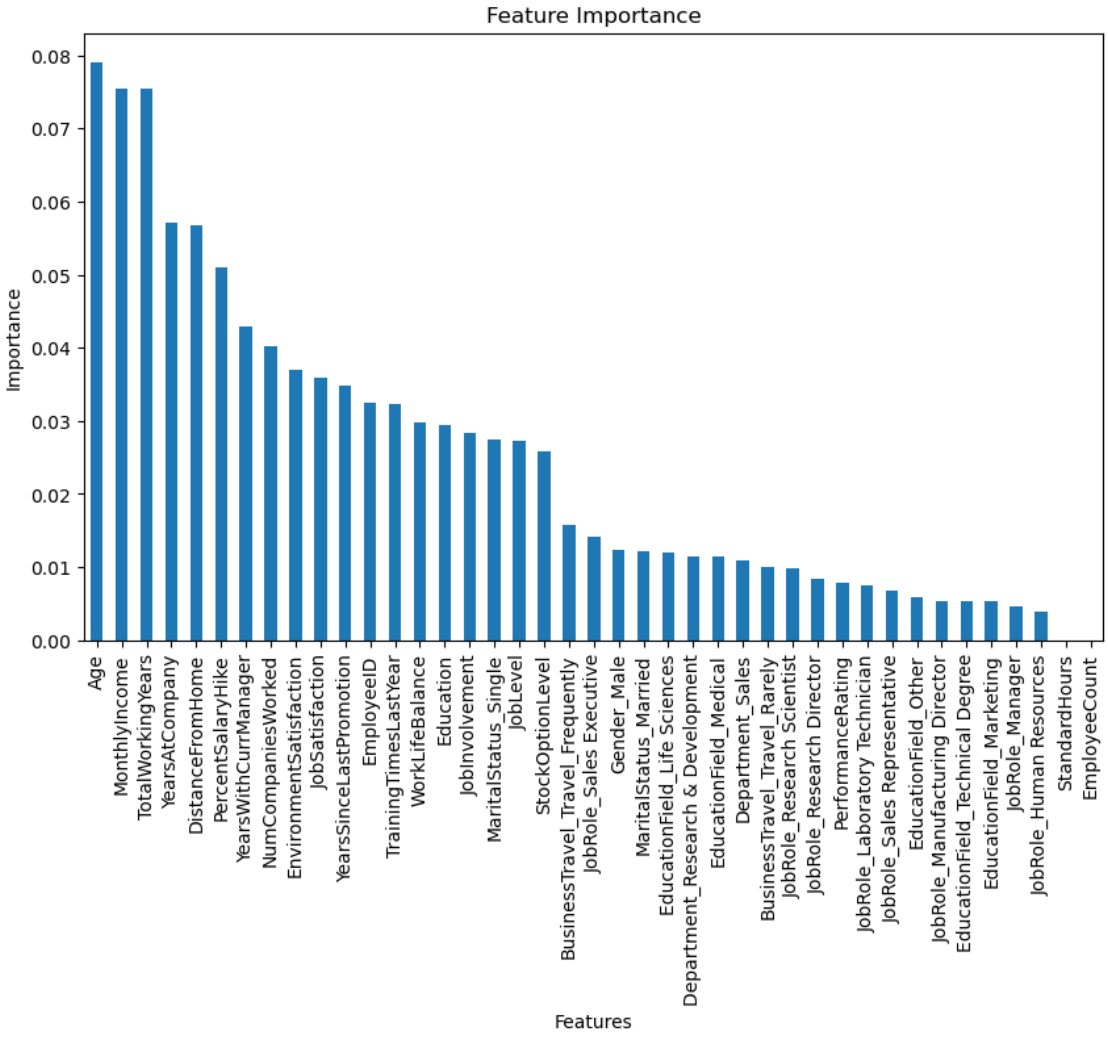
```
Accuracy: 0.9775193798449613
Confusion Matrix:
 [[1076    5]
```

```
 [  24  185]]
Classification Report:
              precision    recall  f1-score   support

           0       0.98      1.00      0.99      1081
           1       0.97      0.89      0.93       209

    accuracy                           0.98      1290
   macro avg       0.98      0.94      0.96      1290
weighted avg       0.98      0.98      0.98      1290
```

[11]:
```
# Feature importance
feature_importance = pd.Series(model.feature_importances_, index=X.columns).
 ↪sort_values(ascending=False)
print('Feature Importance:\n', feature_importance)

# Visualize feature importance
plt.figure(figsize=(10, 6))
feature_importance.plot(kind='bar')
plt.title('Feature Importance')
plt.xlabel('Features')
plt.ylabel('Importance')
plt.show()
```

```
Feature Importance:
 Age                                 0.079104
MonthlyIncome                        0.075500
TotalWorkingYears                    0.075389
YearsAtCompany                       0.057098
DistanceFromHome                     0.056803
PercentSalaryHike                    0.050964
YearsWithCurrManager                 0.042979
NumCompaniesWorked                   0.040158
EnvironmentSatisfaction              0.037003
JobSatisfaction                      0.035980
YearsSinceLastPromotion              0.034843
EmployeeID                           0.032481
TrainingTimesLastYear                0.032308
WorkLifeBalance                      0.029723
Education                            0.029483
JobInvolvement                       0.028358
MaritalStatus_Single                 0.027537
JobLevel                             0.027302
StockOptionLevel                     0.025752
BusinessTravel_Travel_Frequently     0.015757
JobRole_Sales Executive              0.014108
Gender_Male                          0.012379
```

```
MaritalStatus_Married                0.012122
EducationField_Life Sciences         0.011912
Department_Research & Development    0.011460
EducationField_Medical               0.011458
Department_Sales                     0.010956
BusinessTravel_Travel_Rarely         0.010070
JobRole_Research Scientist           0.009865
JobRole_Research Director            0.008327
PerformanceRating                    0.007911
JobRole_Laboratory Technician        0.007438
JobRole_Sales Representative         0.006854
EducationField_Other                 0.005913
JobRole_Manufacturing Director       0.005411
EducationField_Technical Degree      0.005346
EducationField_Marketing             0.005263
JobRole_Manager                      0.004698
JobRole_Human Resources              0.003987
StandardHours                        0.000000
EmployeeCount                        0.000000
dtype: float64
```


Feature Importance
Importance
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0.08
Age
MonthlyIncome
TotalWorkingYears
YearsAtCompany
DistanceFromHome
PercentSalaryHike
YearsWithCurrManager
NumCompaniesWorked
EnvironmentSatisfaction
JobSatisfaction
YearsSinceLastPromotion
EmployeeID
TrainingTimesLastYear
WorkLifeBalance
Education
JobInvolvement
MaritalStatus_Single
JobLevel
StockOptionLevel
BusinessTravel_Travel_Frequently
JobRole_Sales Executive
Gender_Male
MaritalStatus_Married
EducationField_Life Sciences
Department_Research & Development
EducationField_Medical
Department_Sales
BusinessTravel_Travel_Rarely
JobRole_Research Scientist
JobRole_Research Director
PerformanceRating
JobRole_Laboratory Technician
JobRole_Sales Representative
EducationField_Other
JobRole_Manufacturing Director
EducationField_Technical Degree
EducationField_Marketing
JobRole_Manager
JobRole_Human Resources
StandardHours
EmployeeCount
Features