ML PROJECT-2 (SEM-IV)

BY-
DEVIKA JAIN (01101192023)

TO - DEBENDRA DHIR SIR

Predicting Employee Attrition Using Machine Learning

Objective: Develop a machine learning model to predict employee attrition for proactive HR strategies.

Dataset: IBM HR Analytics Employee Attrition Dataset
(IBM HR Analytics Employee Attrition & Performance | Kaggle)

Tech Stack: Python, Pandas, Scikit-learn, Seaborn, Matplotlib

Data Understanding & Preparation

Data Understanding

- Countplot shows imbalance in Attrition (More 'No' than 'Yes')
- Correlation heatmap reveals weak linear correlation between most features
- Job Satisfaction countplot shows lower satisfaction is linked to higher attrition

Data Preparation

- Dropped constant or irrelevant columns (`EmployeeNumber`, `Over18`, etc.)
- Label Encoding for categorical variables
- Imputation of missing values using mean. **Why Mean?**
    - Suitable for **numerical features** with **normal or near-normal distribution**
    - Maintains the **central tendency** of the data
    - Efficient and simple to implement

    Why Not Median/Mode?
    - Median is better for **skewed data** or when **outliers** are present
    - Mode is used for **categorical features**, not ideal for continuous data

    Result: Ensures data completeness without distorting underlying patterns
- Scaling using StandardScaler
- **Handled Class Imbalance** using SMOTE
- Feature selection with `SelectKBest` (Top 10 features)

Feature Selection Technique & Justification

Method Used: `SelectKBest` with **ANOVA F-test** (`f_classif`)

Why This Method?

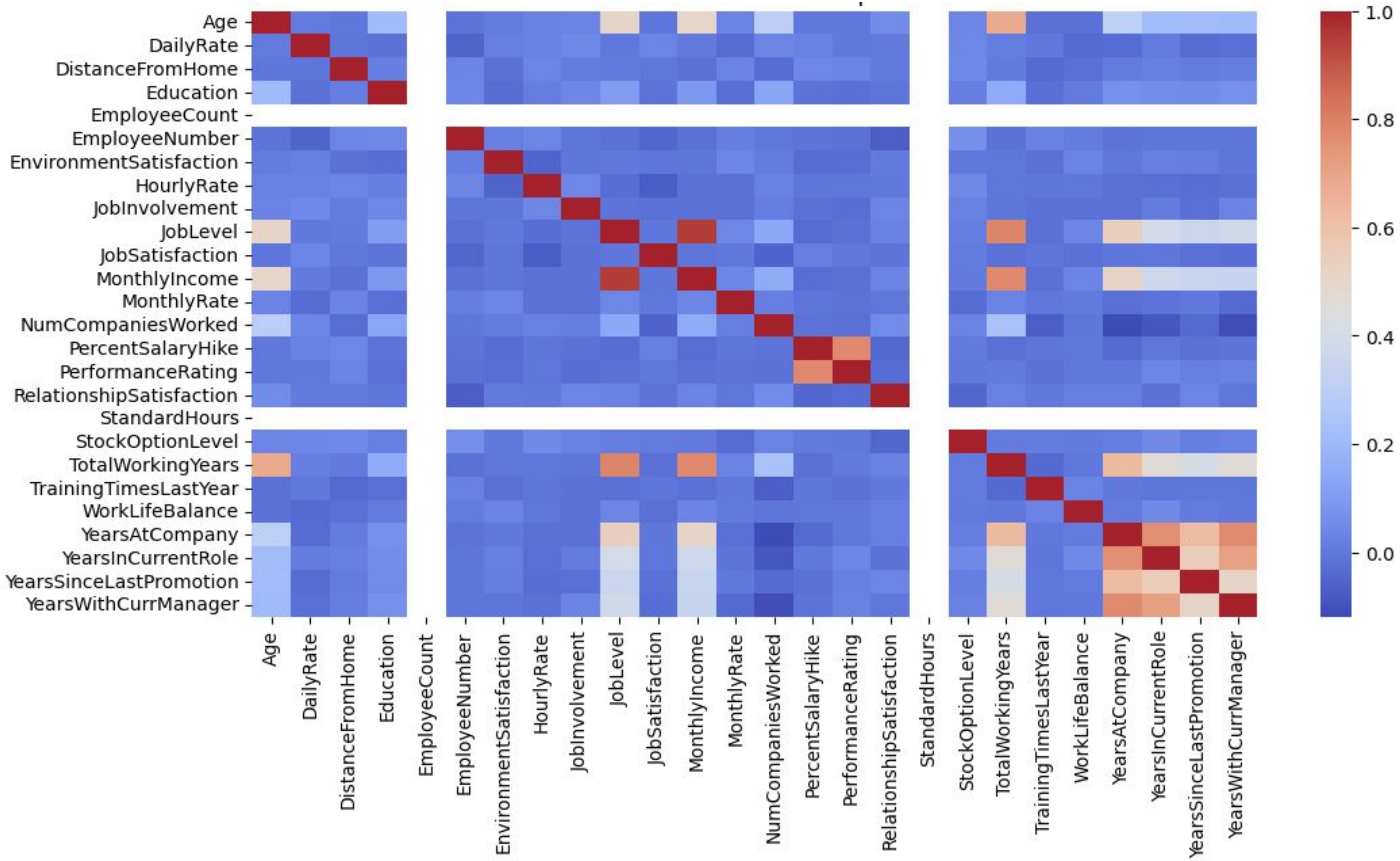
- Focuses on **selecting the top features** that have the strongest relationship with the **target variable (Attrition)**
- The **ANOVA F-test** is ideal for **classification problems** with **numerical input and categorical output**
- Helps reduce **dimensionality**, which:
    i. Improves **model performance**
    ii. Reduces **overfitting**
    iii. Enhances **interpretability**

Outcome:

- Top 10 most relevant features selected out of the entire dataset
- These features were used to train the models, ensuring **efficient learning** and **faster computation**

Conclusion:

- `SelectKBest` with `f_classif` was chosen for its **simplicity, speed**, and **relevance to supervised classification tasks**.


Age
DailyRate
DistanceFromHome
Education
EmployeeCount
EmployeeNumber
EnvironmentSatisfaction
HourlyRate
JobInvolvement
JobLevel
JobSatisfaction
MonthlyIncome
MonthlyRate
NumCompaniesWorked
PercentSalaryHike
PerformanceRating
RelationshipSatisfaction
StandardHours
StockOptionLevel
TotalWorkingYears
TrainingTimesLastYear
WorkLifeBalance
YearsAtCompany
YearsInCurrentRole
YearsSinceLastPromotion
YearsWithCurrManager
1.0
0.8
0.6
0.4
0.2
0.0

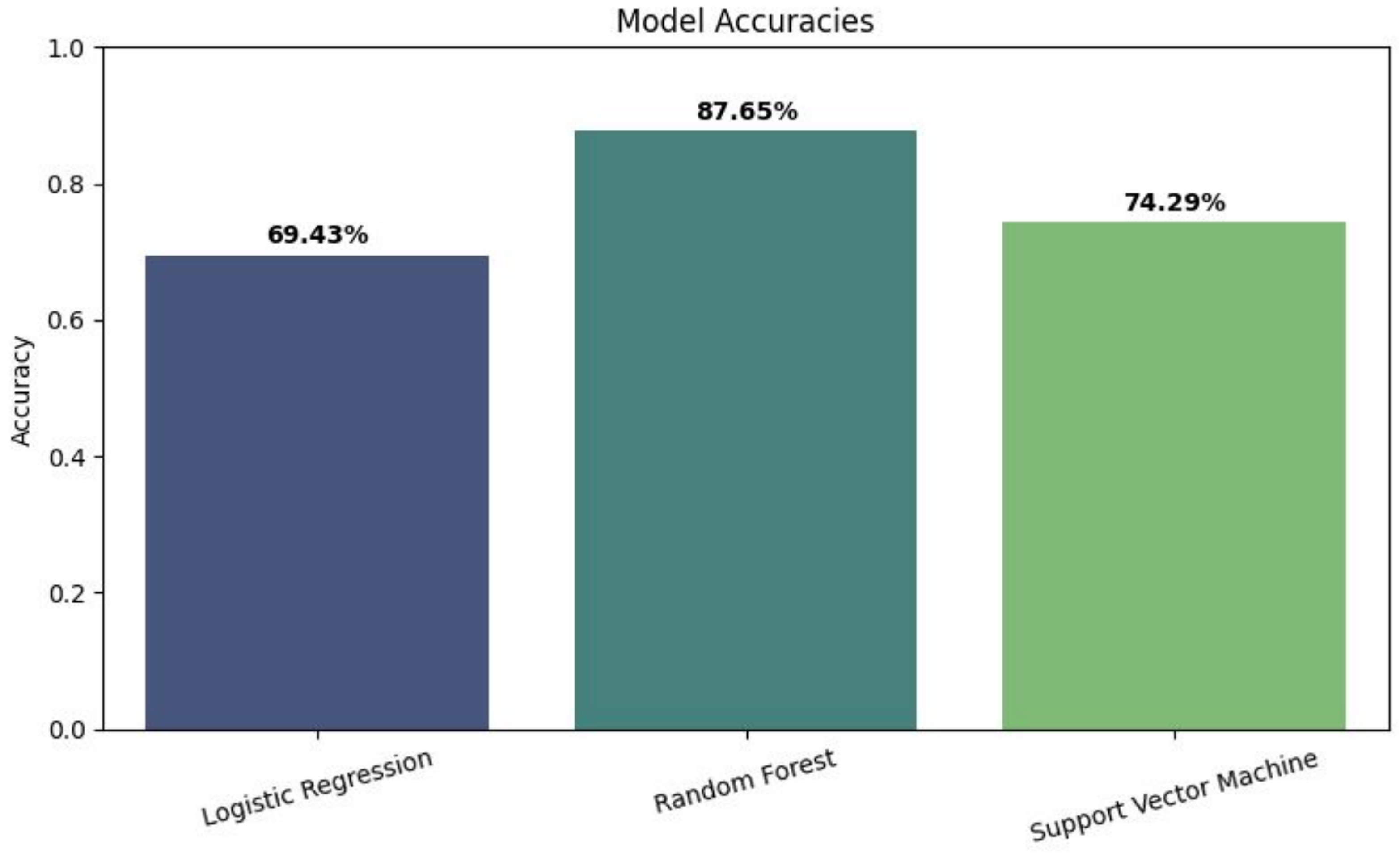
Modeling & Evaluation

Models Used:

- Logistic Regression: **0.69 accuracy**
- Random Forest: **0.88 accuracy**
- SVM: **0.74 accuracy**

Evaluation Method:

- **Accuracy Score** and **Classification Report**
- **Cross-validation (5-fold)** to ensure unbiased performance estimates
    i. Used **5-Fold CV** to evaluate Logistic Regression, Random Forest, and SVM
    ii. Calculated the **average accuracy** across all folds
    iii. Helped in selecting the model with **consistent performance**
- Best Model: **Random Forest** → balances performance and interpretability


Model Accuracies
1.0
0.8
0.6
0.4
0.2
0.0
Accuracy
69.43%
87.65%
74.29%
Logistic Regression
Random Forest
Support Vector Machine

Hyperparameter Tuning with GridSearchCV

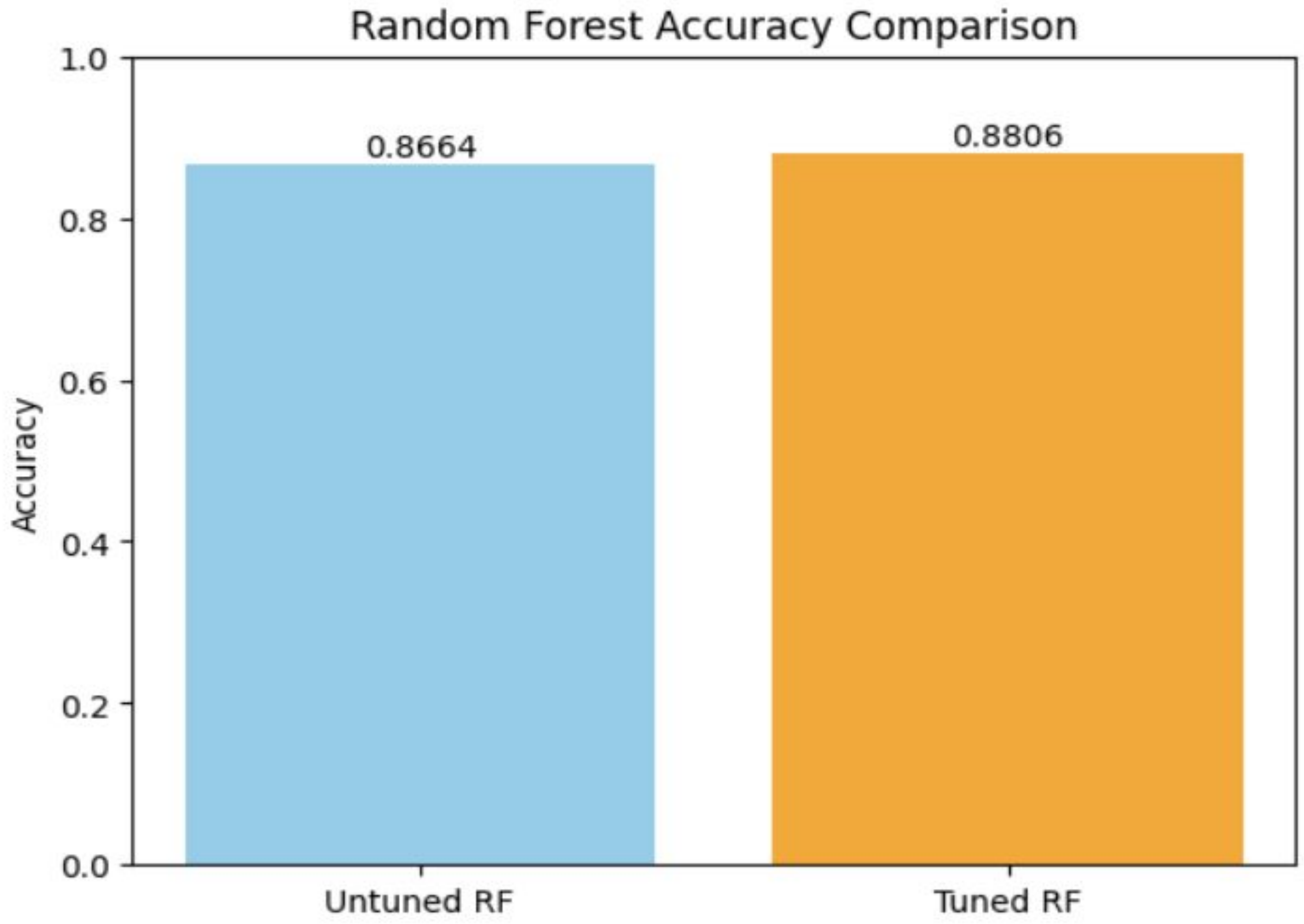
Objective: Enhance the performance of the Random Forest classifier through hyperparameter optimization using **GridSearchCV**.

GridSearchCV – performs an exhaustive search over specified parameter values using cross-validation.

Improved Accuracy: From **86.74%** (default) → **88.2%** (after tuning)

GridSearchCV helped to systematically test different parameter combinations and identify the optimal configuration, leading to improved model performance.

Parameter	Description	Values Tried
`n_estimators`	Number of trees in the forest	[50, 100, 150]
`max_depth`	Maximum depth of each tree	[None, 10, 20]
`min_samples_split`	Minimum samples required to split a node	[2, 5, 10]
`bootstrap`	Whether bootstrap samples are used	[True, False]


Random Forest Accuracy Comparison
0.8664
0.8806
Accuracy
1.0
0.8
0.6
0.4
0.2
0.0
Untuned RF
Tuned RF

SELECTED MODEL: RANDOM FOREST CLASSIFIER

Performance Metrics:

- Highest **Accuracy:** ~88% on test data
- Strong **Precision, Recall, and F1-Score** for both attrition classes
- Consistently high **cross-validation scores**

Why Random Forest?

- **Handles non-linearity** and interactions between features effectively
- Robust to **outliers** and **noise**
- Performs automatic **feature importance analysis**
- Works well even with **imbalanced data** (with SMOTE applied)

Why It Fits Our Problem:

- Employee attrition is influenced by multiple interacting factors (e.g., job satisfaction, environment, income)
- Random Forest's **ensemble approach** captures these relationships better than simpler models like logistic regression
- Offers high **predictive power**, crucial for actionable insights in HR analytics

Managerial Implications & Insights

Insights:

- Low job satisfaction = higher attrition risk
- Model can help HR proactively manage workforce

Actions for HR:

- Early identification of high-risk employees
- Employee engagement & retention programs
- Use model output for data-driven decisions

Novelty and Innovation

Real-World Focus: Tackles the business-critical issue of employee attrition using ML for proactive HR decisions.

Class Imbalance Solved with SMOTE: Balanced dataset using synthetic oversampling to improve minority class prediction.

Comparative Modeling: Evaluated Logistic Regression, Random Forest, and SVM to identify the best-performing model.

Statistical Feature Selection: Used SelectKBest with ANOVA F-value to reduce noise and improve model performance.

Visual Insights for Stakeholders: EDA with meaningful plots like Attrition vs. Job Satisfaction aids managerial understanding.

Reliable Evaluation: Applied K-Fold Cross-Validation for unbiased performance metrics

Managerial Recommendations: Model insights guide HR teams on retention strategies based on top influencing factors.

REFERENCES

IBM Dataset from Kaggle- IBM HR Analytics Employee Attrition & Performance | Kaggle

Scikit-learn documentation- https://scikit-learn.org/stable/

Pandas documentation- https://pandas.pydata.org/docs/

Medium articles on SMOTE and model evaluation-

Tackling Imbalanced Datasets with SMOTE (Synthetic Minority Over-sampling Technique) | by Husein Ghadiali | Medium

THANK YOU !!

SUBMITTED BY :- DEVIKA JAIN (01101192023)

SUBMITTED TO :- PROF DEBENDRA DHIR