Table of Contents

Week 4 Project: Spam Email Detection

1. Introduction

Employee attrition refers to the gradual reduction in workforce due to employees leaving the organization. Predicting attrition is important for HR departments to identify key risk factors and implement retention strategies. The goal of this project is to build machine learning models that classify whether an employee will leave (Yes) or stay (No).

2. Dataset

The dataset used in this project is the HR Employee Attrition dataset. The target variable is Attrition with two classes: Yes (237 samples) and No (1233 samples). Since the dataset is imbalanced, resampling techniques were applied to ensure balanced training data.

3. Data Preprocessing

- **Resampling:** The dataset was imbalanced, so the minority class (Yes) was oversampled using bootstrapping (resample) to balance both classes (1233 each).
- **Label Encoding:** Label Encoding was applied to categorical features.
- **Scaling:** Features were normalized using StandardScaler.
- **Splitting:** Data was split into training (70%) and testing (30%) sets.

4. Models Implemented

4.1 Decision Tree

A Decision Tree with criterion=gini and max_depth=3 was trained.
Accuracy: ~71%
f1-scores indicated weaker recall for Attrition=Yes.

4.2 Bagging Classifier

A Bagging classifier with Decision Trees as base learners (n_estimators=50).
Accuracy: ~95%

4.3 Random Forest + Grid Search

- Random Forest classifier implemented (n_estimators=50 initially).
- **Accuracy:** ~97%

- GridSearchCV was applied for hyperparameter tuning with parameters:
    - n_estimators: [50, 100, 150]
    - max_depth: [None, 5, 10]
    - min_samples_split: [2, 5, 10]
- Best hyperparameters: {'max_depth': None, 'min_samples_split': 5, 'n_estimators': 150}
- Best F1 Score: ~0.95

4.4 k-Nearest Neighbors (k-NN)

KNN classifier with n_neighbors=9 and metric=euclidean.
Accuracy: ~74%

5. Evaluation Metrics

The models were evaluated using Accuracy, Precision, Recall, and F1-score.

- Decision Tree: Accuracy ~71%; weaker recall for Attrition = Yes.
- Bagging: Accuracy ~95%; strong improvement over baseline.
- Random Forest: Accuracy ~97%; best F1-score (~0.95) after hyperparameter tuning with GridSearchCV.
- KNN: Accuracy ~74%; moderate performance, less effective than ensemble methods.

Class balancing through resampling ensured fair evaluation across both classes.

6. Results

- The baseline Decision Tree performed moderately, highlighting the challenges of predicting attrition directly.
- Bagging improved predictive stability, while Random Forest provided the strongest performance (~97% accuracy, ~0.95 F1-score) after tuning.
- Grid Search optimization further improved the Random Forest's generalization ability.
- KNN achieved moderate results but was less effective compared to ensemble methods.
- Resampling played a critical role in balancing the dataset, preventing bias toward the majority "No" class and enabling better detection of attrition cases ("Yes").

7. Conclusion

Employee attrition prediction is a valuable task for organizations. This project demonstrated that ensemble models, particularly Random Forest, are highly effective for this problem. Future improvements could include boosting algorithms (XGBoost, AdaBoost), advanced feature selection, and incorporating domain-specific HR insights.