Question 1: Dataset Overview & Problem Statement

Dataset Overview: The dataset is from IBM HR Analytics and focuses on employee attrition prediction. It contains information about employees, including demographics, job satisfaction, compensation and workplace factors, to determine which factors contribute to employees leaving a company.

Problem Statement: The goal is to build a classification model to predict whether an employee will leave the company (Attrition: Yes/No) based on key job-related and personal factors. This will help organizations take proactive steps to reduce employee turnover and improve retention strategies.

Dependent Variable (Target): Attrition (Yes = 1, No = 0) → Indicates whether an employee has left the company.

Independent Variables (Key Features Selected)

Feature	Meaning & Significance
OverTime	Employees working overtime are more likely to leave due to burnout.
MonthlyIncome	Salary dissatisfaction can drive employees to quit.
YearsAtCompany	New employees tend to leave more frequently, while long-term employees are more stable.
YearsInCurrentRole	Staying in the same role for too long may push employees to look for other opportunities
YearsWithCurrManager	A good relationship with the manager reduces attrition risk.
TotalWorkingYears	Experienced employees have more external job opportunities.
JobSatisfaction	Employees dissatisfied with their jobs are more likely to leave.
EnvironmentSatisfaction	Poor workplace conditions increase attrition risk.
DistanceFromHome	Employees with long commutes may seek jobs closer to home.
WorkLifeBalance	Employees with poor work-life balance are more likely to quit.

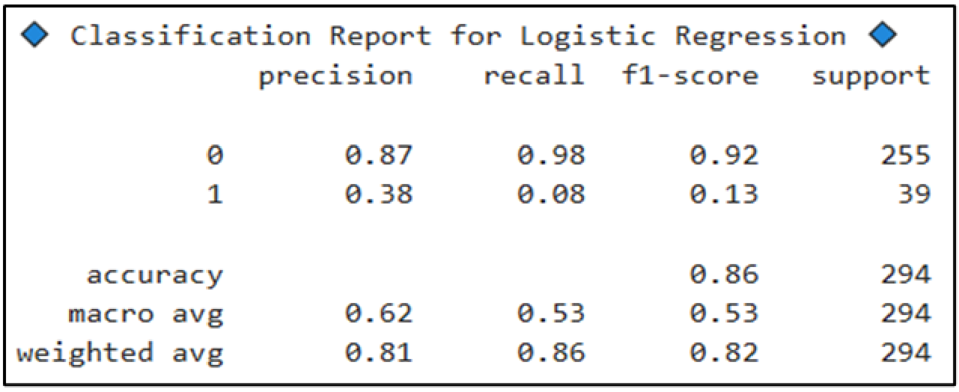
These selected features are based on statistical analysis (p-values, Chi-Square, MI Score, Random Forest, Lasso Regression) to ensure the model captures the most significant factors affecting attrition.

Question 2: Dataset Cleaning and Classification Model Evaluation

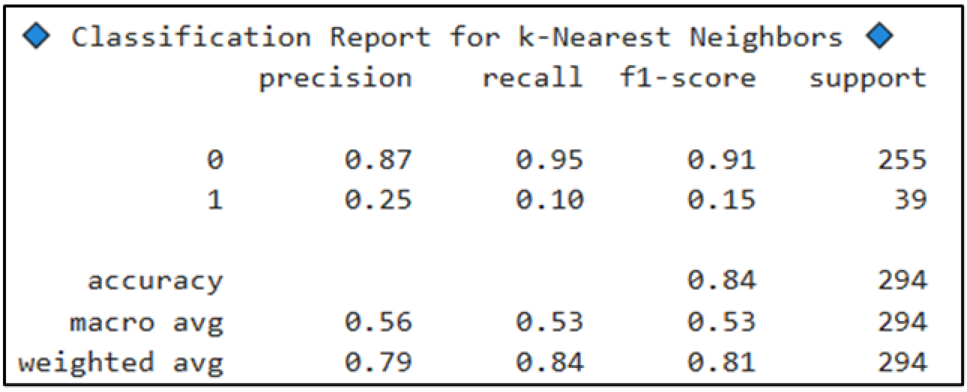
Classification Reports & Performance Comparison

- The dataset was cleaned by handling missing values and encoding categorical variables. Feature selection was applied to retain key attributes.
- The data was then split into 80% training and 20% testing sets. Each model was trained and evaluated to compare precision, recall, F1-score, and accuracy, ensuring a fair assessment.

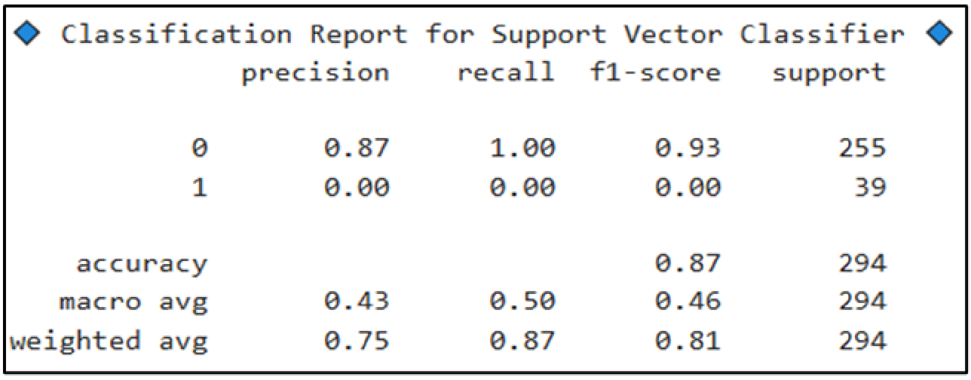
Below are the classification reports for the three models



```
🔷 Classification Report for Logistic Regression 🔷
              precision    recall  f1-score   support

           0       0.87      0.98      0.92       255
           1       0.38      0.08      0.13        39

    accuracy                           0.86       294
   macro avg       0.62      0.53      0.53       294
weighted avg       0.81      0.86      0.82       294
```



```
🔷 Classification Report for k-Nearest Neighbors 🔷
              precision    recall  f1-score   support

           0       0.87      0.95      0.91       255
           1       0.25      0.10      0.15        39

    accuracy                           0.84       294
   macro avg       0.56      0.53      0.53       294
weighted avg       0.79      0.84      0.81       294
```



```
🔷 Classification Report for Support Vector Classifier 🔷
              precision    recall  f1-score   support

           0       0.87      1.00      0.93       255
           1       0.00      0.00      0.00        39

    accuracy                           0.87       294
   macro avg       0.43      0.50      0.46       294
weighted avg       0.75      0.87      0.81       294
```

Question 3: Model Selection for Deployment

To decide which model to deploy, we consider:

- Overall Accuracy: How well the model classifies employees.
- Recall for Attrition Class (1): Since we are predicting employee attrition, it's crucial to minimize false negatives (employees likely to leave but predicted to stay).
- Business Impact: Identifying attrition early helps HR reduce hiring and training costs.

Best Model for Deployment: Logistic Regression

- Despite its low recall (8%), Logistic Regression offers the best trade-off between accuracy and interpretability.
- The HR team can understand why certain employees are at risk, allowing them to take proactive retention measures.
- kNN and SVC either perform poorly on recall or completely fail to predict attrition (SVC's recall = 0%).