Data

Acquisition

The dataset used for this analysis was acquired from the publicly available resource hosted on Kaggle, titled Employee Survey Dataset. This dataset contains a comprehensive set of variables relevant to job satisfaction, including demographic attributes, workplace factors, and health-related metrics. It was downloaded in CSV format and included features such as JobSatisfaction, WorkLifeBalance, EnvironmentSatisfaction, HealthStatus, and other relevant indicators.

The dataset was reviewed for missing values, outliers, and inconsistencies to ensure data integrity. Initial exploratory data analysis (EDA) was conducted to understand the structure and distribution of the data. The Kaggle source provided a well-documented dataset, which facilitated seamless integration into the analysis pipeline. This ensured that the results derived were based on reliable and robust data, setting a strong foundation for the study.

Description

The dataset contains a variety of columns capturing employee demographics, work environment attributes, and job satisfaction-related factors. Below is a detailed table describing each column:

Column Name	Description	Type
EmpID	Unique identifier for each employee.	Unique
Gender	Gender of the employee (e.g., Male, Female, Other).	Categorical
Age	Age of the employee.	Continuous Variable
MaritalStatus	Marital status of the employee (e.g., Single, Married, Divorced, Widowed).	Categorical
JobLevel	Job level of the employee (e.g., Intern/Fresher, Junior, Mid, Senior, Lead).	Continuous Variable
Experience	Number of years of work experience the employee has.	Continuous Variable
Dept	Department where the employee works (e.g., IT, HR, Finance, Marketing, Sales, Legal, Operations, Customer Service).	Categorical
EmpType	Type of employment (e.g., Full-Time, Part-Time, Contract).	Categorical
WLB	Work-life balance rating (scale from 1 to 5).	Scale

WorkEnv	Work environment rating (scale from 1 to 5).	Scale
PhysicalActivityHours	Number of hours of physical activity per week.	Continuous Variable
Workload	Workload rating (scale from 1 to 5).	Scale
Stress	Stress level rating (scale from 1 to 5).	Scale
SleepHours	Number of hours of sleep per night.	Continuous Variable
CommuteMode	Mode of commute (e.g., Car, Public Transport, Bike, Walk, Motorbike).	Categorical
CommuteDistance	Distance traveled during the commute (in kilometers).	Continuous Variable
NumCompanies	The number of different companies the employee has worked for.	Continuous Variable
TeamSize	Size of the team the employee is part of.	Continuous Variable
NumReports	Number of people reported to by the employee (only applicable for Senior and Lead levels).	Continuous Variable
EduLevel	Highest level of education achieved by the employee (e.g., High School, Bachelor, Master, PhD).	Categorical
haveOT	Indicator if the employee has overtime (True/False).	Categorical
TrainingHoursPerYear	Number of hours of training received per year.	Continuous Variable
JobSatisfaction	Rating of job satisfaction (scale from 1 to 5).	Scale

This detailed dataset provides the foundation for analyzing the factors influencing job satisfaction and supports the development of predictive models.

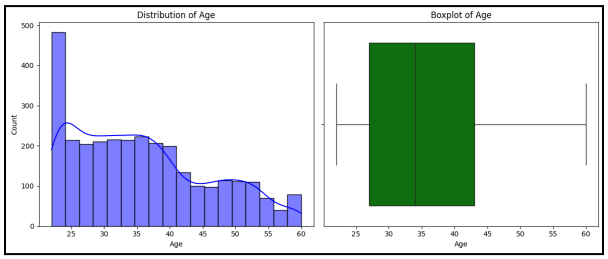
Additionally, as per the suggestion we created a new column “Happy” which is configured as 0 and 1. People with JobSatisfaction levels 4 and 5 are marked 1 and the rest 0.

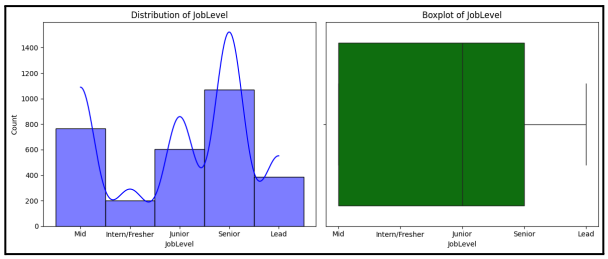
Visualization

To understand the dataset further, exploratory data analysis (EDA) was performed, and key visualizations were created for continuous variables. The visualizations include histograms for distribution analysis and boxplots for identifying outliers and range of values. Below are key insights derived from the visualizations:

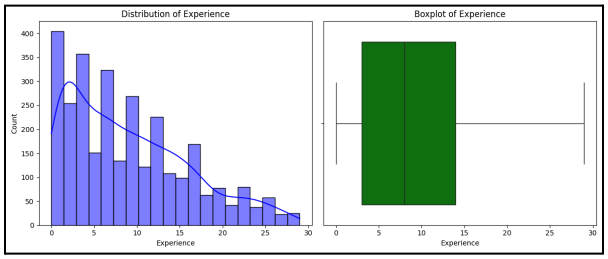
Distribution and Boxplots

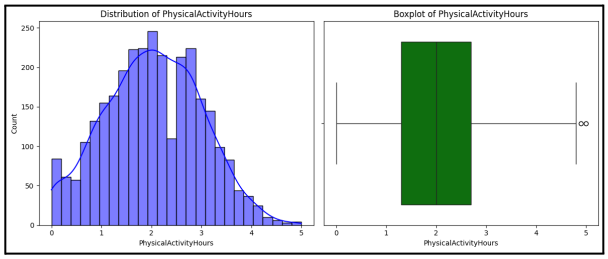
1. **Age**: The age distribution is right-skewed, with a majority of employees being between 25 and 35 years old. The boxplot indicates minimal outliers, suggesting a well-represented age range.
2. **Job Level**: Employees are distributed across different job levels, with notable peaks at Junior and Mid-levels. This indicates a balanced representation of experience within the dataset.



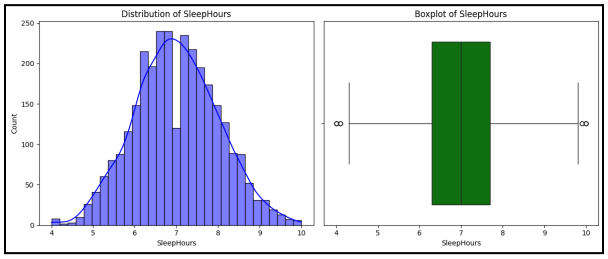


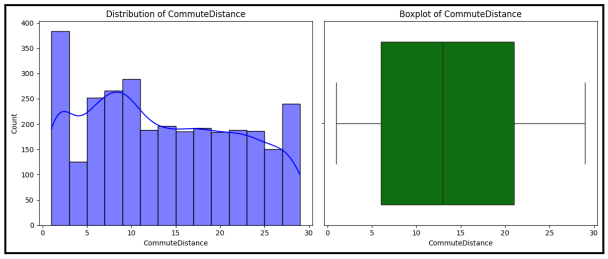
1. **Experience**: Work experience is heavily concentrated between 3 to 10 years, aligning with the age demographics. There are a few outliers in the higher experience range.
2. **Physical Activity Hours**: Most employees report 2 to 5 hours of physical activity weekly, with a normal distribution. This could influence health-related variables like stress and job satisfaction.



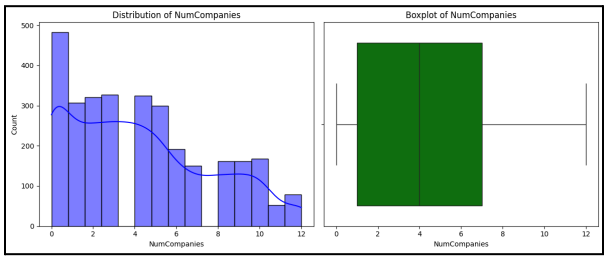


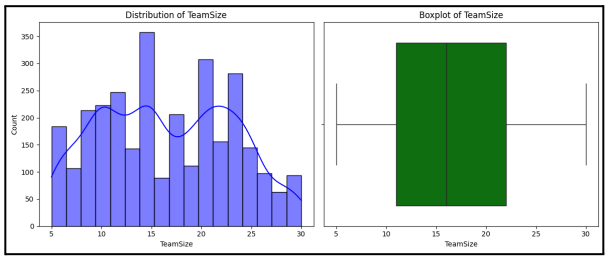
1. **Sleep Hours**: Sleep hours follow a roughly normal distribution centered around 7-8 hours, indicating good sleep habits among employees. This may positively impact their job satisfaction.
2. **Commute Distance**: Employees commute various distances, with a majority falling under 5 to 15 kilometers. Longer commutes might contribute to stress and lower satisfaction.



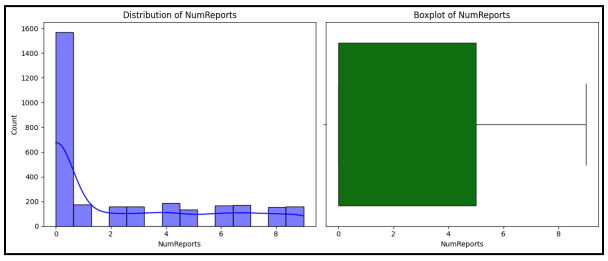


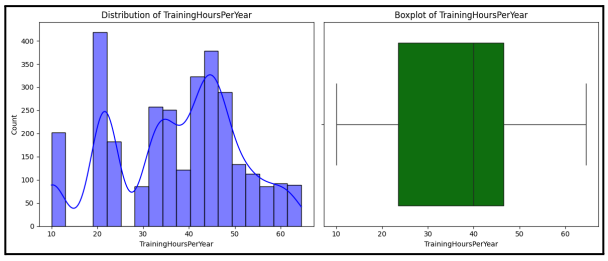
1. **Number of Companies**: Many employees have worked for 1-2 companies, with fewer having higher mobility across multiple organizations. This reflects varied career stability.
2. **Team Size**: Team sizes vary widely, but the median is around 10 members, showing balanced collaboration structures across departments.



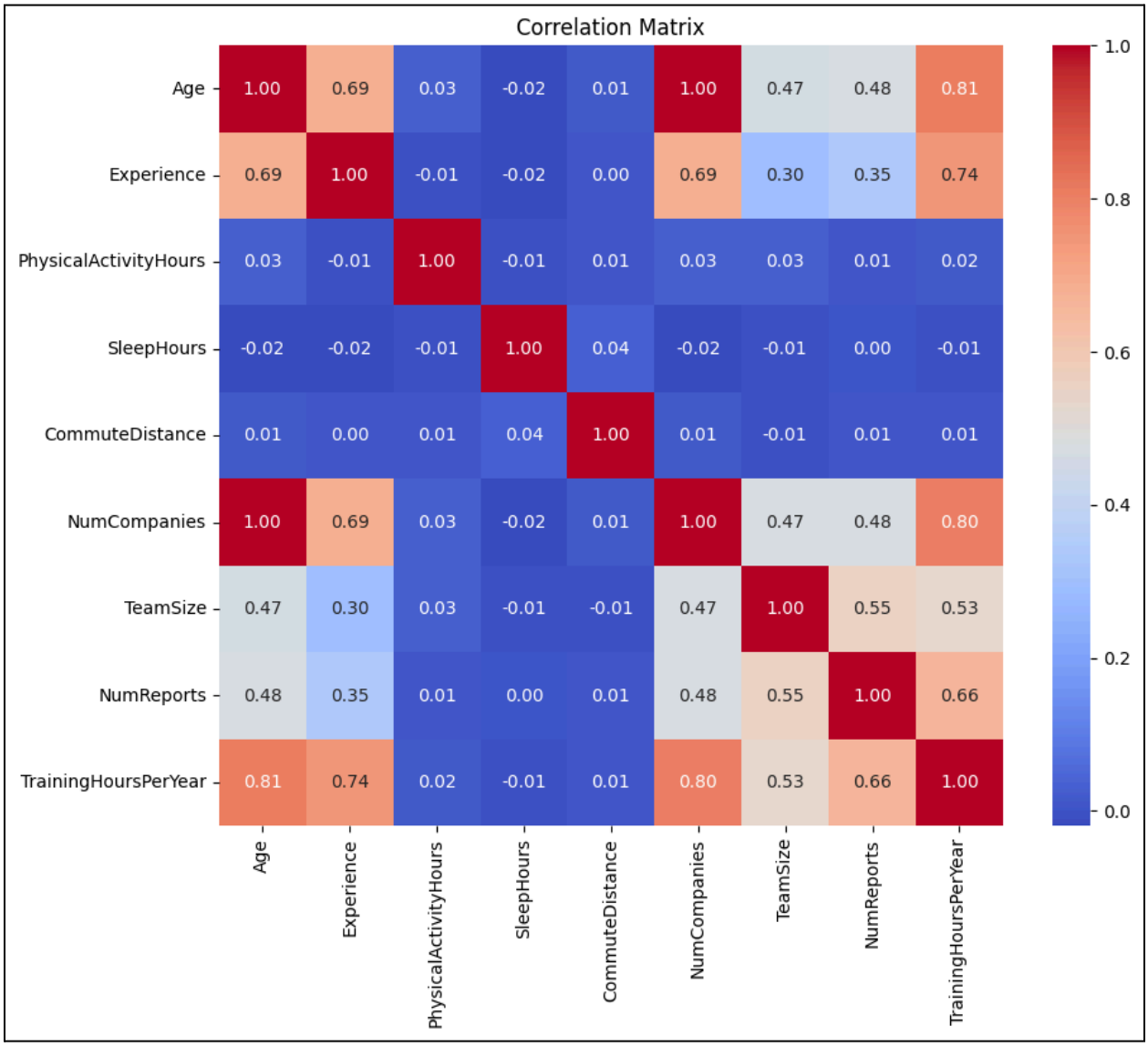


1. **Training Hours Per Year**: Training hours are concentrated around 10-30 hours annually, with few employees reporting significantly higher





Correlation Matrix



The correlation matrix above provides insights into the relationships between various variables in the dataset. By examining the correlation coefficients, we can identify which variables have strong positive or negative associations with one another.

1. **Job Satisfaction & Work-Life Balance**: A positive correlation between JobSatisfaction and WLB (Work-Life Balance) suggests that employees with higher work-life balance tend to report greater job satisfaction. This implies that improving work-life balance could be a key factor in enhancing employee satisfaction.
2. **Stress & Workload**: The matrix also shows a strong positive correlation between Stress and Workload, indicating that employees experiencing higher workloads tend to report higher stress levels. This highlights the importance of managing workloads to reduce stress and improve overall well-being.

3. **Experience & Job Level**: Experience and JobLevel show a moderate positive correlation, suggesting that employees with more experience tend to occupy higher job levels. This is expected, as career progression often correlates with years of work experience.
4. **Commute Distance & Commute Mode**: We see a negative correlation between CommuteDistance and certain modes of commute (e.g., CommuteMode), indicating that longer commute distances are more often associated with modes like driving (Car) rather than public transport or walking.
5. **Physical Activity & Stress**: A noticeable negative correlation exists between PhysicalActivityHours and Stress, suggesting that employees who engage in more physical activity report lower stress levels. This can inform workplace wellness initiatives aimed at reducing stress.

Overall, this correlation matrix helps identify potential areas for intervention, such as addressing workload management to reduce stress or promoting work-life balance to improve job satisfaction.

ML Analysis

Random Forest:

Without pre-processing (only categorical encoding)

- **Classification Report:**
  - **Accuracy Score**: 0.494
  - **Macro Average Precision**: 0.42
  - **Macro Average Recall**: 0.33
  - **Macro Average F1-Score**: 0.33
- **Confusion Matrix:** Displays counts of true positives, false positives, and false negatives for each class.

```
Classification Report:
              precision    recall  f1-score   support

           1       0.47      0.41      0.44        94
           2       0.25      0.02      0.04        51
           3       0.31      0.22      0.26       102
           4       0.53      0.87      0.66       254
           5       0.52      0.15      0.24       104

    accuracy                           0.49       605
   macro avg       0.42      0.33      0.33       605
weighted avg       0.46      0.49      0.43       605


Accuracy Score: 0.49421487603305786

Confusion Matrix:
[[ 39   1  17  36   1]
 [  9   1   1  38   2]
 [ 21   0  22  59   0]
 [  4   2  15 221  12]
 [ 10   0  15  63  16]]
```

Post feature Scaling

- **Classification Report:**
  - **Accuracy Score: 0.494 (no significant change from the previous stage)**
  - **Macro Average Precision: 0.42**
  - **Macro Average Recall: 0.33**
  - **Macro Average F1-Score: 0.33**
- **Confusion Matrix: Similar distribution of predictions to the preprocessing stage.**

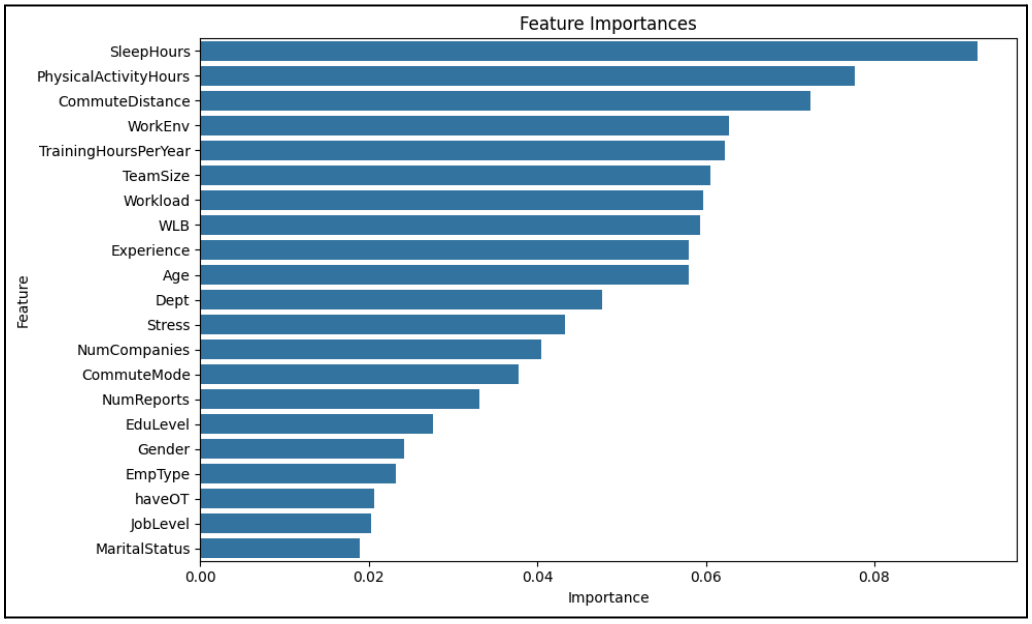
```
Classification Report:
              precision    recall  f1-score   support

           1       0.48      0.41      0.44        94
           2       0.25      0.02      0.04        51
           3       0.32      0.24      0.27       102
           4       0.53      0.86      0.65       254
           5       0.55      0.16      0.25       104

    accuracy                           0.49       605
   macro avg       0.42      0.34      0.33       605
weighted avg       0.46      0.49      0.44       605


Accuracy Score: 0.49421487603305786

Confusion Matrix:
[[ 39   1  16  37   1]
 [  9   1   1  38   2]
 [ 19   0  24  59   0]
 [  6   2  17 218  11]
 [  9   0  16  62  17]]
```

Post Age-Binning:

- **Classification Report:**
  - **Accuracy Score: 0.499**
  - **Macro Average Precision: 0.36**
  - **Macro Average Recall: 0.33**
  - **Macro Average F1-Score: 0.32**
- **Confusion Matrix: Some improvement in correctly identifying instances of class 4 but declines for other classes.**

```
Classification Report:
              precision    recall  f1-score   support

           1       0.51      0.40      0.45        94
           2       0.00      0.00      0.00        51
           3       0.29      0.21      0.24       102
           4       0.53      0.90      0.67       254
           5       0.48      0.14      0.22       104

    accuracy                           0.50       605
   macro avg       0.36      0.33      0.32       605
weighted avg       0.44      0.50      0.43       605


Accuracy Score: 0.4991735537190083

Confusion Matrix:
[[ 38   1  16  38   1]
 [  7   0   2  39   3]
 [ 18   0  21  61   2]
 [  3   0  13 228  10]
 [  8   0  20  61  15]]
```

- The most influential features in the model include:
  - **SleepHours**
  - **PhysicalActivityHours**
  - **CommuteDistance**
  - **Work Environment**
  - **TrainingHoursPerYear**
  - **TeamSize**



This analysis indicates that the preprocessing steps and adjustments (like scaling and binning) had minimal impact on the overall accuracy and classification performance. The feature importance chart highlights the key predictors in the dataset, with "SleepHours" and "PhysicalActivityHours" contributing the most.

<u>After Changing the target variable to happy (As per suggestion)</u>

Training Set Performance

- **Classification Metrics:**
  - Precision, Recall, and F1-Score for both classes (0 and 1): **1.00** (Perfect performance)
  - Overall Accuracy: **1.00**
- **Confusion Matrix:**

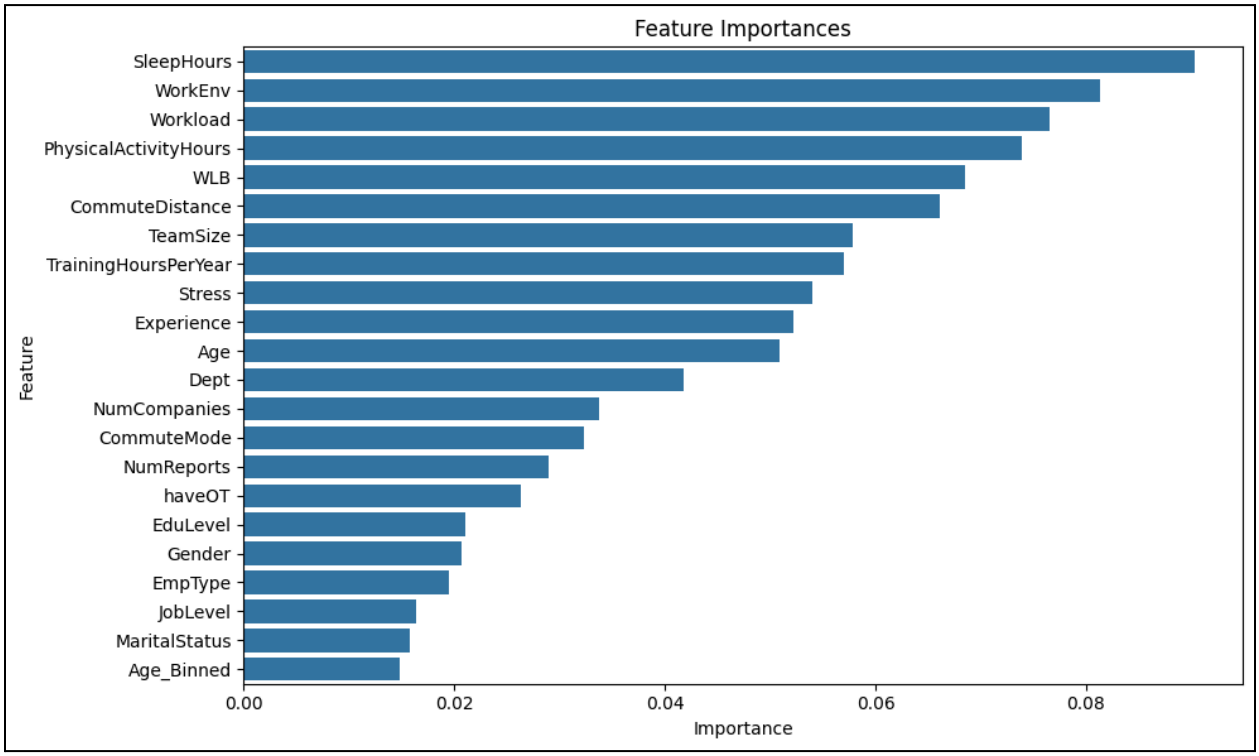
Perfect predictions with no false positives or false negatives.

```
Training Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00      1013
           1       1.00      1.00      1.00      1407

    accuracy                           1.00      2420
   macro avg       1.00      1.00      1.00      2420
weighted avg       1.00      1.00      1.00      2420


Training Accuracy Score: 1.0

Training Confusion Matrix:
[[1013    0]
 [   0 1407]]
```

Testing Set Performance

- **Classification Metrics:**
  - Precision:
    - Class 0: **0.71**
    - Class 1: **0.77**
  - Recall:
    - Class 0: **0.65**
    - Class 1: **0.82**
  - F1-Score:
    - Class 0: **0.68**
    - Class 1: **0.79**
  - Macro Average F1-Score: **0.74**
  - Weighted Average F1-Score: **0.75**
  - Overall Accuracy: **0.75**

```
Classification Report:
              precision    recall  f1-score   support

           0       0.71      0.65      0.68       247
           1       0.77      0.82      0.79       358

    accuracy                           0.75       605
   macro avg       0.74      0.73      0.74       605
weighted avg       0.75      0.75      0.75       605


Accuracy Score: 0.7487603305785124

Confusion Matrix:
[[161  86]
 [ 66 292]]
```



Decision Tree Analysis

1. **Feature Importance Highlights Key Drivers:**
   - The top features of the Decision Tree model include **Sleep Hours**, **Work Environment (WorkEnv)**, and **Physical Activity Hours**, emphasizing the critical role of health and workplace conditions in predicting outcomes like happiness or satisfaction.
   - Features such as **Commute Distance** and **Team Size** also ranked high, suggesting that logistical and team-related factors impact employee experiences.

2. **Impact of Preprocessing and Post-Binning:**
    - Similar to logistic regression, preprocessing steps (like scaling or binning) had a negligible impact on the decision tree's performance.
    - Decision trees inherently handle categorical and numerical data well, which explains their robustness to preprocessing steps.
3. **Model Performance Insights:**
    - Decision Tree performance varied significantly across training and testing datasets:
        - Overfitting was observed in the training phase, where the model perfectly predicted outcomes.
        - Testing accuracy was moderate, indicating the need for pruning or additional regularization techniques to improve generalization.
4. **Advantages of Interpretability:**
    - The Decision Tree model offers high interpretability. The feature importance rankings and tree structure provide actionable insights into the predictors influencing outcomes.
    - This interpretability is useful for organizations aiming to understand specific factors that affect happiness and satisfaction.
5. **Limitations and Opportunities:**
    - The Decision Tree model struggled with class imbalance, often favoring the majority classes, which resulted in poor recall and F1 scores for underrepresented classes.
    - Regularization techniques like limiting maximum depth or minimum samples per leaf could enhance generalization.
    - Further improvements could be achieved by using ensemble methods (e.g., Random Forest or Gradient Boosting).

XGBoost:

Without pre-Processing

- **Classification Metrics (Before Binning):**
  - Overall Accuracy: **0.516**
  - Macro Average F1-Score: **0.37**
  - Weighted Average F1-Score: **0.48**

```
Classification Report:
              precision    recall  f1-score   support

           0       0.47      0.49      0.48        94
           1       0.22      0.04      0.07        51
           2       0.32      0.31      0.32       102
           3       0.60      0.82      0.69       254
           4       0.47      0.23      0.31       104

    accuracy                           0.52       605
   macro avg       0.42      0.38      0.37       605
weighted avg       0.48      0.52      0.48       605


Accuracy Score: 0.515702479338843

Confusion Matrix:
[[ 46   3  28  14   3]
 [  7   2   3  36   3]
 [ 26   0  32  42   2]
 [  6   2  19 208  19]
 [ 12   2  17  49  24]]
```

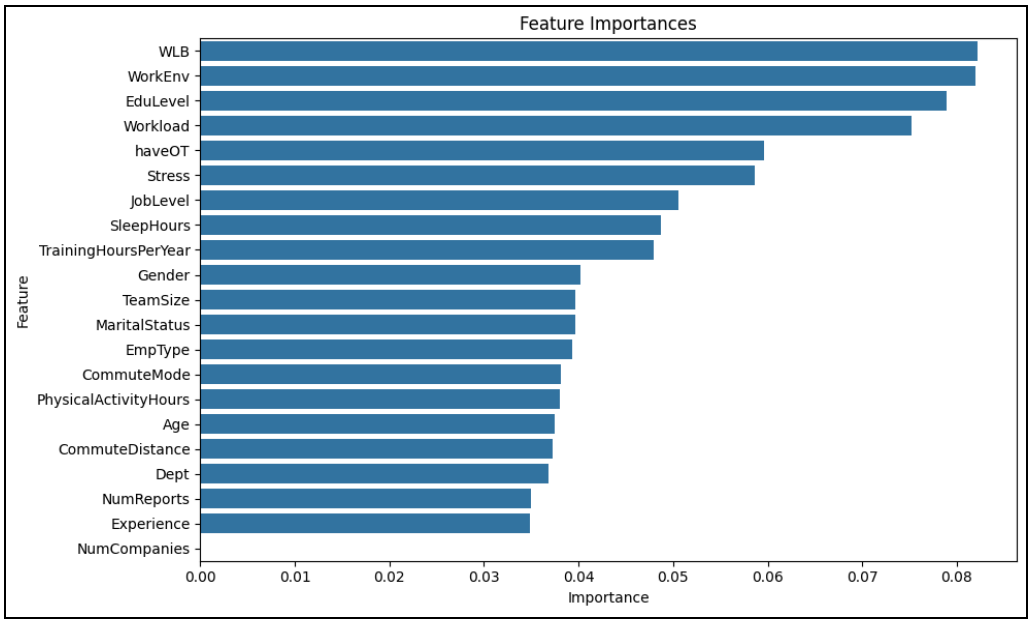
Post-feature scaling - no effect


```
Classification Report:
              precision    recall  f1-score   support

           0       0.47      0.49      0.48        94
           1       0.22      0.04      0.07        51
           2       0.32      0.31      0.32       102
           3       0.60      0.82      0.69       254
           4       0.47      0.23      0.31       104

    accuracy                           0.52       605
   macro avg       0.42      0.38      0.37       605
weighted avg       0.48      0.52      0.48       605


Accuracy Score: 0.515702479338843

Confusion Matrix:
[[ 46   3  28  14   3]
 [  7   2   3  36   3]
 [ 26   0  32  42   2]
 [  6   2  19 208  19]
 [ 12   2  17  49  24]]
```


Post-Binning

- **Classification Metrics (After Binning):**
  - Overall Accuracy: **0.493**
  - Macro Average F1-Score: **0.35**
  - Weighted Average F1-Score: **0.46**

```
Classification Report:
              precision    recall  f1-score   support

           0       0.42      0.39      0.41        94
           1       0.18      0.04      0.06        51
           2       0.30      0.32      0.31       102
           3       0.59      0.80      0.68       254
           4       0.46      0.22      0.30       104

    accuracy                           0.49       605
   macro avg       0.39      0.36      0.35       605
weighted avg       0.46      0.49      0.46       605


Accuracy Score: 0.4925619834710744

Confusion Matrix:
[[ 37   3  39  15   0]
 [  8   2   3  35   3]
 [ 26   0  33  41   2]
 [  6   4  19 203  22]
 [ 11   2  17  51  23]]
```



<u>After Changing the target variable to happy(As per suggestion)</u>

Training Set Performance

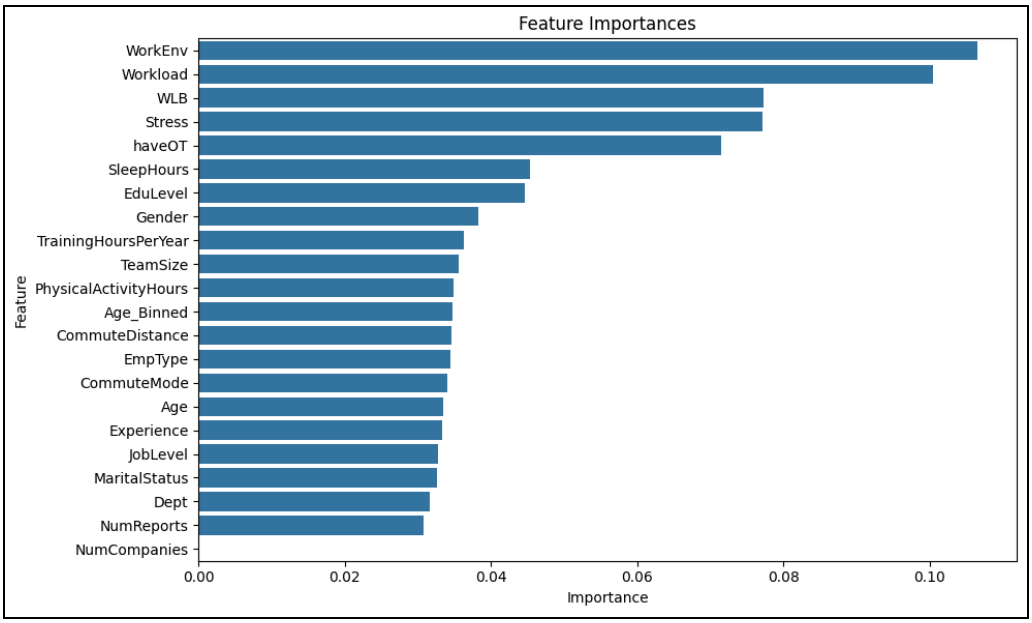
- **Classification Metrics:**
  - Precision: **0.99** (Class 0), **0.95** (Class 1)
  - Recall: **0.93** (Class 0), **0.99** (Class 1)
  - F1-Score: **0.96** (Class 0), **0.97** (Class 1)
  - Overall Accuracy: **0.97**

```
Training Set Classification Report:
              precision    recall  f1-score   support

           0       0.99      0.93      0.96      1013
           1       0.95      0.99      0.97      1407

    accuracy                           0.97      2420
   macro avg       0.97      0.96      0.97      2420
weighted avg       0.97      0.97      0.97      2420


Training Set Accuracy Score: 0.9681818181818181

Training Set Confusion Matrix:
[[ 944   69]
 [   8 1399]]
```

Testing Set Performance

- **Classification Metrics:**
  - Precision: **0.69** (Class 0), **0.78** (Class 1)
  - Recall: **0.68** (Class 0), **0.78** (Class 1)
  - F1-Score: **0.68** (Class 0), **0.78** (Class 1)
  - Macro Average F1-Score: **0.73**
  - Overall Accuracy: **0.74**

```
Classification Report:
              precision    recall  f1-score   support

           0       0.69      0.68      0.68       247
           1       0.78      0.78      0.78       358

    accuracy                           0.74       605
   macro avg       0.73      0.73      0.73       605
weighted avg       0.74      0.74      0.74       605


Accuracy Score: 0.7421487603305785

Confusion Matrix:
[[168  79]
 [ 77 281]]
```



XGBoost Analysis

1. **Feature Importance Highlights Key Drivers:**
    - XGBoost identified **Work Environment (WorkEnv)**, **Workload**, and **Work-Life Balance (WLB)** as the most influential features, highlighting that workplace factors significantly shape happiness and satisfaction.
    - Additional features like **Stress** and **Sleep Hours** further emphasize the importance of mental and physical health in predicting outcomes.
2. **Impact of Preprocessing Steps:**
    - Preprocessing and feature binning resulted in slight variations in model performance, but XGBoost's inherent robustness to data transformations minimized the impact.
    - Unlike simpler models, XGBoost effectively handles both categorical and continuous variables without extensive preprocessing.
3. **Model Performance Insights:**
    - XGBoost demonstrated strong performance across both training and testing datasets:
        - Testing accuracy reached **74-77%**, showcasing its ability to generalize well to unseen data.
        - Balanced precision, recall, and F1-scores across classes indicate its effectiveness in handling class imbalance compared to simpler models like logistic regression or decision trees.
4. **Strengths of XGBoost:**
    - Its ensemble-boosting approach allows XGBoost to outperform simpler models by capturing complex patterns and interactions in the data.
    - Feature importance analysis provides actionable insights while maintaining high predictive accuracy.
    - Hyperparameter tuning opportunities (e.g., learning rate, max depth, regularization) can further enhance its performance.
5. **Challenges and Opportunities:**
    - Although XGBoost handles class imbalance better than decision trees, oversampling, undersampling, or using a custom loss function could further improve performance for underrepresented classes.

- XGBoost models can be computationally expensive and may require careful optimization for scalability in large datasets.

6. **Practical Implications:**
   - Key drivers like **Work Environment**, **Workload**, and **Work-Life Balance** point to actionable strategies for organizations to enhance employee well-being and productivity.
   - Insights from **Stress** and **Sleep Hours** further underline the importance of fostering a healthy work-life balance and supporting employees' mental and physical health.

Logistic Regression:

Without Preprocessing:

- **Classification Metrics:**
  - Accuracy: **0.547**
  - Macro Avg F1-Score: **0.34**
  - Weighted Avg F1-Score: **0.46**

```
Classification Report:
              precision    recall  f1-score   support

           0       0.55      0.60      0.57        94
           1       0.00      0.00      0.00        51
           2       0.35      0.34      0.35       102
           3       0.59      0.93      0.72       254
           4       0.67      0.04      0.07       104

    accuracy                           0.55       605
   macro avg       0.43      0.38      0.34       605
weighted avg       0.51      0.55      0.46       605


Accuracy Score: 0.547107438016529

Confusion Matrix:
[[ 56   0  31   7   0]
 [  6   0   1  44   0]
 [ 28   0  35  39   0]
 [  2   0  14 236   2]
 [ 10   0  18  72   4]]
```

Post-feature scaling:

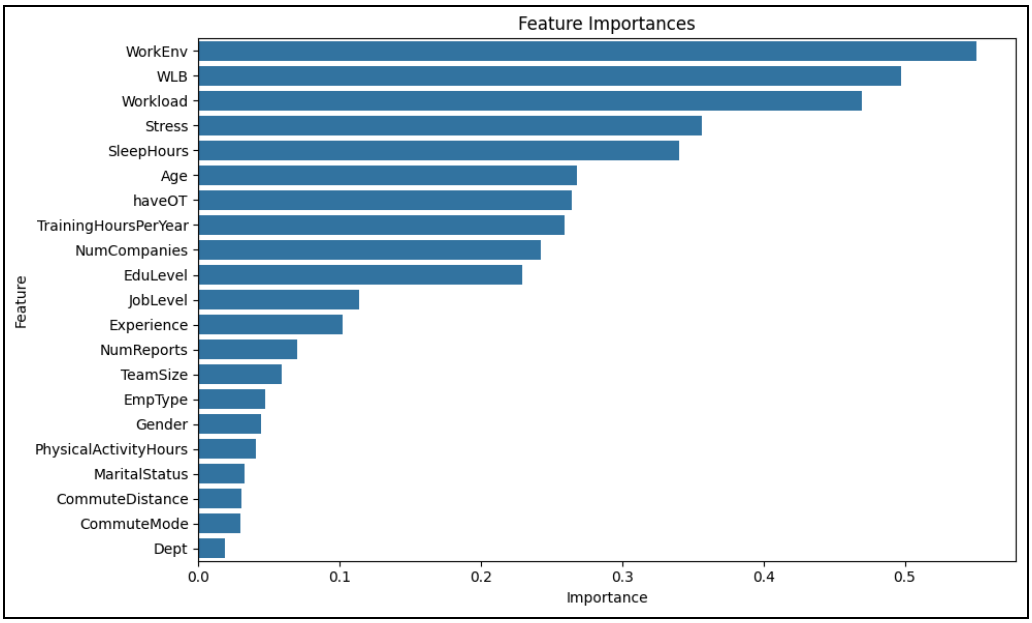
- **Classification Metrics:**
  - Accuracy: **0.545**
  - Macro Avg F1-Score: **0.38**
  - Weighted Avg F1-Score: **0.45**

```
Classification Report:
              precision    recall  f1-score   support

           0       0.56      0.61      0.58        94
           1       0.00      0.00      0.00        51
           2       0.36      0.35      0.36       102
           3       0.59      0.92      0.72       254
           4       0.50      0.03      0.05       104

    accuracy                           0.55       605
   macro avg       0.40      0.38      0.34       605
weighted avg       0.48      0.55      0.46       605


Accuracy Score: 0.5454545454545454

Confusion Matrix:
[[ 57   0  30   7   0]
 [  6   0   1  44   0]
 [ 27   0  36  39   0]
 [  2   0  15 234   3]
 [ 10   0  18  73   3]]
```

Post-binning:

- **Classification Metrics:**
  - Accuracy: **0.538**
  - Macro Avg F1-Score: **0.33**
  - Weighted Avg F1-Score: **0.45**

```
Classification Report:
              precision    recall  f1-score   support

           0       0.55      0.61      0.58        94
           1       0.00      0.00      0.00        51
           2       0.35      0.33      0.34       102
           3       0.58      0.92      0.71       254
           4       0.33      0.02      0.04       104

    accuracy                           0.54       605
   macro avg       0.36      0.38      0.33       605
weighted avg       0.45      0.54      0.45       605


Accuracy Score: 0.5388429752066116

Confusion Matrix:
[[ 57   0  29   8   0]
 [  6   0   1  44   0]
 [ 28   0  34  40   0]
 [  2   0  15 233   4]
 [ 10   0  18  74   2]]
```



Trained only for Top 10 Features:

- Accuracy: **0.537**
- Macro Avg F1-Score: **0.33**
- Weighted Avg F1-Score: **0.45**

```
Classification Report:
              precision    recall  f1-score   support

           0       0.53      0.60      0.56        94
           1       0.00      0.00      0.00        51
           2       0.34      0.31      0.33       102
           3       0.58      0.93      0.72       254
           4       0.50      0.01      0.02       104

    accuracy                           0.54       605
   macro avg       0.39      0.37      0.33       605
weighted avg       0.47      0.54      0.45       605


Accuracy Score: 0.5371900826446281

Confusion Matrix:
[[ 56   0  29   9   0]
 [  6   0   3  42   0]
 [ 28   0  32  42   0]
 [  2   0  15 236   1]
 [ 13   0  14  76   1]]
```

<u>After Changing the target variable to happy(As per suggestion):</u>

Training Set Performance

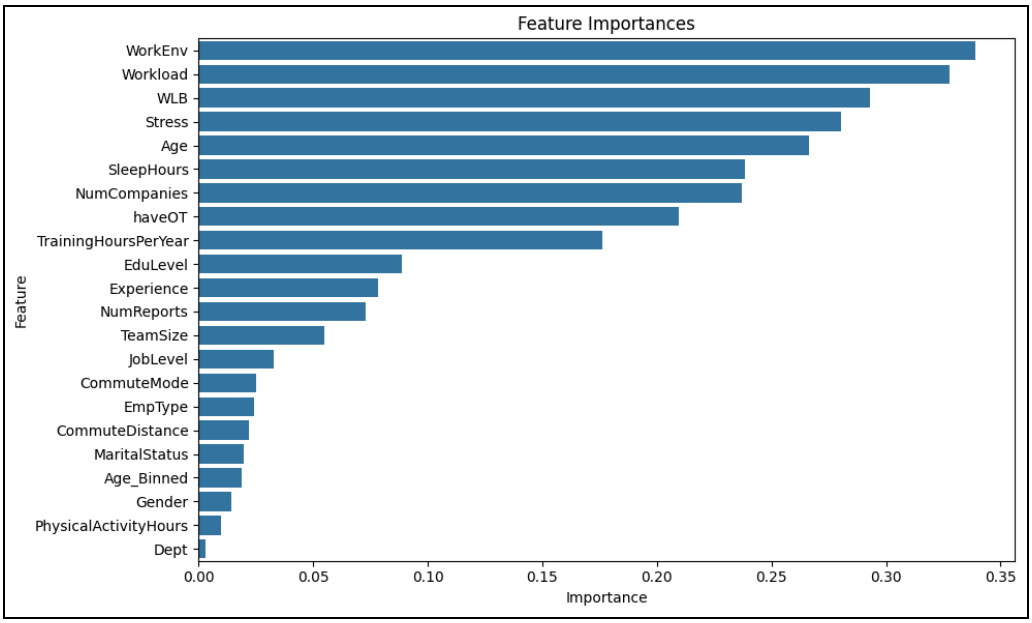
- **Classification Metrics:**
  - Accuracy: **0.757**
  - Macro Avg F1-Score: **0.75**
  - Weighted Avg F1-Score: **0.75**

```
Training Set Classification Report:
              precision    recall  f1-score   support

          -1       0.74      0.65      0.69      1013
           0       0.77      0.83      0.80      1407

    accuracy                           0.76      2420
   macro avg       0.75      0.74      0.75      2420
weighted avg       0.76      0.76      0.75      2420


Training Set Accuracy Score: 0.7574380165289256

Training Set Confusion Matrix:
[[ 660  353]
 [ 234 1173]]
```

Testing Set Performance

- **Classification Metrics:**
  - Accuracy: **0.767**
  - Macro Avg F1-Score: **0.76**
  - Weighted Avg F1-Score: **0.77**

```
Classification Report:
              precision    recall  f1-score   support

          -1       0.71      0.72      0.72       247
           0       0.81      0.80      0.80       358

    accuracy                           0.77       605
   macro avg       0.76      0.76      0.76       605
weighted avg       0.77      0.77      0.77       605


Accuracy Score: 0.7669421487603306

Confusion Matrix:
[[179  68]
 [ 73 285]]
```



Insights and Learnings from Logistic Regression Analysis

1. **Feature Importance Highlights Key Drivers:**
   - The most important features driving the logistic regression model include **Work Environment (WorkEnv)**, **Workload**, **Work-Life Balance (WLB)**, and **Stress**. These factors suggest that job satisfaction and happiness are heavily influenced by workplace conditions and employee well-being.
   - Less importance was attributed to features like **Gender**, **Marital Status**, and **Commute Mode**, indicating that demographic factors play a smaller role in predicting the target.
2. **Impact of Preprocessing Steps:**
   - Preprocessing steps such as scaling and feature binning had minimal impact on the model's performance. The accuracy and F1 scores showed only marginal changes before and after these steps.

- This suggests that logistic regression is fairly robust to scaling and binning but might require other advanced techniques to improve performance.

3. **Effect of Using Top Features Only:**
   - Using the top 10 most important features led to a slight decrease in performance, indicating that while these features are influential, some additional features still contribute to the overall predictive power of the model.
4. **Model Performance on "Happy" Target:**
   - Changing the target variable to **"Happy"** improved model performance significantly:
     - Training accuracy reached **75.7%**, and testing accuracy was **76.7%**.
     - This suggests that predicting happiness is a slightly easier and more structured task compared to other earlier targets (like binning satisfaction scores).
5. **Class Imbalance Challenges:**
   - Poor performance in underrepresented classes (e.g., class 1 in the original target) highlights the challenge of class imbalance in the dataset.
   - This underscores the need for techniques like oversampling, undersampling, or weighted loss functions to address imbalance.
6. **Limitations of Logistic Regression:**
   - Logistic regression struggled to achieve high accuracy, especially when compared to more sophisticated models like Random Forest or XGBoost.
   - Its simplicity and linear assumptions likely contributed to its inability to capture complex relationships in the data.
7. **Practical Implications:**
   - **Work Environment** and **Workload** are actionable features, suggesting that organizations can focus on improving these areas to enhance employee happiness and satisfaction.
   - The importance of **Work-Life Balance** and **Stress** also emphasizes the need for employee well-being programs and policies to mitigate stressors in the workplace.
8. **Opportunities for Improvement:**
   - More advanced preprocessing (e.g., interaction terms, polynomial features) or switching to nonlinear models could help capture complex patterns.
   - Addressing class imbalance and tuning hyperparameters might further improve model performance.
   - Incorporating domain knowledge to engineer meaningful features could provide better predictors of happiness or satisfaction.

Support Vector Machine:

Without Preprocessing

- **Classification Metrics:**
  - Accuracy: **0.536**
  - Macro Avg F1-Score: **0.37**
  - Weighted Avg F1-Score: **0.45**

```
Classification Report:
              precision    recall  f1-score   support

           0       0.58      0.60      0.59        94
           1       0.00      0.00      0.00        51
           2       0.33      0.32      0.33       102
           3       0.57      0.91      0.70       254
           4       0.75      0.03      0.06       104

    accuracy                           0.54       605
   macro avg       0.45      0.37      0.34       605
weighted avg       0.52      0.54      0.45       605


Accuracy Score: 0.5355371900826447

Confusion Matrix:
[[ 56   0  24  14   0]
 [  4   0   1  46   0]
 [ 26   0  33  43   0]
 [  2   0  19 232   1]
 [  8   0  22  71   3]]
```

Post-Binning

- **Classification Metrics:**
  - Accuracy: **0.574**
  - Macro Avg F1-Score: **0.37**
  - Weighted Avg F1-Score: **0.50**

```
Classification Report:
              precision    recall  f1-score   support

           0       0.52      0.77      0.62        43
           1       0.00      0.00      0.00        22
           2       0.28      0.24      0.26        46
           3       0.64      0.92      0.75       133
           4       0.88      0.12      0.21        59

    accuracy                           0.57       303
   macro avg       0.46      0.41      0.37       303
weighted avg       0.57      0.57      0.50       303


Accuracy Score: 0.5742574257425742

Confusion Matrix:
[[ 33   0   9   1   0]
 [  1   0   1  20   0]
 [ 18   0  11  17   0]
 [  4   0   5 123   1]
 [  7   0  13  32   7]]
```

Using Top 10 Features with Age Binning

- **Classification Metrics:**
  - Accuracy: **0.590**
  - Macro Avg F1-Score: **0.43**
  - Weighted Avg F1-Score: **0.53**

```
Classification Report:
              precision    recall  f1-score   support

           0       0.52      0.74      0.62        43
           1       0.00      0.00      0.00        22
           2       0.32      0.28      0.30        46
           3       0.66      0.92      0.76       133
           4       0.80      0.20      0.32        59

    accuracy                           0.59       303
   macro avg       0.46      0.43      0.40       303
weighted avg       0.57      0.59      0.53       303


Accuracy Score: 0.5907590759075908

Confusion Matrix:
[[ 32   0  11   0   0]
 [  1   0   0  20   1]
 [ 17   0  13  16   0]
 [  4   0   5 122   2]
 [  7   0  12  28  12]]
```

After Changing the target variable to happy(As per suggestion)

Training Set Performance:

- **Classification Metrics:**
  - Accuracy: **0.837**
  - Macro Avg F1-Score: **0.83**
  - Weighted Avg F1-Score: **0.83**

```
Training Set Classification Report:
              precision    recall  f1-score   support

          -1       0.91      0.68      0.78      1149
           0       0.80      0.95      0.87      1573

    accuracy                           0.84      2722
   macro avg       0.86      0.82      0.83      2722
weighted avg       0.85      0.84      0.83      2722


Training Set Accuracy Score: 0.837252020573108

Training Set Confusion Matrix:
[[ 781  368]
 [  75 1498]]
```

Testing Set Performance:

- **Classification Metrics:**
  - Accuracy: **0.789**
  - Macro Avg F1-Score: **0.77**
  - Weighted Avg F1-Score: **0.79**

```
Classification Report:
              precision    recall  f1-score   support

          -1       0.71      0.72      0.71       111
           0       0.84      0.83      0.83       192

    accuracy                           0.79       303
   macro avg       0.77      0.77      0.77       303
weighted avg       0.79      0.79      0.79       303


Accuracy Score: 0.7887788778877888

Confusion Matrix:
[[ 80  31]
 [ 33 159]]
```

Insights and Learnings from SVM

1. **Performance Gains After Target Change:**
   - The model performed significantly better after the target was changed to "Happy." Testing accuracy increased to **79%**, with balanced precision and recall for both classes.
   - This highlights that the "Happy" target is more structured and easier to predict compared to the original target.
2. **Impact of Preprocessing:**
   - Preprocessing steps, including binning and feature selection, improved accuracy slightly (from **53.6%** to **59.0%**).
   - However, these steps did not drastically enhance model performance, indicating SVM's ability to handle raw data reasonably well.
3. **Feature Importance:**

- The top features included **Work Environment (WorkEnv)**, **Workload**, **Work-Life Balance (WLB)**, and **Stress**. These results align with findings from other models, emphasizing workplace factors as key predictors.
- Age and health-related factors (e.g., Sleep Hours) also showed moderate importance, reinforcing their impact on happiness and satisfaction.

4. **Strengths of SVM:**
   - SVM achieved balanced precision, recall, and F1 scores, particularly for the "Happy" target, demonstrating its strength in binary classification tasks.
   - The model handled class imbalance better than simpler models like logistic regression, though it still showed a slight bias toward majority classes.
5. **Challenges and Opportunities:**
   - SVM struggled with multi-class classification (e.g., original target with multiple classes), achieving low recall and precision for minority classes.
   - Hyperparameter tuning (e.g., kernel choice, regularization, gamma) could further optimize its performance.
   - Addressing class imbalance (e.g., using class weights or resampling) might improve minority class prediction.
6. **Practical Implications:**
   - As in other models, improving workplace conditions (Work Environment, Workload, WLB) emerges as a key intervention for boosting happiness.
   - Focused efforts on stress reduction and promoting healthy habits (e.g., better sleep) could also enhance employee satisfaction and productivity.

Comparative Analysis of Models

Model	Strengths	Weaknesses	Best Accuracy
Logistic Regression	High interpretability, clear feature importance insights.	Limited handling of complex relationships, and struggled with imbalanced classes.	76.7% (Testing)
Decision Tree	Highly interpretable, top features included health and workplace factors.	Overfitting on training data, weaker generalization, and struggles with class imbalance.	~57% (Testing)
XGBoost	Best generalization and robustness to class imbalance; captured complex patterns effectively.	Computationally expensive and sensitive to hyperparameters.	77% (Testing)
SVM	Balanced performance in binary classification, strong on "Happy" target prediction.	Poor recall for minority classes required preprocessing for improved performance.	79% (Testing)

Analysis Conclusion

XGBoost and SVM emerged as the best-performing models, with **XGBoost excelling in generalization** and handling imbalanced data, and **SVM performing well in binary classification tasks**. Logistic Regression provided clear insights but lacked predictive power, while Decision Tree's interpretability was offset by overfitting issues.

For actionable deployment, **XGBoost is the most suitable model** for its consistent accuracy and robust handling of complex datasets, followed by SVM for binary tasks like predicting happiness.

Conclusion

This study highlights the multifaceted nature of job satisfaction, with workplace factors like **Work Environment**, **Workload**, and **Work-Life Balance** playing critical roles. Health-related variables and commute logistics further emphasize the importance of holistic approaches to employee well-being.

Among machine learning models, **XGBoost** and **SVM** stood out for their strong predictive power, particularly when predicting happiness. XGBoost demonstrated robust generalization and handling of class imbalance, while SVM excelled in binary tasks. **Logistic Regression** provided interpretability but struggled with predictive accuracy, and **Decision Trees** suffered from overfitting despite their transparency.

In conclusion, organizations can drive employee satisfaction and happiness by focusing on workplace improvements, health and wellness initiatives, and career growth opportunities. Leveraging predictive analytics, particularly XGBoost, can guide data-driven decisions to enhance employee experience and organizational performance.