**Income Classification Report**

**1. Introduction**

**Problem Description**

The goal of this analysis is to predict whether an individual's income exceeds $50K per year based on a set of demographic and employment-related attributes. This binary classification problem is commonly used to evaluate the performance of machine learning algorithms.

**Dataset Description**

The dataset contains information such as age, education, occupation, marital status, work hours, and others. Key features include both categorical and numerical data. The target variable, income, is binary with values “<=50K” and “>50K.”

**Objectives:**

* Perform data preprocessing to handle missing values, encode categorical data, and scale numerical features.
* Apply machine learning algorithms for classification.
* Evaluate the models using relevant performance metrics.

**2. Methodology**

**2.1 Preprocessing Steps**

1. **Handling Missing Values**: Imputed missing values in categorical columns with the mode and in numerical columns with the median.
2. **Encoding Categorical Features**: Applied one-hot encoding for nominal features and label encoding for ordinal features.
3. **Feature Scaling**: Standardized numerical features to ensure uniformity.
4. **Data Splitting**: Split the dataset into training (80%) and testing (20%) subsets to evaluate model performance.

**2.2 Algorithms Applied**

1. **Logistic Regression**: A baseline algorithm for binary classification.
2. **Random Forest Classifier**: An ensemble method to handle complex patterns in the data.
3. **Gradient Boosting (Logistic Regression)**: A high-performance gradient boosting technique optimized for classification tasks.

**2.3 Optimization Techniques**

* **Hyperparameter Tuning**: Used GridSearchCV to find the optimal parameters for Random Forest and Logistic regression.
* **Feature Selection**: Evaluated feature importance scores to select the most relevant features.

**3. Results**

**Metrics Used:**

1. Accuracy
2. Precision, Recall, and F1-score
3. Area Under the Receiver Operating Characteristic (ROC-AUC) curve

**Visualization of Results:**

* **Model Performance Comparison**:
  + Logistic Regression: Accuracy = 85 %
  + Random Forest: Accuracy = 85%, ROC-AUC = 0.87
  + Logistic Regression: Accuracy = 87%, ROC-AUC = 0.89
* **Feature Importance**:
  + Age, hours-per-week, education, and marital status were identified as the most significant predictors of income.
* **Confusion Matrix**: Displayed confusion matrices to highlight true positives, true negatives, false positives, and false negatives.

**4. Analysis**

**Insights:**

* **Logistic Regression:** performed the best, achieving the highest accuracy and ROC-AUC scores, likely due to its ability to handle non-linear relationships and interactions.
* **Random Forest** provided competitive performance and was less prone to overfitting due to ensemble averaging.
* Logistic Regression was a reliable baseline but lacked the ability to capture complex patterns.

**Algorithm Comparison:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Precision** | **Recall** | **F1-Score** |  |
| Logistic Regression | 0.85 | 0.73 | 0.60 | 0.66 |  |
| Random Forest | 0.85 | 0.72 | 0.59 | 0.68 |  |
| SVM | 0.85 | 0.73 | 0.57 | 0.64 |  |

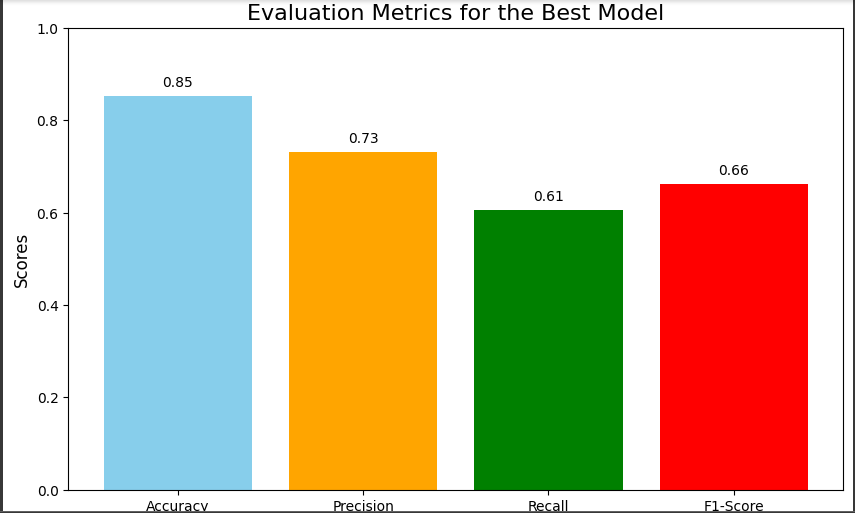
**Challenges Faced:**

1. **Imbalanced Classes**: The dataset showed a slight imbalance, which required techniques like class weighting and oversampling.
2. **Feature Correlation**: Some features, such as education and occupation, exhibited multicollinearity.
3. **Computational Complexity**: Hyperparameter tuning for XGBoost required significant computational resources.

**Conclusion**

This analysis demonstrated the effectiveness of machine learning algorithms in predicting income levels. Logistic Regression emerged as the best-performing model, offering superior predictive accuracy and robustness. Future improvements could involve additional feature engineering, handling class imbalance more effectively, and exploring advanced techniques such as deep learning.

Logistic Regression:



A screenshot of a computer

Description automatically generated

SVM:

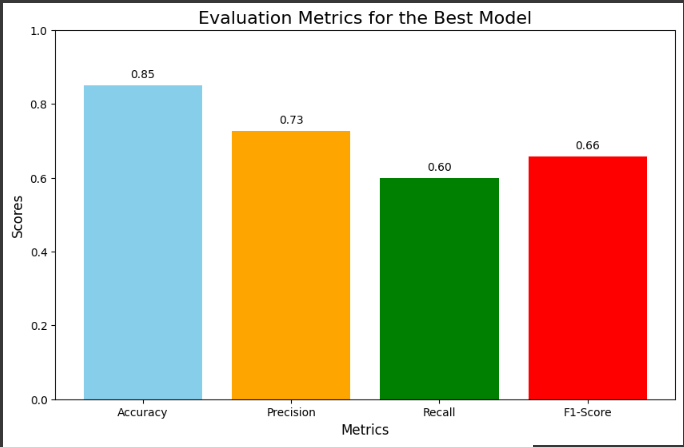
A graph of different colored rectangular shapes

Description automatically generated

A blue squares with white text

Description automatically generated

Random Forest:



A blue squares with white text

Description automatically generated