📑 APR Assignment 1: Predicting Income Using Logistic Regression and SVM on the Adult Dataset (Pranjal Chamaria 2201CS55)

---

1. Introduction

The goal of this experiment is to apply machine learning models taught in class to a real-world dataset obtained via Google Dataset Search. We chose the **Adult (Census Income)** dataset from the UCI Machine Learning Repository. The task is a **binary classification** problem: predicting whether an individual's income exceeds $50K/year based on demographic and employment-related attributes.

We implemented and evaluated two supervised models:

- **Logistic Regression** (baseline, interpretable linear model)
- **Support Vector Machine (SVM)** (margin-based classifier)

The models were compared in terms of classification performance, cross-validation results, and error analysis.

---

2. Dataset Description

- **Source:** UCI Machine Learning Repository, Adult Dataset.
- **Size:** ~48,842 rows (32,561 training + 16,281 testing in original; here combined and re-split).
- **Features:** 14 attributes including both numeric (age, education-num, hours-per-week, capital-gain, etc.) and categorical variables (workclass, education, marital status, occupation, race, sex, native-country).
- **Target variable:** `income` (binary: `<=50K` = 0, `>50K` = 1).
- **Preprocessing:**
    - Removed rows with missing values.

    - Encoded categorical features using **one-hot encoding**.
    - Scaled numerical features using **StandardScaler**.
    - Final split: 80% training, 20% test (stratified).

---

3. Models Implemented

3.1 Logistic Regression

- A linear classification model that predicts the probability of an individual earning >50K.
- Uses the **sigmoid function** and optimizes a log-likelihood objective with **L2 regularization**.
- **Grid search hyperparameters tested:**
    - Regularization strength: C ∈ {0.01, 0.1, 1, 10}
    - Penalty: L2
    - Solver: lbfgs
- **Best Parameters:** C=0.1, penalty=L2, solver=lbfgs

3.2 Support Vector Machine (SVM)

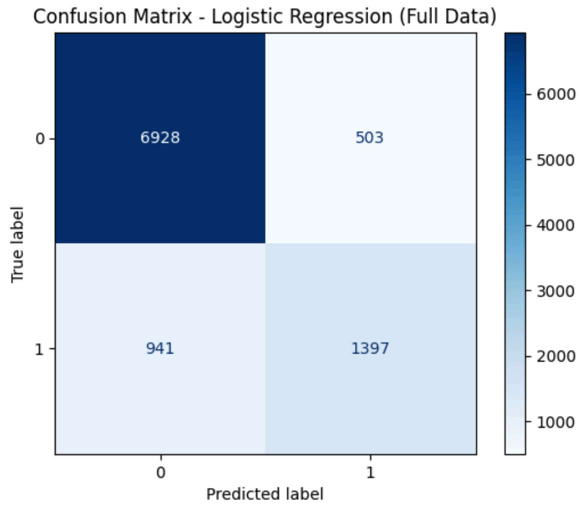
- A margin-based classifier that separates classes by maximizing the decision boundary margin.
- Due to computational constraints, SVM was trained on a **10k subset** of the dataset.
- Only **linear kernel** was used (scales better than RBF for high-dimensional sparse data).
- **Grid search hyperparameters tested:**
    - C ∈ {0.1, 1, 10}
    - Kernel = linear

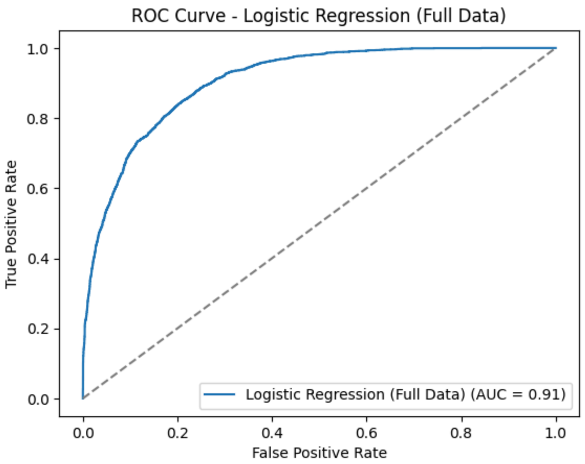
- **Best Parameters:** C=0.1, kernel=linear

---

4. Results and Evaluation

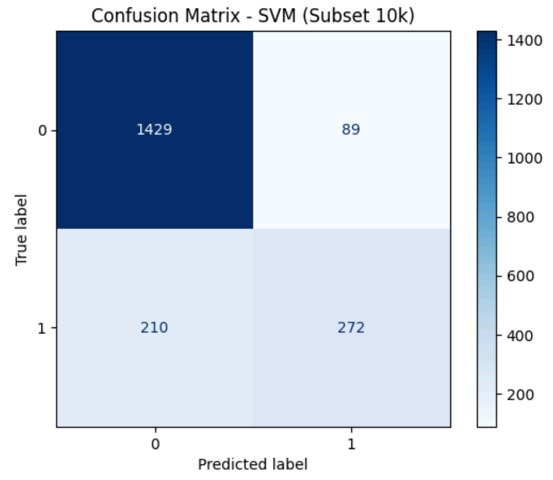
4.1 Logistic Regression (Full Dataset)

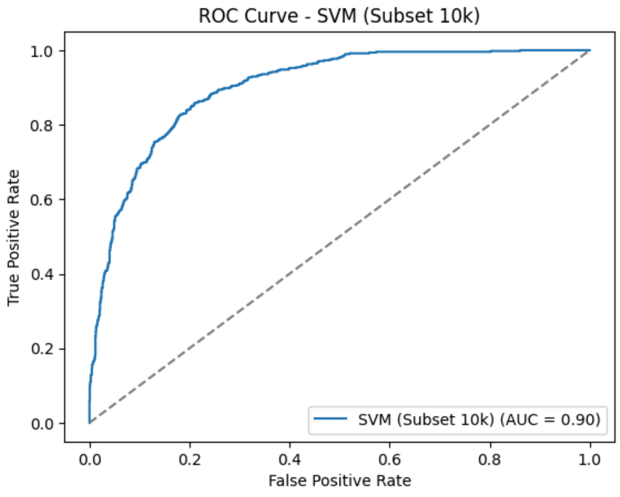




- **Best CV Accuracy:** 85.27%
- **Test Results (9,769 samples):**
    - Accuracy: 85%
    - Precision (class 1): 0.74
    - Recall (class 1): 0.60
    - F1-score (class 1): 0.66
- **Confusion Matrix:**
    - True Negatives: 6,928
    - True Positives: 1,397
    - False Negatives: 941
    - False Positives: 503
- **ROC-AUC:** 0.91 (from ROC curve)

4.2 SVM (Subset 10k)





- **Best CV Accuracy:** 84.88%
- **Test Results (2,000 samples):**
  - Accuracy: 85%
  - Precision (class 1): 0.75
  - Recall (class 1): 0.56
  - F1-score (class 1): 0.65
- **Confusion Matrix:**
  - Similar distribution as Logistic Regression, with slightly more false negatives.
- **ROC-AUC:** 0.90 (from ROC curve)

---

5. Comparative Analysis

- Both models achieved **very similar accuracy (~85%)**, showing that linear models are well-suited for this dataset.
- **Logistic Regression** had slightly better recall for high-income individuals (class 1), meaning it identified more true positives.
- **SVM** showed marginally higher precision for class 1, meaning when it predicts >50K, it's slightly more reliable.

- ROC-AUC values were nearly identical, indicating both models rank predictions well.

- Logistic Regression is preferable here due to:

  - **Faster training** (scales better on full dataset).

  - **Interpretability** (coefficients provide insights into feature importance).

  - Similar or slightly better recall for the minority class.

---

6. Conclusion

This study demonstrated the application of **Logistic Regression and SVM** on the Adult Census Income dataset. Both models achieved **~85% accuracy** with similar performance. Logistic Regression emerged as the more practical and interpretable choice for this dataset, while SVM provided comparable results on a reduced subset. The results highlight the strengths of linear classifiers for structured tabular data, while also indicating potential gains from more advanced models in future work.