📘 Technical Report: Adult Census Income Prediction

Internship Program: AICTE & Edunet Foundation (Employability Skills & Digital Literacy with AI)

Student Name: Kunal Singh Jayant

Student ID: STU68406b02b70631749052162

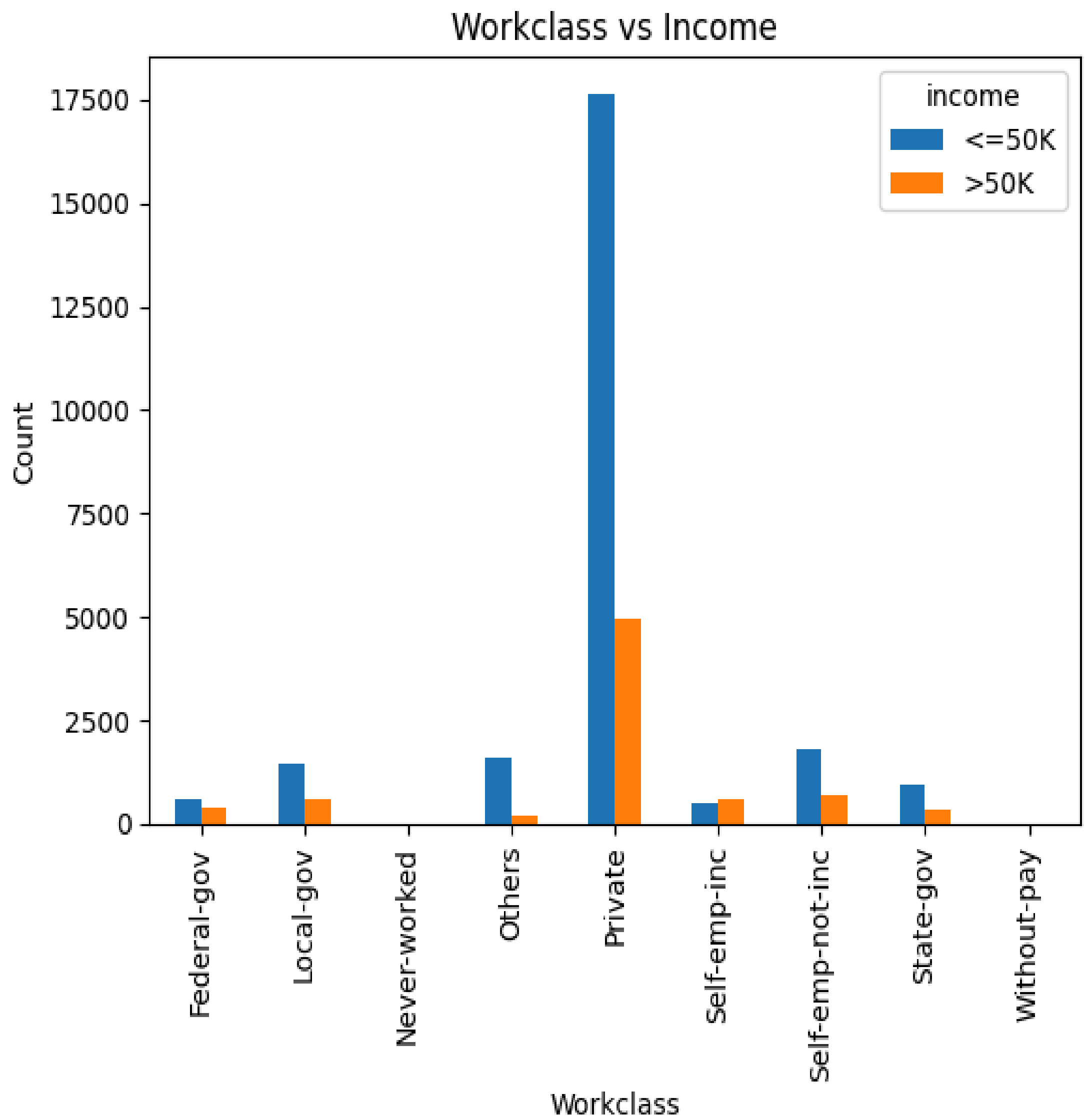
Date: December 2025

1. 📌 Project Abstract

This project focuses on building a Binary Classification system using the Adult Census Income dataset (48k+ records). The goal is to predict individual income levels based on demographic features. The project follows a rigorous Data Science lifecycle, including deep Exploratory Data Analysis (EDA), statistical cleaning, and the implementation of a Multi-Layer Perceptron (MLP) Neural Network.

2. 📊 Initial Data Distribution

The first step involved understanding our target variable: income. We identified a significant class imbalance between individuals earning ≤50K and those earning >50K.



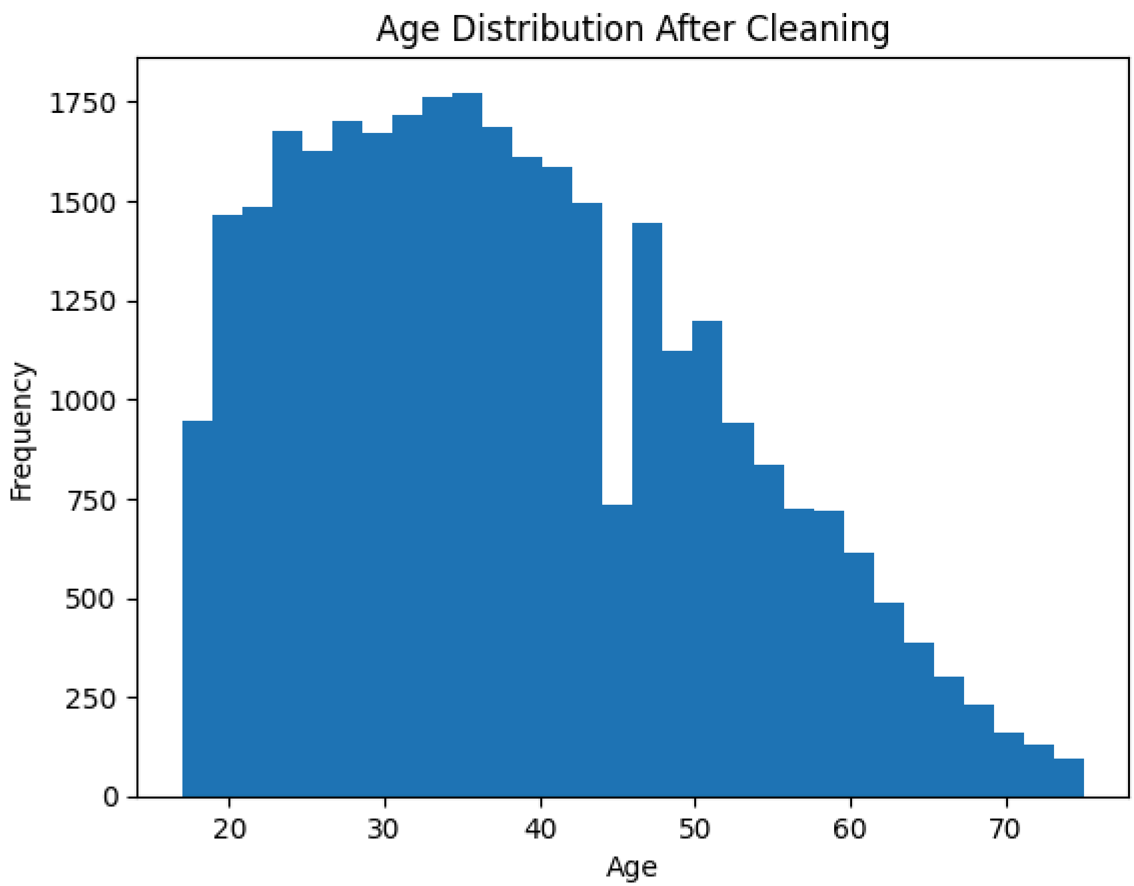
Bar Chart of Income Distribution

Analysis: This chart highlights the majority class (<=50K), which informs our decision to use Precision, Recall, and F1-Score rather than just simple Accuracy.

3. 🔍 Exploratory Data Analysis & Outlier Detection

To ensure model stability, we used box plots to identify statistical anomalies in the numerical features.

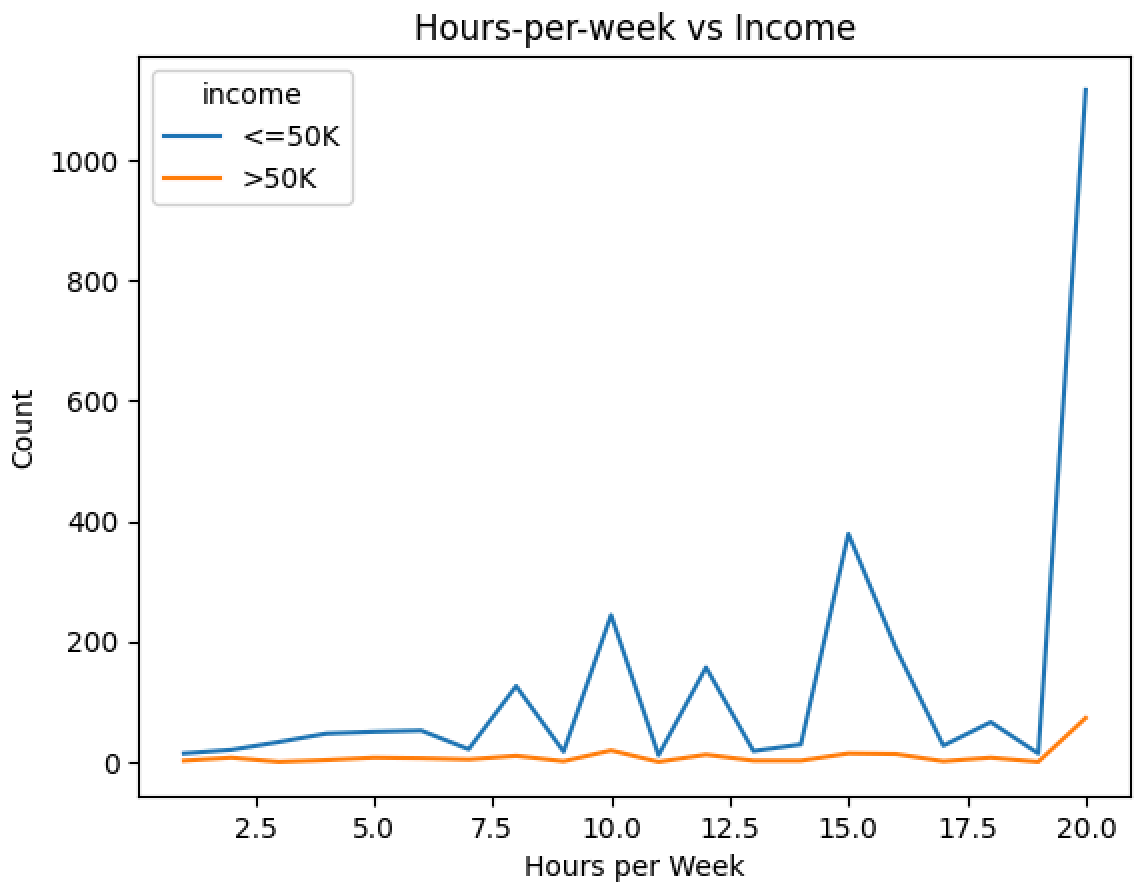
3.1 Age Analysis



Box Plot or Histogram for Age (After Cleaning)

Result: Post-cleaning, the age distribution is more concentrated within the active workforce range (17-75), reducing noise and improving model reliability.

3.2 Working Hours Analysis

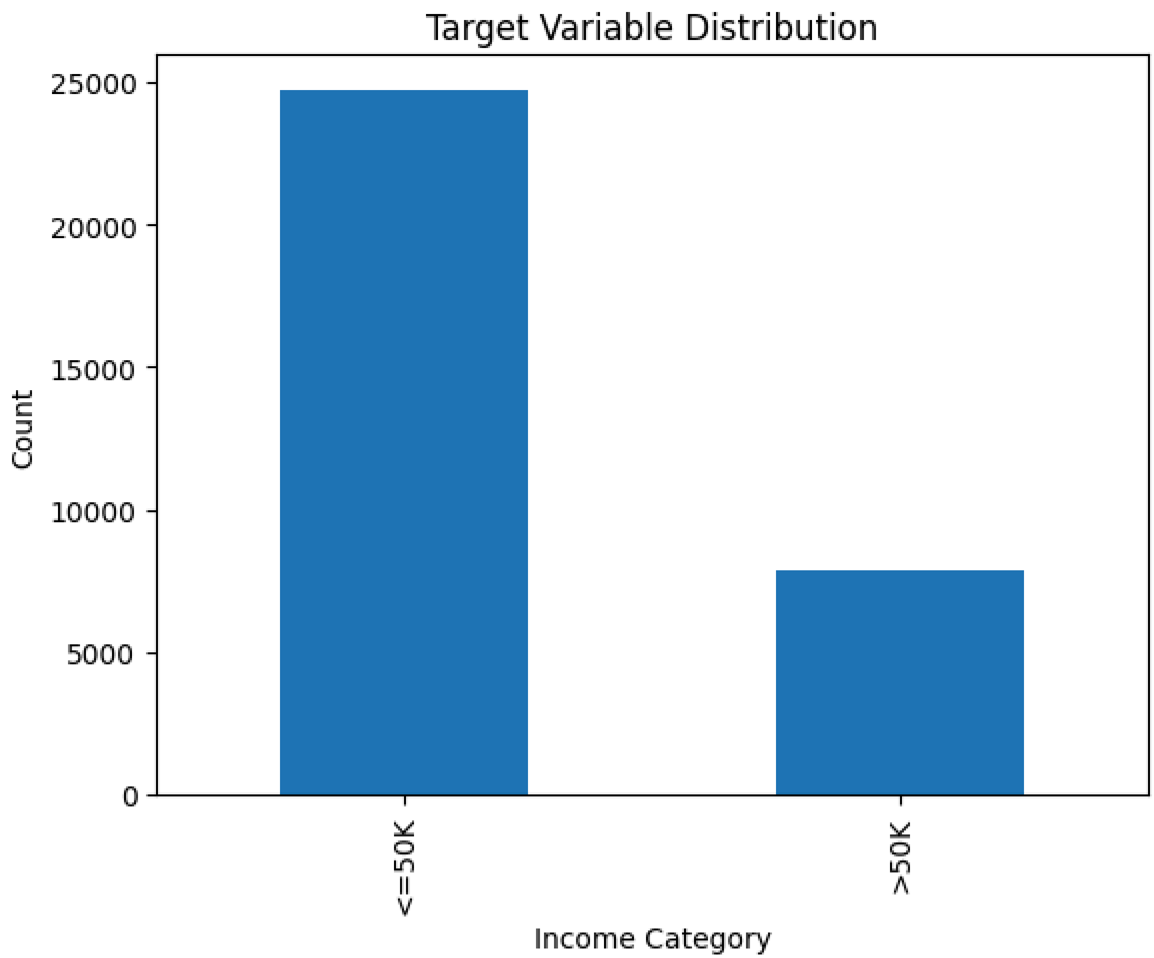


Box Plot for Hours-per-week

Finding: This visualization helped identify unrealistic labor data, which was filtered to normalize the input range.

4. 🧬 Categorical Insights

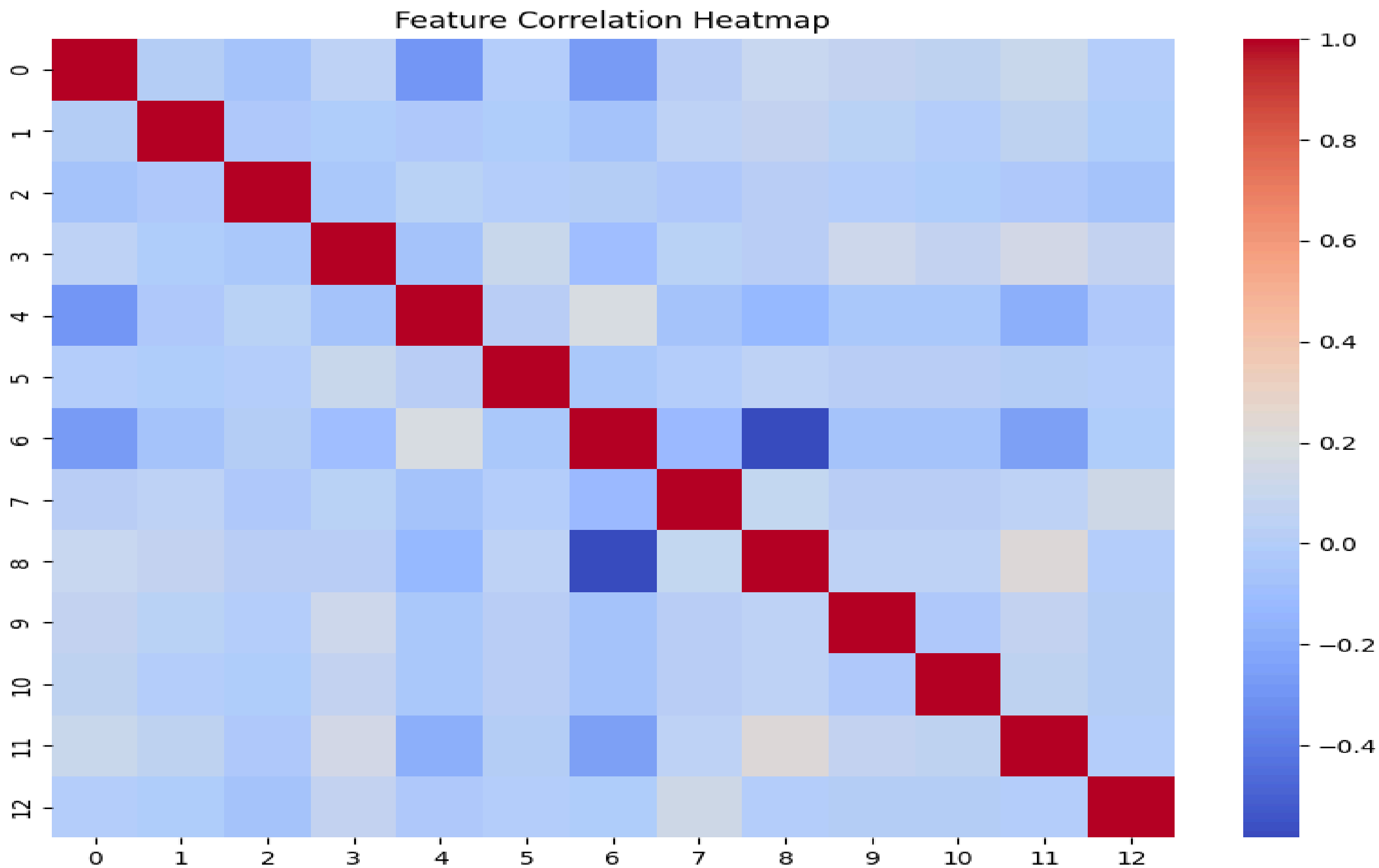
Understanding the impact of education and occupation on income levels was key to feature selection.



Countplot of Education or Workclass *Observation: Private sector employment and Bachelor's degrees emerged as dominant features in the dataset.*

5. 🛠️ Feature Engineering & Correlation

We performed feature reduction to avoid multicollinearity and improve processing speed.

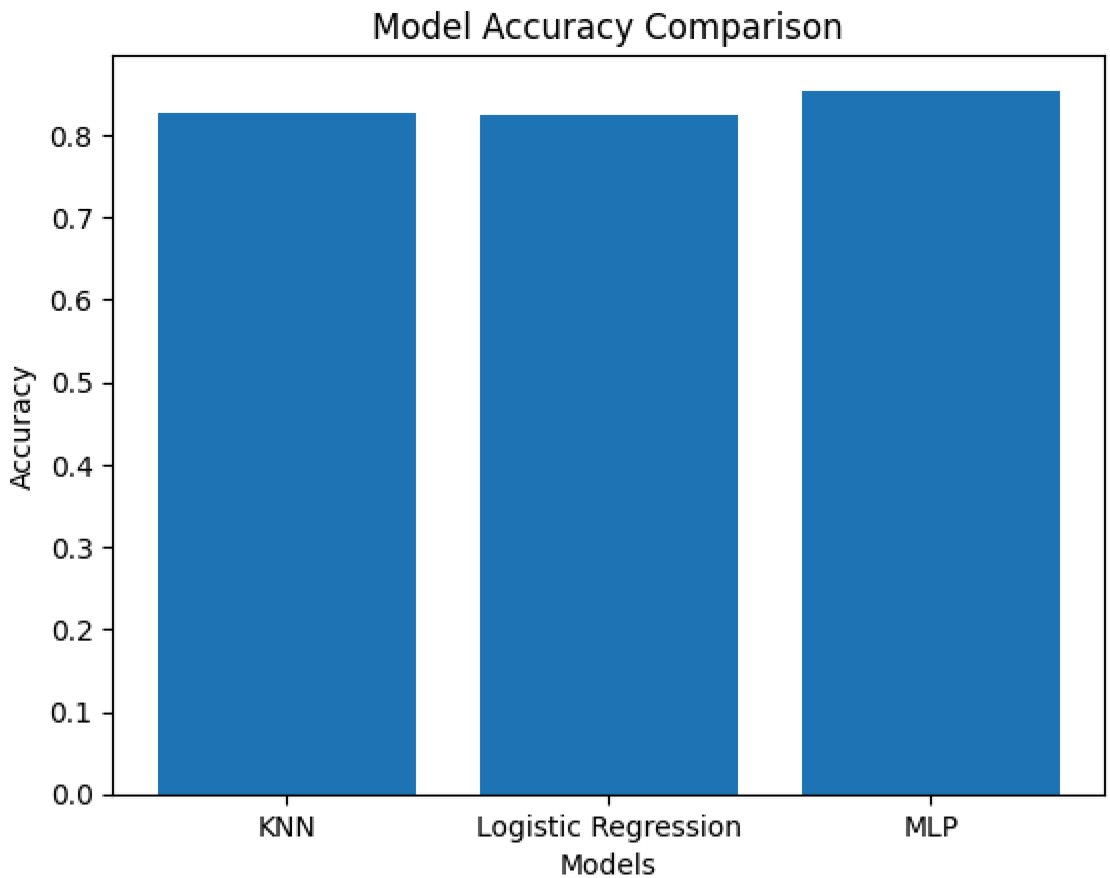


Feature Correlation Heatmap

Technical Note: The heatmap confirmed a perfect correlation between 'Education' and 'Educational-num'. Consequently, the categorical 'Education' column was dropped to simplify the model.

6. 🏆 Model Benchmarking & Performance

I implemented and compared three architectures: K-Nearest Neighbors (KNN), Logistic Regression, and a Multi-Layer Perceptron (MLP) Neural Network

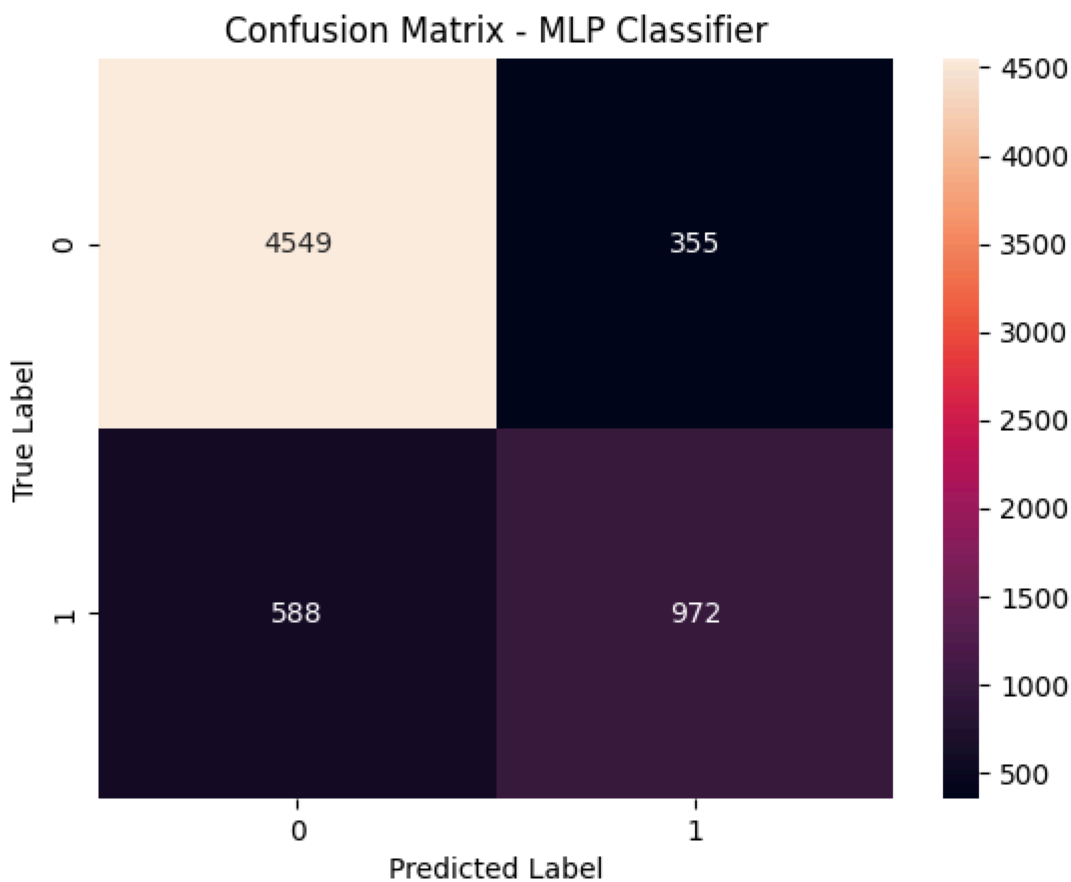


Comparison Bar Chart of Model Accuracies

Result: The MLP Classifier achieved the highest accuracy of ~82.1%, proving the effectiveness of neural hidden layers in capturing non-linear demographic patterns.

7. 📉 Final Evaluation (MLP)

The performance of the winning model was evaluated using a Confusion Matrix to visualize True Positives and True Negatives.



Confusion Matrix Heatmap (MLP Results)

Conclusion: The model demonstrated high reliability in predicting the majority class while maintaining a competitive F1-score for the high-income class.

8. 🏁 Conclusion

By combining rigorous statistical preprocessing with a Multi-Layer Perceptron architecture, this project successfully predicts income levels with an accuracy of **82.1%**. The internship experience at **Edunet Foundation** provided the necessary technical and digital literacy skills to bridge the gap between raw data and actionable AI insights.

Certified by Edunet Foundation in collaboration with AICTE & IBM SkillsBuild.