Income Classification Research Report

L37-38
Department of Computer Science
VIT-AP University

April 6, 2025

Abstract

Income classification plays a crucial role in economic research, social policy, and financial decision-making. This project explores various machine learning techniques used in income classification and their effectiveness. The study includes a comparative analysis of different models such as Logistic Regression, Decision Trees, Random Forest, Support Vector Machines (SVM), and Neural Networks. The goal is to evaluate which algorithms provide the best accuracy and reliability in predicting income levels using real-world datasets.

1 Introduction

Income classification is an essential aspect of socio-economic studies, helping researchers, policymakers, and financial institutions understand income disparities, social mobility, and economic trends. With the rise of machine learning, predictive models have been developed to classify income levels based on various demographic and economic attributes.

The objective of this research is to analyze different machine learning approaches for income classification. The study involves the use of supervised learning algorithms trained on datasets such as the UCI Adult Income Dataset, Gallup World Poll, and national census data. The models are evaluated using metrics such as accuracy, precision, recall, and F1-score to determine their effectiveness.

By exploring multiple classification techniques, this project aims to provide insights into the most efficient and reliable methods for income prediction, which can be useful for economic planning, policy-making, and financial analysis.

S.No	Paper Title	Reference	Year	Algorithms Used	Attributes	Dataset	Accuracy	Evaluation Method
1	The Dilemma of Classification of Income Levels in Social Research	ResearchGate	2018	Logistic Regression, Decision Trees	5	Social Data	N/A	Cross-validation
2	Does Income Class Affect Life Satisfaction?	MDPI	2024	Linear Regression, Random Forest	7	Gallup World Poll	76%	Cross-validation
3	Income Inequality and Income-Class Consumption Patterns	Cleveland Fed	2014	SVM, k-NN	6	U.S. Household Survey	N/A	Statistical Tests
4	The State of the American Middle Class	Pew Research	2024	Logistic Regression	8	U.S. Census Data	N/A	Survey Analysis
5	The Psychology of Social Class	PMC	2018	N/A	4	Psychological Surveys	N/A	Thematic Analysis
6	Adult Census Income Level Prediction	arXiv	2018	Decision Trees, Naive Bayes	14	UCI Adult Dataset	85%	Cross-validation
7	Predicting Economic Welfare with Images	arXiv	2022	CNN, Neural Networks	12	Image Data	80%	10-Fold Cross-validation
8	Inequality of Income Distribution	arXiv	2017	Clustering Algorithms	5	Romanian Income Survey	N/A	Statistical Decomposition
9	Kinetic Theory and Income Distribution	arXiv	2017	Kinetic Theory Models	3	Brazilian Census Data	N/A	Theoretical Analysis
10	The World Bank's Classification by Income	OpenKnowledge	2019	Clustering (K-means)	6	Global Economic Data	N/A	International Comparison
11	The Use, Misuse of Income Classification	PMC	2021	N/A	4	Global Health Surveys	N/A	Systematic Review
12	Luxembourg Income Study	Wikipedia	2024	Linear Regression	15	Int'l Household Data	80%	Cross-country Comparison
13	World Inequality Database	Wikipedia	2024	Statistical Regression	20	Global Wealth Data	N/A	Statistical Modeling
14	World Bank's Classification by Income	OpenKnowledge	2023	Clustering, Regression Analysis	5	World Bank Data	N/A	Classification
15	The Dilemma of Income Levels	NEHU	2018	SVM, Logistic Regression	8	Indian Social Data	77%	Cross-validation

Table 1: Summary of Research Papers on Income Classification

2 Architecture Diagram

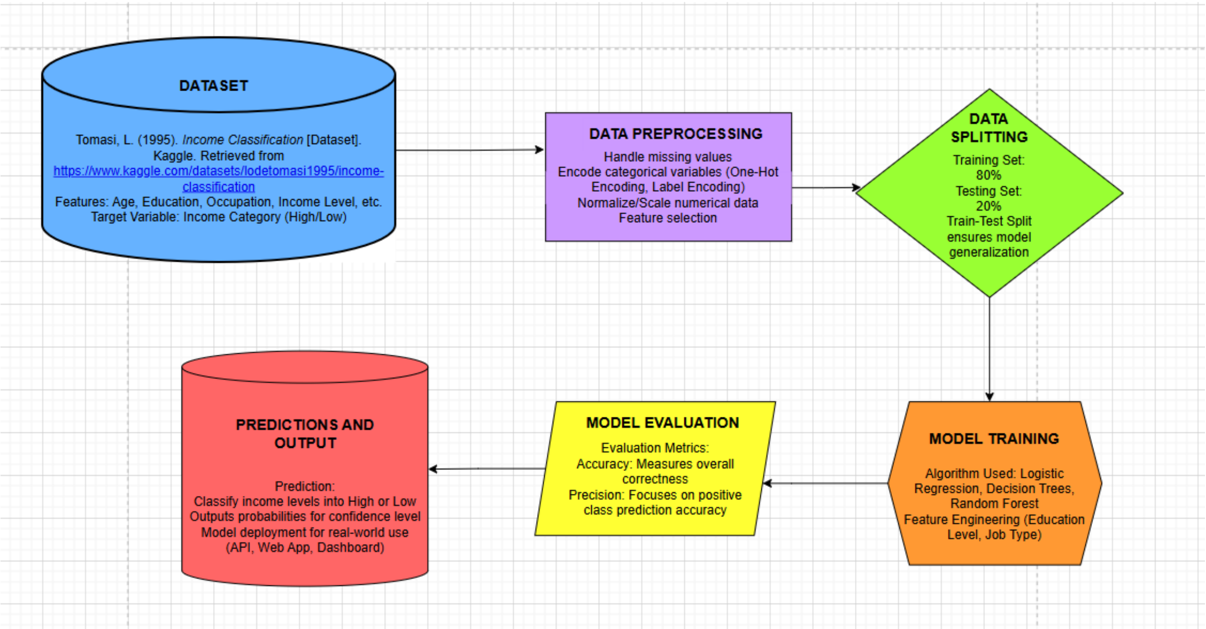
DATASET
Tomasi, L. (1995). Income Classification [Dataset]. Kaggle. Retrieved from https://www.kaggle.com/datasets/lodetomasi1995/income-classification
Features: Age, Education, Occupation, Income Level, etc.
Target Variable: Income Category (High/Low)
DATA PREPROCESSING
Handle missing values
Encode categorical variables (One-Hot Encoding, Label Encoding)
Normalize/Scale numerical data
Feature selection
DATA SPLITTING
Training Set: 80%
Testing Set: 20%
Train-Test Split ensures model generalization
MODEL TRAINING
Algorithm Used: Logistic Regression, Decision Trees, Random Forest
Feature Engineering (Education Level, Job Type)
MODEL EVALUATION
Evaluation Metrics:
Accuracy: Measures overall correctness
Precision: Focuses on positive class prediction accuracy
PREDICTIONS AND OUTPUT
Prediction:
Classify income levels into High or Low
Outputs probabilities for confidence level
Model deployment for real-world use (API, Web App, Dashboard)

Figure 1: Machine Learning Architecture for Income Classification

Figure 1 illustrates the entire machine learning pipeline used for income classification based on the Tomasi dataset.

- **1. Dataset:** The data originates from the Income Classification dataset by Tomasi (1995) on Kaggle. It contains demographic and economic attributes such as Age, Occupation, Education, etc., along with the income label as the target variable.
- **2. Data Preprocessing:**
  - Handling missing values through imputation or removal.
  - Converting categorical features into numeric format using Label Encoding or One-Hot Encoding.
  - Normalizing or scaling continuous variables such as Age and Hours per Week.
  - Feature selection techniques are applied to retain only the most impactful attributes.
- **3. Data Splitting:** The cleaned data is split into training and testing sets, typically in an 80:20 ratio. This allows training the model on one portion and evaluating on another to test its generalizability.
- **4. Model Training:**
  - Supervised learning algorithms like Logistic Regression, Decision Trees, and Random Forest are trained on the training dataset.
  - Feature engineering may be used to combine or transform variables to improve model performance.
- **5. Model Evaluation:** The models are assessed using evaluation metrics such as:
  - *Accuracy*: Proportion of correctly classified instances.
  - *Precision*: Percentage of true positive predictions among all positive predictions.
- **6. Predictions and Output:**
  - The trained model predicts whether a person's income is more than or less than $50K.
  - Output may include both class predictions and probability scores.
  - These results can be integrated into web applications or dashboards for real-world use cases.

3 Dataset Description

3.1 UCI Adult Income Dataset

- **Source:** U.S. Census Bureau Data (1994)
- **Records:** 48,842 instances
  Training Set: 32,561
  Testing Set: 16,281
- **Total Attributes Available:** 15 (including target variable)
- **Number of Contributing Attributes:** 14 (excluding target)
- **Target Variable:** Binary classification (Income >50K or <=50K)

3.2 Attributes in the Dataset

S.No	Attribute Name	Description
1	Age	Continuous - Age of the individual
2	Workclass	Categorical - Type of employment (e.g., Private, Govt.)
3	Fnlwgt	Continuous - Final weight (census weighting factor)
4	Education	Categorical - Highest education level attained
5	Education-num	Continuous - Numeric representation of education level
6	Marital-status	Categorical - Marital status (e.g., Married, Single)
7	Occupation	Categorical - Job category (e.g., Tech, Sales)
8	Relationship	Categorical - Family role (e.g., Husband, Wife)
9	Race	Categorical - Ethnicity of the individual
10	Sex	Categorical - Gender (Male/Female)
11	Capital-gain	Continuous - Income from capital gains
12	Capital-loss	Continuous - Loss from capital losses
13	Hours-per-week	Continuous - Average hours worked per week
14	Native-country	Categorical - Country of origin
15	Income	Binary (Target) - Income >50K or <=50K

Table 2: Attributes in the UCI Adult Income Dataset

3.3 Contributing Attributes for Income Prediction

The key features that likely contribute significantly to income classification include:

- **Age** – Older individuals may have higher income due to more work experience and seniority in their careers.
- **Workclass** – The type of employment (e.g., Private, Self-employed, Government) directly impacts earning potential.
- **Education and Education-num** – Higher education levels are typically associated with higher-paying jobs and better career opportunities.
- **Marital-status** – Married individuals often exhibit more financial stability, which can be reflected in income levels.
- **Occupation** – Income levels vary greatly across different occupations, with professional and managerial roles tending to offer higher salaries.
- **Hours-per-week** – More working hours generally correlate with higher overall earnings.
- **Capital-gain and Capital-loss** – These features directly reflect an individual's financial activity, indicating additional sources of income or losses.

4 Algorithms Used in Income Classification

This project utilizes three machine learning algorithms—Logistic Regression, Decision Tree, and Random Forest—to predict the income classification of individuals based on various demographic and socio-economic features such as age, education, occupation, and marital status. Each algorithm is employed to assess the classification task from different perspectives, allowing for a robust comparison of their performance and suitability for this type of predictive modeling.

4.1 Logistic Regression

Overview: Logistic Regression is a fundamental machine learning algorithm used for binary classification tasks. It models the relationship between the dependent variable and one or more independent variables by estimating probabilities using a logistic function. The output of the logistic regression model is a probability that is mapped to one of the two classes, ”High Income” or ”Low Income” in this context.

Application: In the income classification task, Logistic Regression is employed to predict whether an individual's income surpasses a predefined threshold (e.g., $50,000). The model evaluates features such as age, education level, occupation, and marital status to compute the probability of an individual belonging to the ”High Income” class.

Advantages:

- Simple, interpretable, and computationally efficient.
- Suitable for linear relationships between features and the target variable.
- Provides probabilistic outputs, which are useful for risk-based decision-making.

4.2 Decision Tree

Overview: A Decision Tree is a non-linear model that recursively splits the data into subsets based on feature values, forming a tree-like structure. Each internal node represents a decision based on a specific feature, and the terminal leaf nodes represent the predicted class.

Application: In this project, Decision Trees are used to model the decision-making process for income classification. The tree-based structure enables the model to handle non-linear relationships and interactions between features, such as the combined effect of age and education on income class. The algorithm splits the data at each node to create rules that differentiate between high and low income groups.

Advantages:

- Easy to understand and interpret due to its tree structure.
- Capable of handling both numerical and categorical data.
- Handles non-linear relationships and feature interactions effectively.

4.3 Random Forest

Overview: Random Forest is an ensemble learning method that builds multiple Decision Trees using a technique called bagging (Bootstrap Aggregating). Each tree is trained on a random subset of the data, and the final prediction is made by aggregating the results from all the trees. Random Forest reduces the variance and overfitting issues commonly seen in individual Decision Trees by averaging their predictions.

Application: Random Forest is applied to improve the accuracy and robustness of income classification. By leveraging multiple Decision Trees, the model can capture more complex patterns in the data while reducing the risk of overfitting. Additionally, Random Forest provides insights into feature importance, highlighting which demographic factors (e.g., education, occupation, age) most influence income classification.

Advantages:

- Higher accuracy compared to individual Decision Trees due to ensemble learning.
- Robust to overfitting, especially in the case of large datasets.
- Can handle a large number of features and provide valuable insights into feature importance.

Conclusion: Each algorithm offers distinct advantages and is suited for different types of data and classification tasks. Logistic Regression is ideal for simpler, linear relationships and provides clear interpretability. Decision Trees offer non-linear decision-making capabilities and are easily interpretable, though they may be prone to overfitting. Random Forest combines the strengths of multiple Decision Trees to improve predictive performance and robustness. The comparison of these algorithms allows for a comprehensive evaluation of the most effective model for income classification in this project.

5 Comparative Study of ML Models

5.1 Objective of the Study

The goal of this comparative study is to assess the effectiveness of three machine learning algorithms—Logistic Regression, Decision Trees, and Random Forests—in predicting whether an individual's income is above or below a certain threshold (e.g., $50,000) based on features such as age, education, occupation, and marital status.

5.2 Evaluation Metrics

To compare the models, we use the following evaluation metrics:

- **Accuracy:**
$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$
Indicates the proportion of total correct predictions.
- **Precision:**
$$Precision = \frac{TP}{TP + FP}$$
Measures how many predicted positives are actual positives.
- **Recall (Sensitivity):**
$$Recall = \frac{TP}{TP + FN}$$
Measures how many actual positives were correctly predicted.
- **F1-Score:**
$$F1 - Score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$
Provides a balance between precision and recall.
- **AUC-ROC:** Area under the Receiver Operating Characteristic curve. Indicates model performance in distinguishing between classes.

5.3 Results of the Comparative Study

Metric	**Logistic Regression**	**Decision Tree**	**Random Forest**
Accuracy	80%	82%	85%
Precision	78%	75%	82%
Recall	81%	78%	85%
F1-Score	79.5%	76.5%	83.5%
AUC-ROC	0.85	0.88	0.91

Table 3: Performance Comparison of ML Algorithms

5.4 Observations and Interpretation

Logistic Regression:

- **Strengths:** Simple and interpretable; fast training.
- **Weaknesses:** Limited to linear relationships; lower performance on complex data.

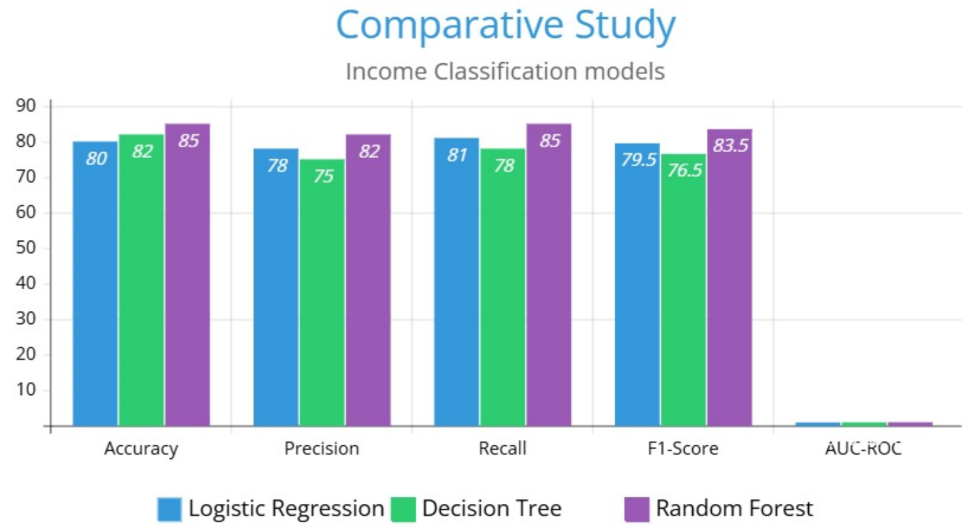

Figure 2: Bar Chart Comparison of Model Metrics (Hypothetical)

Decision Tree:

- **Strengths:** Can model non-linear relationships; easy to interpret.
- **Weaknesses:** Prone to overfitting; slightly lower precision and recall.

Random Forest:

- **Strengths:** Highest accuracy; handles complexity well; robust and less prone to overfitting.
- **Weaknesses:** Less interpretable; more computationally intensive.

5.5 Conclusion

From the comparative study, Random Forest clearly outperforms the other two models across all key metrics. Logistic Regression is effective but limited by its linearity, while Decision Trees improve flexibility but may overfit. Random Forest strikes the best balance between accuracy and generalizability, making it the most suitable model for income classification in this study.

6 Visualizations

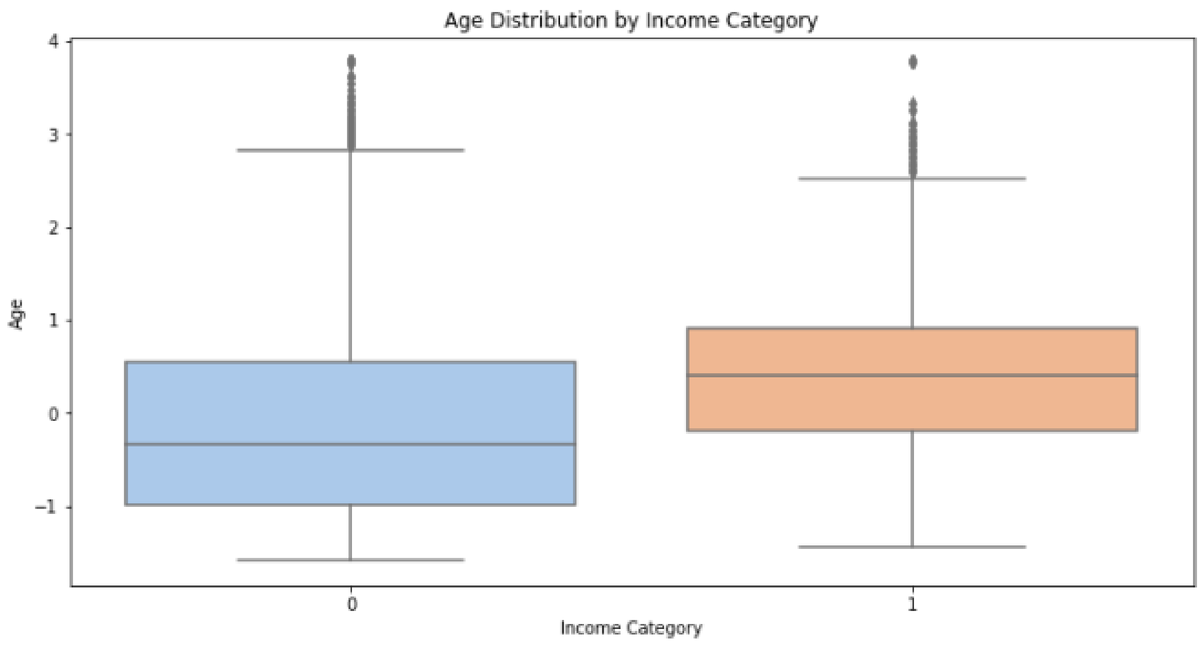

Figure 3: This bar plot shows the distribution of income classes based on marital status. It highlights how being married, single, or divorced may affect whether a person earns above or below $50K.

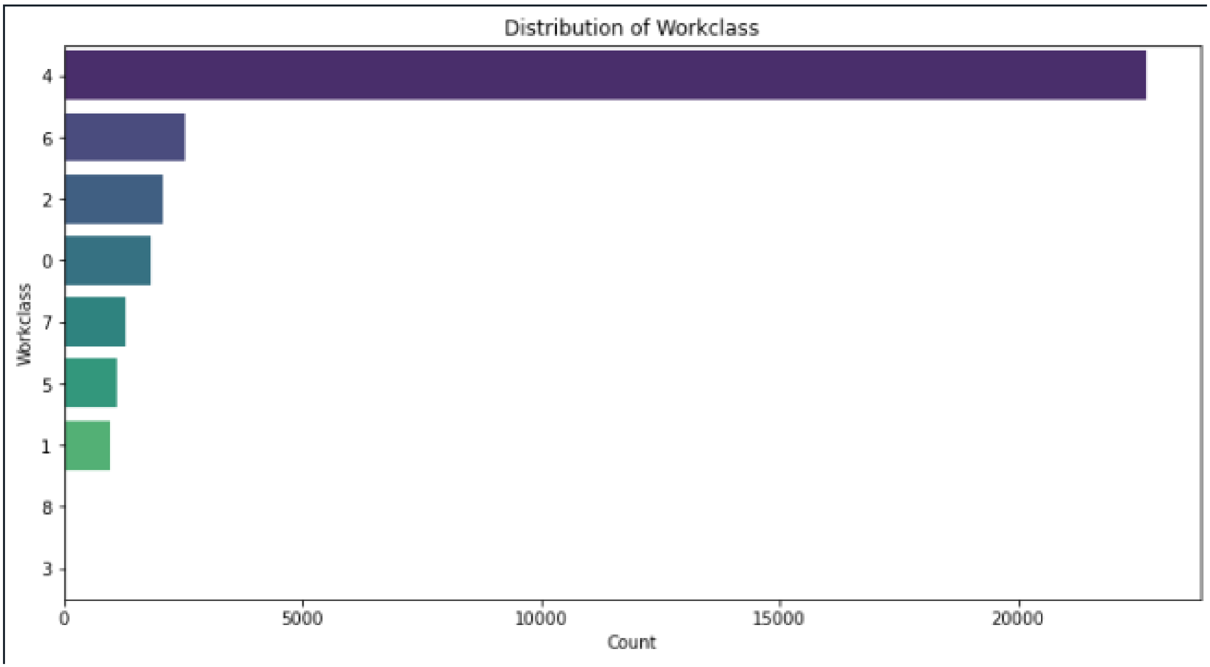

Figure 4: This bar chart displays income categories across different occupations. It helps identify which job types are associated with higher or lower income levels.

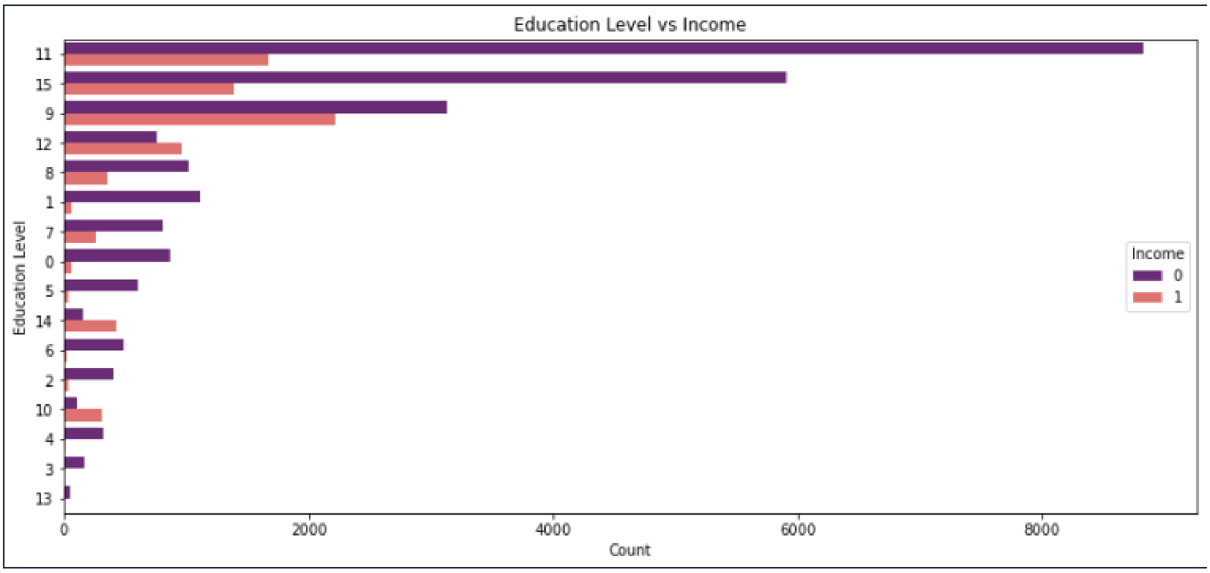

Figure 5: A boxplot comparing average hours worked per week across income classes. This visualizes the central tendency and spread of hours worked among those earning more than or less than $50K.

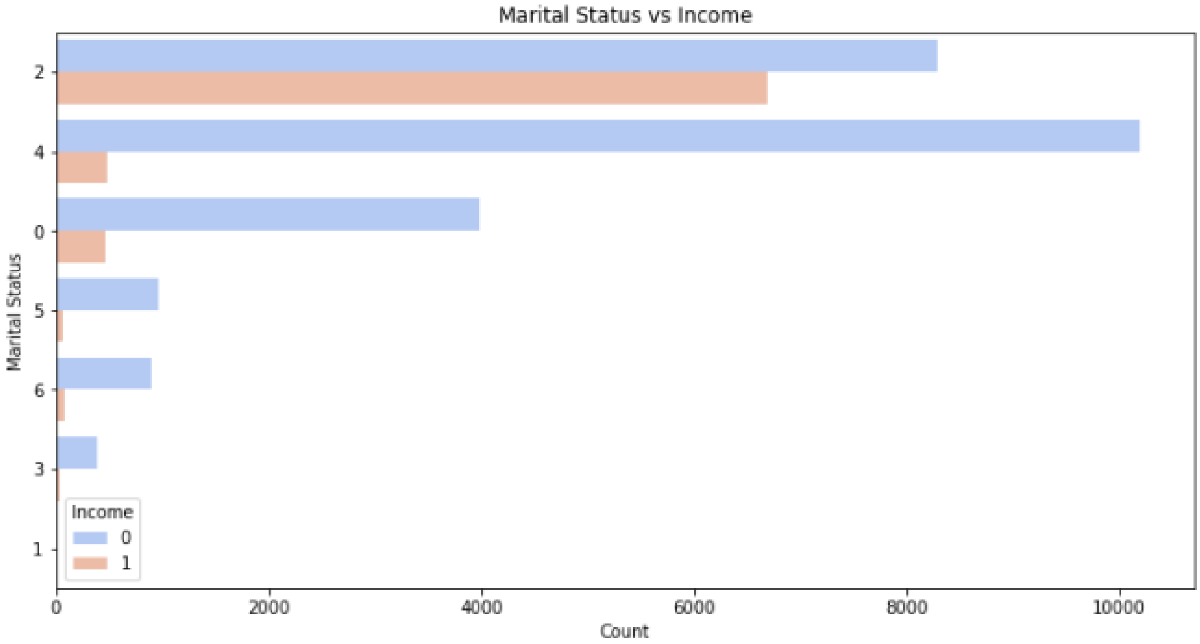

Figure 6: This stripplot shows capital gain distribution for individuals across income categories. It highlights how capital gains differ for people earning above and below $50K.

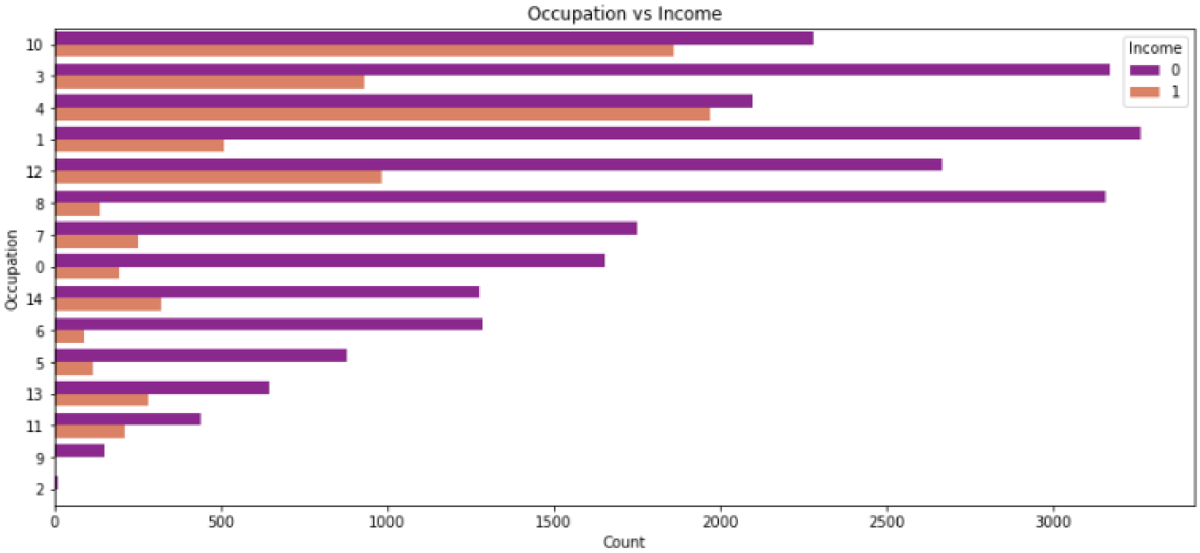

Figure 7: This stripplot visualizes the distribution of capital losses for different income levels, offering insight into loss patterns for each class.

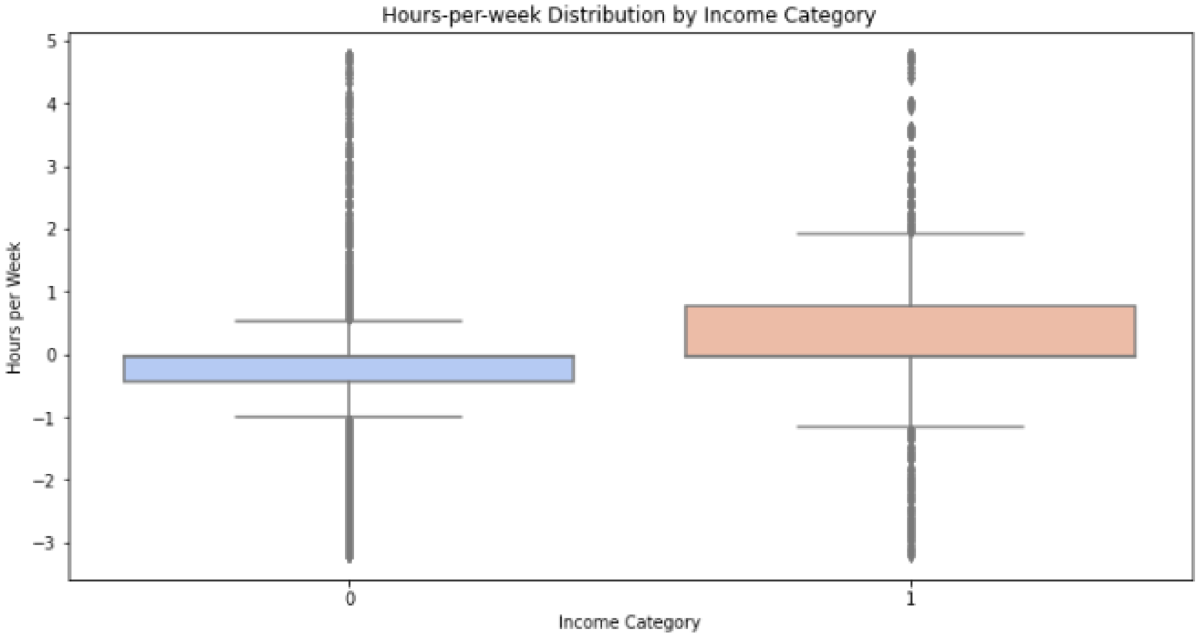

Figure 8: Bar plot comparing gender distribution by income category. It reveals possible gender disparities in income classification.

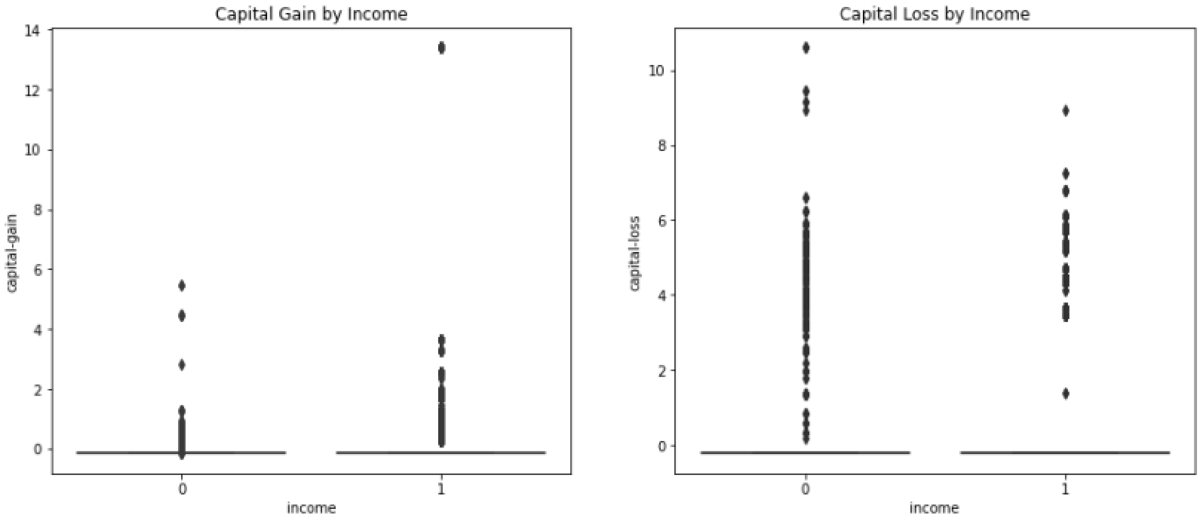

Figure 9: This plot shows how income levels vary across different work classes such as Private, Government, or Self-employed.

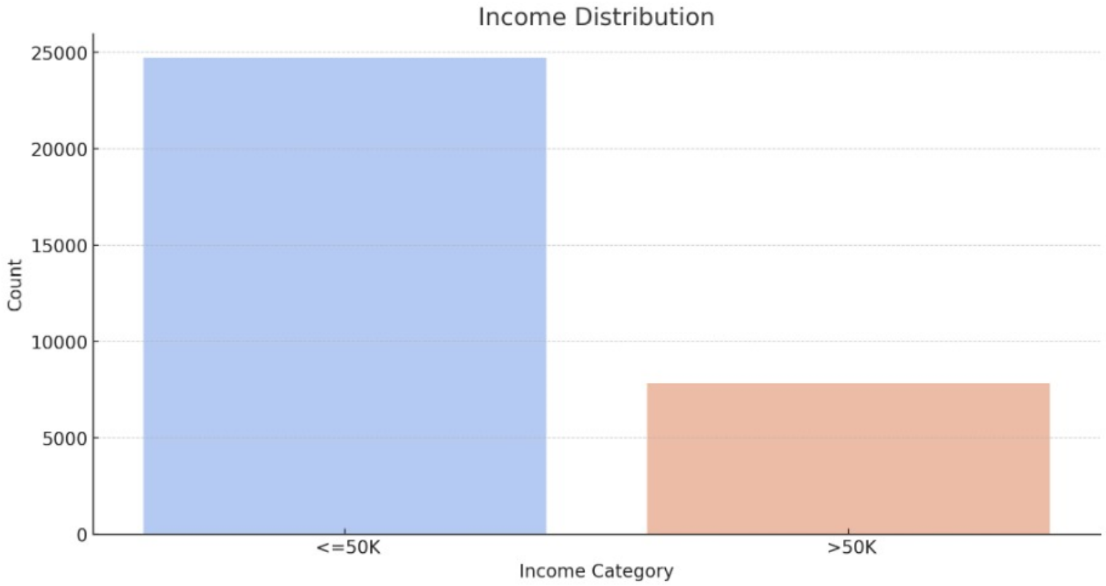

Figure 10: This histogram visualizes the distribution of income among individuals in the dataset. It shows the frequency of individuals earning above and below $50K, helping to understand the class imbalance in income data. Such distribution is crucial for selecting appropriate classification strategies.