Socioeconomic study: Adult Data Set

Paulina Iwach-Kowalska 254362,
Kamil Iwach-Kowalski 262300

June 25, 2024

Contents

1 Introduction

The analysis of income distribution and its determinants continues to be a critical area of research due to its profound implications on economic policies, workforce development, and social welfare programs. Understanding the complex interplay of factors that influence income levels is essential for shaping interventions that promote equitable economic growth. This study employs the Adults dataset from the UC Irvine Machine Learning Repository, a rich source of demographic, educational, and occupational data, to explore these dynamics.

The primary goal of this research is to identify and analyze the key factors that contribute to an individual's likelihood of earning above or below the annual income threshold of \$50 000. This threshold often distinguishes between middle-class and lower-income earners, making it a pivotal point for policy and economic analysis.

The specific research questions addressed in this study are:

1. How do factors such as education, occupation, and marital status correlate with income levels?
2. What role does educational attainment play in influencing an individual's earning potential?
3. How does marital status affect economic stability and income?
4. What insights can be drawn about gender disparities in income?
5. What results are we able to obtain in terms of classifying people by creating different decision boundaries?

By addressing these questions, this study aims to delineate the intricate relationships between various demographic variables and income. The potential benefits of this analysis are manifold. Understanding which occupations lead to higher incomes can guide workforce development initiatives and educational advising, ensuring that training programs align closely with market demands. Meanwhile, the findings identifying personal characteristics that have a significant impact on income level can provide an indication of which social groups require new development programs and benefits. Moreover, gathering information on potential financial disparities across different demographics, such as gender or race, is important for informing policy decisions and educational strategies aimed at promoting equality.

2 Data characteristics

The analysed dataset was retrieved from the UC Irvine Machine Learning Repository [2] and is part of the Census database selected in 1994 by Barry Becker. It contains 48 842 records, of which 3620 rows have missing data. The set includes 14 independent variables and one dependent variable, which are described in detail in Table 1. Additionally, Table 2 shows sample records from the dataset.

Table 1: Description of the variables in the dataset

Feature	Type	Meaning
`workclass`	categorical	Employment status of the individual.
`education`	categorical	Level of education.
`education-num`	categorical	Level of education, represented by the numbers of years of schooling.
`marital-status`	categorical	Marital status of the individual.
`occupation`	categorical	Occupation category.
`relationship`	categorical	Relationship of the individual in terms of a family member.
`race`	categorical	Race of the person.
`native-country`	categorical	Country of origin for the individual.
`age`	continuous	Age of the individual.
`fnlwgt`	continuous	Weight of the record that is affected by certain socio-economic conditions. It can be interpreted as the number of people with similar characteristics.
`capital-gain`	continuous	Capital gains in dollars.
`capital-loss`	continuous	Capital loss in dollars.
`hours-per-week`	continuous	Number of working hours per week.
`sex`	boolean	Biological sex of the individual
`income`	boolean	Dependent variable indicating whether an individual earns more than \$50 000 per year. True if > 50K, False if ≤ 50K.

Table 2: Exemplary records from the dataset

age	workclass	fnlwgt	education	education-num	marital-status	occupation
39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical
50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial
38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners
28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty

relationship	race	sex	capital-gain	capital-loss	hours-per-week	native-country	income
Not-in-family	White	Male	2174	0	40	United-States	$\leq$ 50K
Husband	White	Male	0	0	13	United-States	$\leq$ 50K
Not-in-family	White	Male	0	0	40	United-States	$\leq$ 50K
Wife	Black	Female	0	0	40	Cuba	$\leq$ 50K

The dataset is unbalanced. There are 37 155 records where income is $\leq$ 50K, accounting for approximately 76% of the dataset. The remaining 24% of the set (exactly 11 687 rows) consist of individuals with income $>$ 50K.

3 Methodology

3.1 Data Preparation

A few anomalies were noted in the dataset, which were corrected before detailed analysis began. One of the errors identified was in the income column. This variable is expected to contain only two unique values; however, four distinct values were observed: "$\leq$ 50K", "$>$ 50K", "$\leq$ 50K.", and "$>$ 50K.". The data was corrected by removing the dots, resulting in the appropriate values of "$\leq$ 50K" and "$>$ 50K". Further attention was drawn to the handling of missing data. In several columns, missing values were indicated with a "?". These occurrences have been converted to NaN values to standardize the dataset. Overall, data gaps were identified across three variables, as detailed in Table 3.

Table 3: Variables with number of missing values

Variable	Number of missing values
workclass	2799
occupation	2809
native-country	857

Figure 1 shows the relationship between capital gain values and age. The value of 99999 for capital gains significantly deviates from the rest of the observations. This may have been the way missing values were marked. Additionally, these outliers are observed in the subset of individuals earning more than \$50 000 annually. Consequently, these values have been replaced with the average capital gain, calculated only for the group with earnings above \$50 000, after excluding records containing the value of 99999.

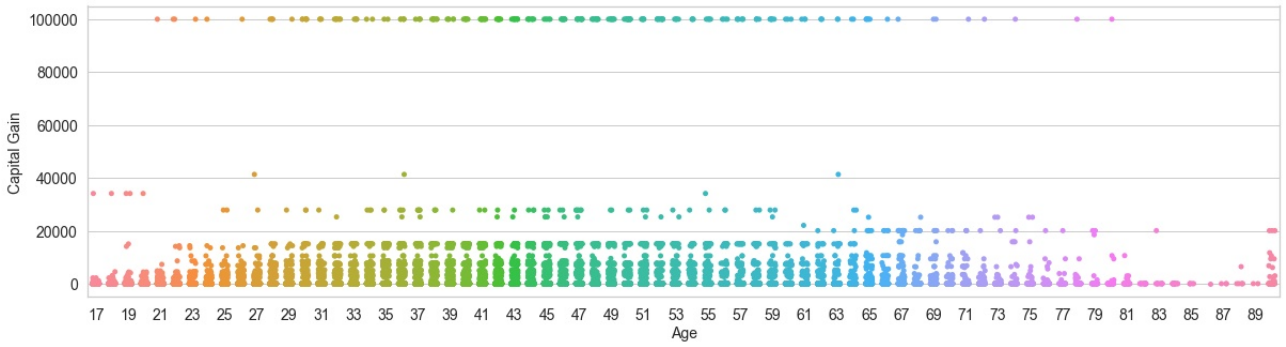

Figure 1: Dependence of capital gain on age

The effect of removing outliers can be seen in Figure 2. A clear reduction in the maximum values can be seen, the pattern and correlations in the values are more noticeable.

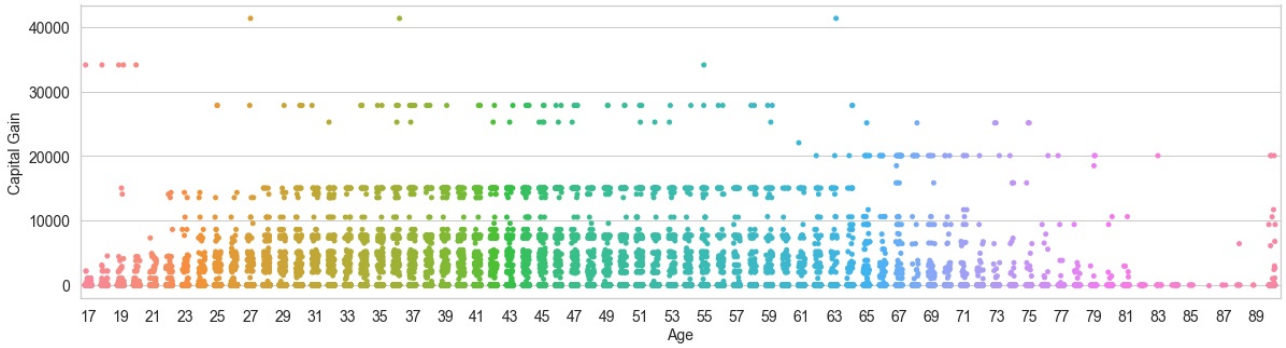

Figure 2: Dependence of capital gain on age after removal of outliers

3.2 Exploratory Data Analysis

3.2.1 Univariate Analysis

To analyze the properties of the independent variables included in the dataset, key statistical measures were computed. For numerical variables, arithmetic means, standard deviations, minimum and maximum values, as well as quartile values, were examined. For qualitative variables, the number of unique values and the most frequent value along with its frequency were determined. The results are described in detail in the Section 4.1

3.2.2 Analysis of Relationships between Variables

To explore the relationships between variables and the influence of individual factors on the income target, a detailed analysis was conducted using the exploratory techniques listed below. Results of this examination are presented in Section 4.2.

Pearson correlation [3] coefficient, denoted as r, is a measure of the linear correlation between two variables X and Y. It quantifies the degree to which a relationship between the two variables can be described by a line. The coefficient ranges from -1 to 1, where:

- $r = 1$ indicates a perfect positive linear relationship,
- $r = -1$ indicates a perfect negative linear relationship,
- $r = 0$ indicates no linear relationship.

The formula for the Pearson correlation coefficient is

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}},$$

where x_i, y_i are the individual sample points with i index, n is sample size and $\overline{x}$, $\overline{y}$ are the means of x and y.

Cramér's V correlation [4] is a measure of the relationship between two nominal variables, providing a value in the range from 0 to 1, where values close to 1 indicate a high correlation. The statistic is based on the chi-squared statistic χ^2. The formula for Cramér's V is

$$V = \sqrt{\frac{\chi^2}{n \cdot \min(k-1, r-1)}},$$

where n – number of observations, k – the number of columns, r – number of rows and χ^2 statistic is derived from Pearson's chi-squared test

$$\chi^2 = \sum_{i=1}^{n} = \frac{(O_i - E_i)^2}{E_i},$$

where O_i – number of type i observations and E_i – the expected count of type i.

Box plots were used to compare the distribution of numeric variables, specifically focusing on the median, quartiles, and the number of outliers.

Bar plots were utilized to analyze the proportions between income groups across various quality characteristics.

Density distributions plots were used to visualise differences in distributions for various income groups. These plots are particularly useful for comparing underlying patterns, trends, and for identifying skewness and peaks.

3.3 Classification

In this study, a systematic approach was adopted for the creation of classifiers, centered around the exploration of optimal scaling techniques and classification algorithms through a robust cross-validation framework. Initially rows containing missing values were dropped to ensure the integrity of the analysis. Subsequently, categorical variables were transformed into dummy variables using the `drop_first=True` option to avoid multicollinearity, effectively reducing the dimensionality of the data. Also the `relationship_Wife` column was removed to prevent duplication of information (because of the `marital-status` column). The process involved constructing pipelines comprising two components: a scaling method and a classification estimator. The entire analysis was then performed on this processed data, ensuring that the models developed were based on clean and appropriately formatted inputs. Results of the classifier evaluations and optimizations are detailed in Section 4.3.

The depicted Figure 3 outlines a methodology for parameter and feature selection to avoiding data leakage. Initially, the dataset is divided into training and test sets, with the test set reserved strictly for the final evaluation to prevent any influence on the training process. After cross-validation, the model is retrained on the complete training dataset and subsequently undergoes evaluation using the previously untouched test data, ensuring that the assessment is unbiased and reflects the model's performance on new, unseen data.

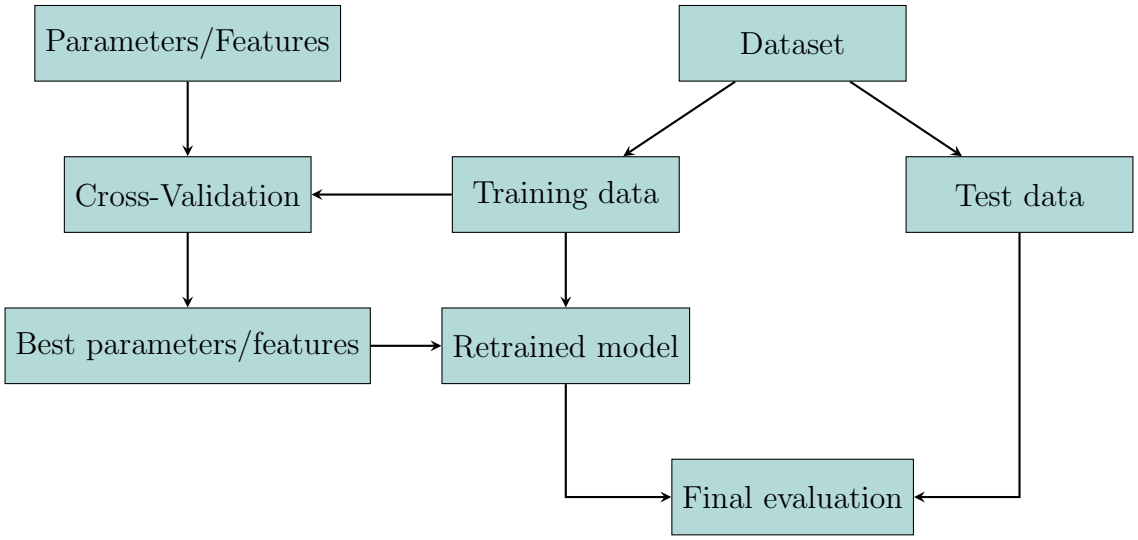

Figure 3: A schematic overview of the model training and evaluation process. This flowchart demonstrates the steps involved in selecting the best parameters and features through cross-validation, training, and final evaluation, ensuring the integrity and generalizability of the model without data leakage.

3.3.1 Scaling Methods

The scaling component of the pipelines utilized two distinct methods: `StandardScaler` and `MinMaxScaler`. The `StandardScaler` standardizes features by removing the mean and scaling to unit variance. Conversely, `MinMaxScaler` transforms features by scaling each feature to a range $[0, 1]$.

3.3.2 Classification Estimators and Hyperparameters

Multiple classification algorithms were evaluated to determine the most effective models for our dataset. The selection process involved tuning various hyperparameters for each classifier

using a structured grid search methodology. The algorithms and their respective hyperparameters are detailed below:

Linear Discriminant Analysis (LDA) For LDA, the main hyperparameter considered was `shrinkage`, which is used to improve the estimator's accuracy by regularizing the covariance matrix. The possible values for `shrinkage` included:

- None (no shrinkage),
- "auto" (automatic shrinkage using an estimator),
- Fixed values ranging from 0.01 to 0.99 at various intervals.

Quadratic Discriminant Analysis (QDA) QDA's primary tuning parameter was the `reg_param` (regularization parameter), which helps prevent overfitting by adjusting the conditioning of the problem and smoothing the likelihood estimations. The grid for `reg_param` included values from 0.0 to 1.0 at 0.1 increments.

Logistic Regression Logistic Regression was fine-tuned using multiple hyperparameters:

- `C`$\in [0.01, 0.1, 1, 10, 100]$, the inverse of regularization strength.
- `solver`, for which algorithms "newton-cg", "lbfgs", "sag", "liblinear", "saga" were considered based on their suitability for different penalties.
- `penalty`, included "l1", "l2", "elasticnet", or "none", defining the type of regularization applied.
- `max_iter`$\in [100, 200, 300]$, the maximum number of iterations for solvers to converge.

Decision Tree Classifier The Decision Tree Classifier's hyperparameters included:

- `max_depth`$\in$[None, 10, 20, 30, 40, 50], allowing the tree to expand until a specified depth or fully if None, to control model complexity.
- `min_samples_split`$\in [2, 5, 10, 20]$, defining the minimum number of samples required to split an internal node.
- `min_samples_leaf`$\in [1, 2, 4, 10]$, specifying the minimum number of samples required to be at a leaf node.
- `max_features`$\in$["auto", "sqrt", "log2", None], determining the number of features to consider when looking for the best split.
- `criterion`, with options "gini" or "entropy" for measuring the quality of a split.

K-Neighbors Classifier For the K-Neighbors Classifier, parameters tuned were:

- `n_neighbors`$\in [3, 5, 7, 10, 15]$, the number of neighbors to use for k-nearest neighbors voting.
- `weights`, either "uniform" where all points in each neighborhood are weighted equally, or "distance" where closer neighbors have a greater influence on the outcome.
- `p`, defining the power parameter for the Minkowski metric used in distance calculations, exploring values 1 through 3.

Random Forest Classifier Random Forest involved an extensive array of hyperparameters aimed at optimizing the ensemble's predictive performance:

- `n_estimators`, number of trees in the forest, tested at 100, 200, 300, 500, and 1000.
- `max_features`, the number of features to consider when looking for the best split, with options "auto", "sqrt", and "log2".
- `max_depth`, the maximum depth of the trees, with options ranging from 10 to 50 and an option for unrestricted growth (None).
- `min_samples_split`, the minimum number of samples required to split an internal node, tested at 2, 5, and 10.
- `min_samples_leaf`, the minimum number of samples required at a leaf node, tested at 1, 2, and 4.
- `criterion`, the function to measure the quality of a split, with options "gini", "entropy", and "log_loss".

Table 7 summarizes the optimal hyperparameters for each classifier, as determined through grid search analysis.

3.3.3 Cross-Validation and Hyperparameter Tuning

Tuning of hyperparameters was conducted using `GridSearchCV` combined with `RepeatedStratifiedKFold`, which involves stratified sampling to ensure that each fold is a good representative of the whole. Specifically, the data was split into 10 folds, and the process was repeated three times to mitigate the variability in the cross-validation results. The primary metric for optimization was the F1 score, a harmonic mean of precision and recall, providing a balance between the two in scenarios with uneven class distributions.

3.3.4 Evaluation Metrics

The results from the cross-validation process were meticulously analyzed to ascertain the best combination of scaler and estimator, based on the F1 score (see Figure 9). The superior performance of the `StandardScaler` was identified, leading to its selection for subsequent model training. The effectiveness of all classifiers with the standard scaler was then thoroughly assessed on both training and test datasets. Metrics such as accuracy, precision, recall, and F1 score were computed, providing a comprehensive view of model effectiveness. Additionally, the confusion matrix, for example shown in Figure 11, illustrates the number of correct and incorrect predictions made by the model, which is crucial for understanding its performance in a nuanced manner.

A confusion matrix is a table used to describe the performance of a classification model on a set of data for which the true values are known. It allows for the visualization of the accuracy of a model by comparing the actual values with the predictions. The matrix is divided into four parts: true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN).

- **Accuracy** measures the overall correctness of the model and is defined as the ratio of correct predictions (both true positives and true negatives) to the total number of cases examined. The equation for accuracy is given by

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}.$$

- **Precision** is the ratio of correctly predicted positive observations to the total predicted positives and is a measure of the accuracy of the positive predictions. It is defined as

$$\text{Precision} = \frac{TP}{TP + FP}.$$

- **Recall** (or sensitivity) measures the ability of the model to find all the relevant cases (all actual positives). It is defined as

$$\text{Recall} = \frac{TP}{TP + FN}.$$

- **F1 Score** is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account. It is especially useful when the class distribution is uneven. The F1 score is the harmonic mean of precision and recall

$$\text{F1 Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}.$$

3.3.5 Feature Selection

Feature selection was systematically conducted using Weight of Evidence (WoE) and Information Value (IV) [1]. The Weight of Evidence indicates the predictive power of an independent variable in relation to the dependent variable. This method provides a measure for assessing the strength of the relationship between each feature and the outcome variable. The Weight of Evidence is defined as:

$$WoE_i = \ln\left(\frac{N_i}{P_i}\right) - \ln\left(\frac{\sum N_i}{\sum P_i}\right),$$

where P represents the occurrence of an event, N represents the non-occurrence of the event, and i indexes the evaluated feature in a given class.

Information Value helps quantify the overall predictive power of a feature. It is calculated from the sum of the products of the WoE of each category and the difference in the proportions of goods and bads. Typically, features with an IV value greater than 0.1 are considered strong predictors, whereas those below 0.02 are regarded as weak. The Information Value is defined as:

$$IV = \sum_{i=1}^{n} \left(\frac{N_i}{\sum N_i} - \frac{P_i}{\sum P_i}\right) \cdot WoE_i,$$

where P represents the occurrence of an event, N represents the non-occurrence of the event, n is the total number of evaluated features in the class, and WoE_i is the Weight of Evidence for the i-th feature.

Table 4: Interpretation of Information Value

Information Value (IV)	**Predictive Power**
< 0.02	Useless for prediction
0.02 to 0.1	Weak predictor
0.1 to 0.3	Medium predictor
0.3 to 0.5	Strong predictor
> 0.5	Suspicious or too good to be true

As shown in the table 4, IV values less than 0.02 are considered useless for prediction, indicating that the feature does not contribute significantly to the outcome variable. Values

between 0.02 and 0.1 are deemed weak predictors, having a minor influence. Features with IV values from 0.1 to 0.3 are considered medium predictors, while those between 0.3 and 0.5 are strong predictors, likely to be very useful in a predictive model. An IV greater than 0.5 might be suspicious and suggest that the variable is too good to be true. However, in the case of unbalanced datasets (like ours), IV values greater than 0.5 are not unusual and may not necessarily indicate a problem, reflecting instead the disparity in the distribution of the outcome variable.

In this study, new models based on Random Forest and Quadratic Discriminant Analysis (QDA) were trained by iteratively selecting the most important feature based on IV. Each selected feature was added one at a time to the training set to observe the incremental improvement in model performance. This stepwise approach allowed for the careful evaluation of each feature's contribution to the model's predictive accuracy. The results from these models are described in Section 4.4.

3.3.6 Cost Sensitive Learning

In this study, cost-sensitive learning was implemented to address the challenge of class imbalance within the dataset, which often leads to biased predictive models that favor the majority class. To mitigate this issue, a Random Forest Classifier was employed with varying class weights, enhancing the model's sensitivity towards the minority class. The Repeated Stratified K-Fold cross-validation method was used on training set. Calculated metrics such as the F1 score, precision, and recall for each model configuration. The results are described in the section 4.5.

4 Results

4.1 Univariate Analysis

To explore the characteristics of the features included in the dataset, key statistical measures were calculated for each variable. Table 5 shows the results for the numerical type attributes.

The age of participants ranges from 17 to 90 years, with an average age of 38.64. The standard deviation of 13.71 indicates a moderately diverse age group. The quartile values reveal that 25% of the dataset is under 28 years old, 50% are below 37 years, and 75% are younger than 48 years. This distribution suggests that the dataset consists mainly of working-age adults, which is the expected correct age for persons in this type of dataset, focusing on earnings.

The fnlwgt statistics show a wide range from 12 285 to 1 490 400. For this feature, the standard deviation is high at 105 604, indicating a significant variability in weights among the study participants. However, while the maximum value is very high, the third quartile shows that the majority of the population has a weight less than 23 7642. It is possible that the greater part of the set are individuals with similar socio-economic conditions whose fnlwgt do not differ significantly.

Educational attainment, measured in years, is an average of 10.08 with a standard deviation of 2.57. Thus, the majority of adults have attained at least a middle school level of education.

Quartiles equal to 0 for the capital gain and capital loss variables indicate that a larger proportion of individuals in the dataset are not investing capital. The average capital gain of \$589.47, the large maximum value of \$41 310 and the standard deviation of \$2 532.27 highlights the presence of significant outliers. Compared to the average gain, the average capital loss is relatively low at \$87.5, with the maximum loss reaching \$4 356.

The average weekly working time is approximately 40.42 hours, which corresponds to a standard full-time working week. However, the range of values from 1 to 99 hours per week

indicates the presence of both part-time and more than full-time workers.

Table 5: Basic statistics for numerical variables

	age	fnlwgt	education-num	capital-gain	capital-loss	hours-per-week
mean	38.64	189 664.13	10.08	589.47	87.50	40.42
std	13.71	105 604.03	2.57	2 532.27	403.00	12.39
min	17.00	12 285.00	1.00	0.00	0.00	1.00
Q1	28.00	117 550.50	9.00	0.00	0.00	40.00
Q2	37.00	178 144.50	10.00	0.00	0.00	40.00
Q3	48.00	237 642.00	12.00	0.00	0.00	45.00
max	90.00	1 490 400.00	16.00	41 310.00	4 356.00	99.00

For the qualitative variables, Table 6 shows the number of unique categories, the most frequent value and its count for each characteristic. Education levels among the population are categorized into 16 unique classifications, with "HS-grad" being the most prevalent, observed in 15 784 individuals. This suggests that the highest educational attainment for the largest segment of the population is a high school diploma. The dominant sector in the workclass variable is the private sector, with as many as 33 906 individuals employed. Another important observation is that married individuals are the most numerous group. It can also be seen that men represent the majority in the dataset. This may be due to the fact that in the 1990s, a family model in which mainly the man undertook paid work was still popular. An important finding is the great number of groups in the native country, as many as 41 different countries. However, the majority of the observations come from the USA, which is more than 91% of the total set. Therefore, for this analysis, it is important to keep in mind that the conclusions are mainly valid for this country and may not be as accurate for the rest of the world. Also, with the race of the population, the collection is based mainly on white people.

Table 6: Basic statistics for qualitative variables

	workclass	education	marital-status	occupation	relationship	race	sex	native-country
unique	8	16	7	14	6	5	2	41
top	Private	HS-grad	Married-civ-spouse	Prof-specialty	Husband	White	Male	United States
freq	33 906	15 784	22 379	6 172	19 716	41 762	32 650	43 832

Due to the predominant number of people in the USA and the small size of the groups of other countries, it was decided to divide the native-country column into two groups: the "USA", for people from the United States, and "other" for persons from all other countries.

4.2 Analysis of Relationships between Variables

The heat matrix presented in Figure 4 shows the Pearson correlation values between the numerical variables. This allows to check how the individual variables affect each other. The

analysis showed no significant correlation between the characteristics. The highest value was between capital gain and age at and was 0.11. This is a weak correlation, but may suggest a slight tendency for capital gains to increase as individuals age. This could indicate that older peoples accumulate more wealth or engage more successfully in activities that lead to capital gains. Similarly, there is a very weak positive correlation of 0.07 between age and hours per week, indicating that older individuals might work marginally more hours, though this relationship is not strong enough to suggest a reliable trend.

The correlations between fnlwgt and economic factors such as capital gain, capital loss, and hours per week are all very close to zero. This seems to imply that the final weight, which adjusts for the number of people with similar characteristics, has almost no effect on these economic variables. This finding may be important because it confirms the statistical independence of economic results from the weights applied to the records.

Additionally, it can be seen that the correlation between capital gain and capital loss is also very low and hence insignificant. This suggests that individuals with higher capital gains do not necessarily experience capital losses. Perhaps capital gains and losses are independent economic events affected by different factors.

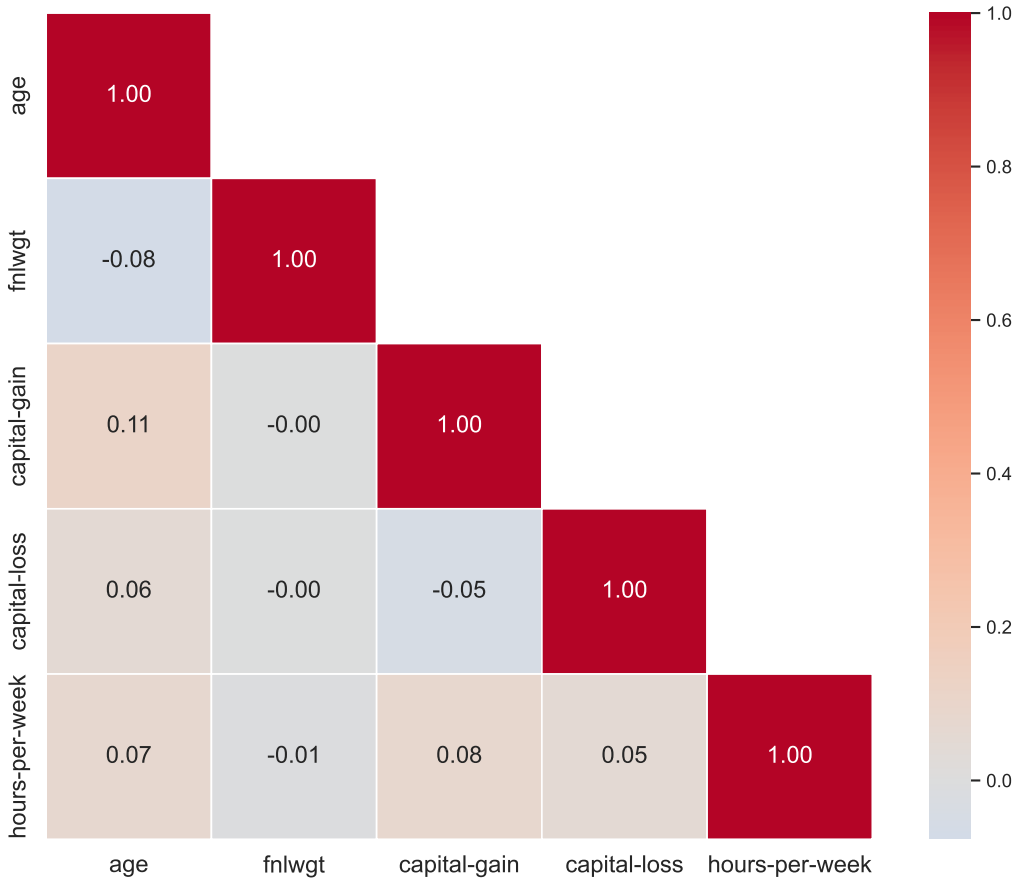

Figure 4: Pearson correlation between numerical variables

The Cramer's V correlation measure between categorical variables is represented by the heat map shown in Figure 5.

One of the important observation is the perfect correlation between the variables "education" and "education-num". This is due to the fact that the variables actually present the same values. For this reason, it was decided to exclude the variable "education-num" from the classification process.

The correlation between sex and relationship is 0.65, indicating a strong association. This suggests that there might be a pattern or trend in relationship status that differs between the

sexes.

Further analysis reveals a moderate correlation between "income" and variables such as education, at a level of 0.37, marital-status with correlation of 0.45, occupation type at 0.35, and relationship at 0.45. These correlations suggest that higher levels of education and specific occupations may contribute to higher income levels. Furthermore, marital status and relationship also appears to influence economic conditions and thus earnings.

The correlation of 0.22 between occupation and workclass indicates that certain jobs are more common in specific types of work.

Demographic variables such as race and native country have a correlation of 0.39, which is intuitive because country of origin is associated with race. These characteristics also show relatively low correlations with "income" (0.10 and 0.03 respectively). This suggests that such factors may not have a strong direct effect on financial status. However, the correlation may be distorted by the fact that the set is dominated by individuals from the US as well as the white race and the weak impact of the rest of the population on the analyses.

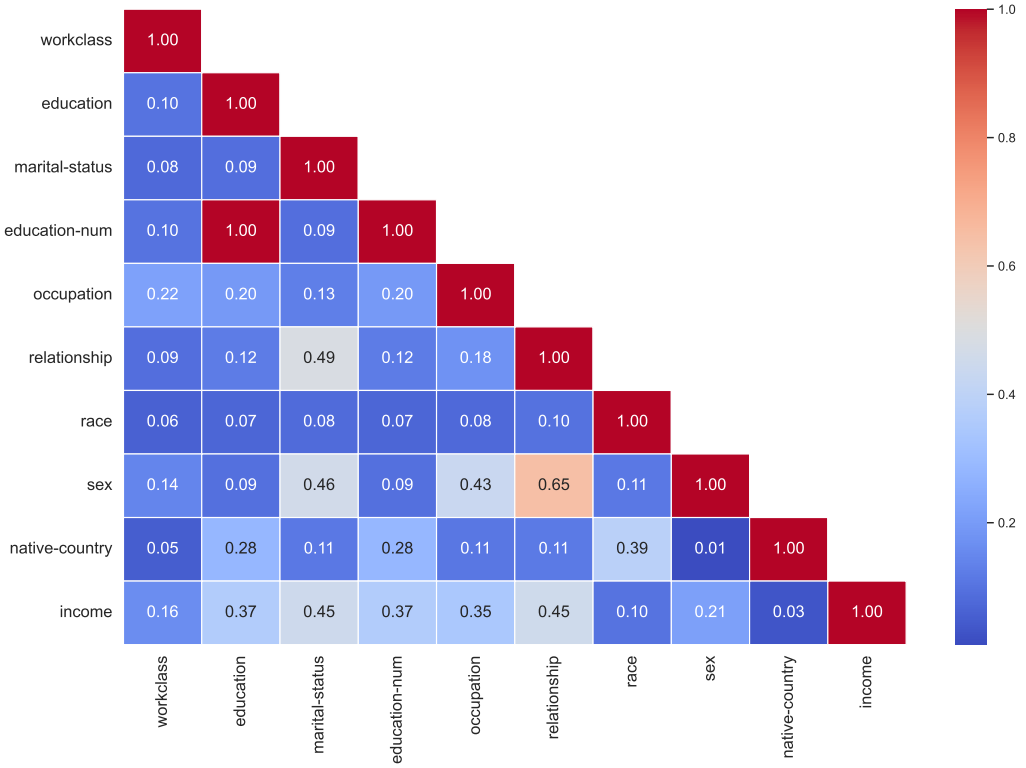

Figure 5: Cramér's V correlation between categorical variables

Boxplots, shown in Figure 6, were created for each numerical type attribute to emphasise the differences between the two income groups.

The upper left plot reveal a noticeable difference in age distribution between the two income groups. Individuals in the higher income category tend to be older, with a median age around 43 years, compared to the median age of approximately 35 years in the lower income group. This difference suggests that older age, which is often associated with more experience and a higher career position, may be a potential factor in higher income levels.

The values observed for fnlwgt between the two income groups overlap significantly, indicating that this variable is not a strong determinant of income levels. This is due to the fact that final weight is calculated from variables other than income.

In the next chart, it can be seen that higher levels of education are associated with higher incomes. The median level of education for those earning more than $50 000 is around 12 years compared to around 9 years for those earning less. It is interesting to note that at the 10 year

education level, there is a third quartile for those with incomes below $50K, while there is also a first quartile for the higher earning group. This highlights the importance of education in earnings.

Due to the dominance of capital gain and loss equal to 0, the values of the quartiles are equal to 0. For this reason, only the median and the distribution of outliers are visible in the graphs. However, it can be noted that higher income earners show a much wider range of capital gains, while the capital gains for the lower income group are lower and unlikely to exceed $10 000. Meanwhile, in the case of capital loss, there are more occurrences in the group earning less than $50K. While in the second group, losses are less frequent but relatively high.

In the case of working hours, in both groups the median is 40 hours per week which corresponds to the most common working model which is full-time. However, there is a wide variation in working hours in both groups, as can be seen from the outliers. Boxplots suggest that higher incomes may be associated with more working hours.

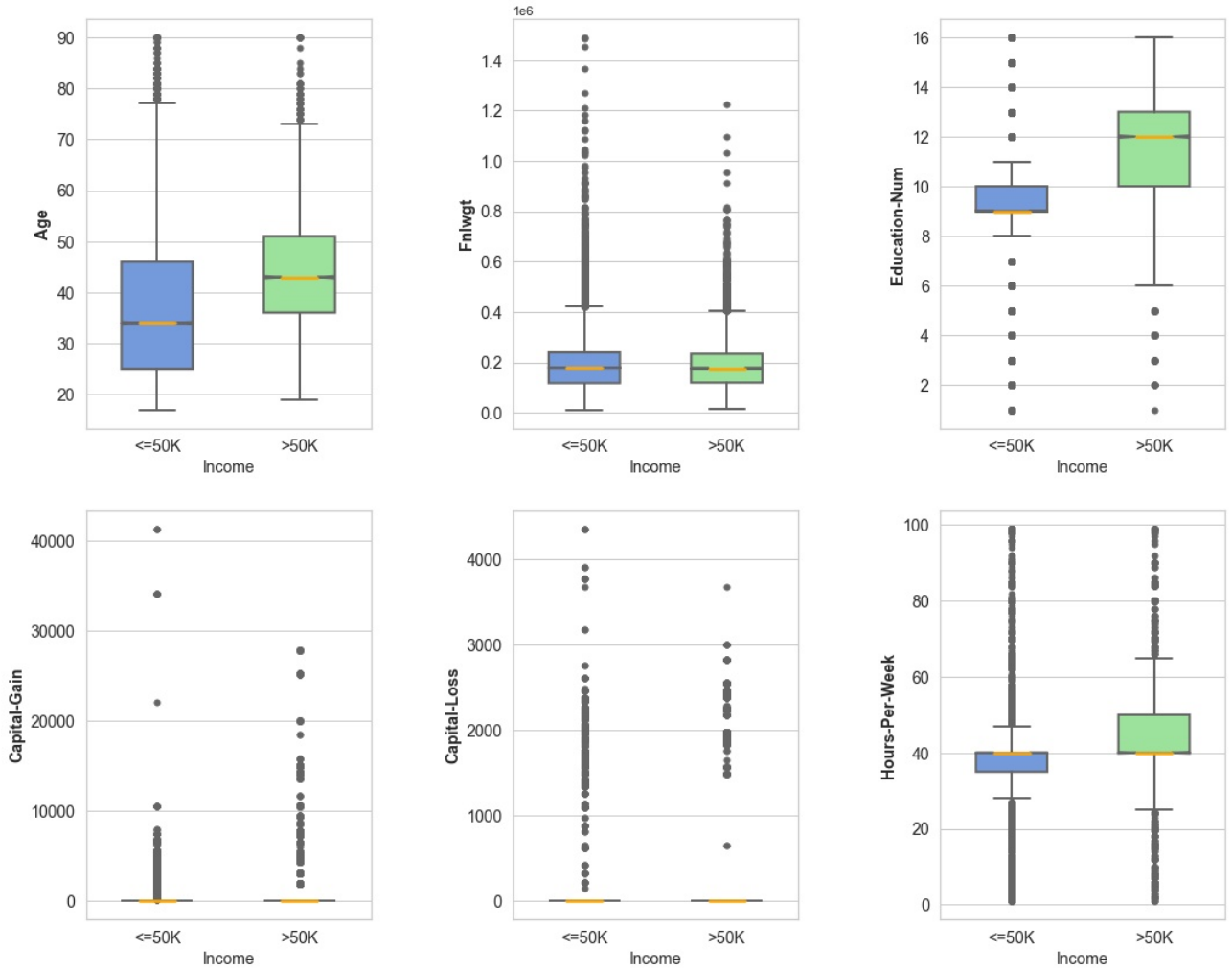

Figure 6: Distribution of age, final weight, education number, capital gain, capital loss, and hours per week across two income groups ($\leq$ $50K and $>$ $50K). Each plot highlights the median, interquartile range, and outliers, illustrating the differences in these variables between lower and higher income individuals.

The analysis of the proportions between the two income groups for each value of qualitative characteristics, as seen in Figure 7, provided interesting insights into the impact of factors on income.

The graph for the work class reveals that Self-employed with income individuals (Self-emp-inc) show a significantly higher probability of earning more than $50 000 per year, reflecting the

potential financial benefits of owning registered businesses. Other sectors show lower income potential and the percentage of people in the higher income group is less than 40%.

In the case of education level, there is clear trend where income tends to increase with higher levels of education. Most individuals with a university degree earn over $50K per year, particularly those with doctoral degrees, where nearly 80% fall into this income bracket. On the other hand, those who have completed only secondary or high school education usually earn less than $50K. This underscores the significant role of education in obtaining better-paying jobs.

It can also be seen that the married population has a significantly greater proportion of people who are in a higher income group than the other marital and relationship statuses. This may reflect total household income or the stability that marriage can provide.

For the occupation characteristic, a wide variety of financial conditions can be observed. Higher income levels are typically associated with positions in specialized fields, management, or military roles. Meanwhile, the smallest proportion of individuals earning over $50 000 annually is found in professions such as house Services, handlers-cleaners, and farmers.

Some racial differences in income are apparent. White and Asian/Pacific Islander groups have a higher proportion of high-income earners. However, it is important to remember that more than 85% of people in the dataset are white so the analysis may not correctly represent the financial situation of other races.

A significant income gap exists between genders. Men are more likely to earn above $50K compared to women, which may indicate gender inequality in earnings. This gap reflects the challenges of achieving gender parity in the workplace that occurred in the 1990s in many countries.

A plot comparing income levels by country of origin shows little difference between the United States and other countries. The proportions of the two income groups are quite similar, with a slight dominance of those earning over $50K in the USA

Figure 7: Proportion of individuals within two income brackets across various categorical variables, including work class, education, marital status, occupation, relationship, race, sex, and native country

The Figure 8 illustrates density plots that compare the distribution of various numerical attributes between two income groups: those earning less than $50 000 and those earning more than $50 000 per year.

A density plot for age reveals that those with incomes above $50 000 tend to be older, with the distribution peaking in their mid-40s. Whereas, a plot for the lower income group shows an skewed curve with a peak at around age 25.

The density curves for fnlwgt variable have a similar shape for both income groups, with a noticeable peak at lower weights and a rapid decline as the weight increases. The similarity in shape suggests that this variable does not differ significantly across income levels, indicating uniform representation in the census data regardless of income. The resemblance in shape suggests that this feature does not vary significantly by income level, indicating a uniform representation in the data regardless of income.

There is a difference in the distribution of years of education between the two income groups. Those earning more than $50K show a peak at higher education levels, particularly around 13 to 16 years, suggesting that a college degree is common. In contrast, the lower income group has a clear spike and dominant education values after around 9-10 years, indicating completion of secondary school.

The distribution of capital gains is highly different for the two groups. For those with lower incomes, capital gains are barely visible, with a density approaching the zero line. For individuals earning more than $50 000, the distribution peaks near gains of $0-$200, but it shows that those with higher incomes are more likely to have capital gains, reflecting greater involvement or access to profitable investments.

Both income groups show that most individuals experiencing no capital loss, with the plots peaking at zero. However, the distribution of the higher income group has longer arms and a sleeker shape, suggesting that these individuals might also engage more frequently in financial markets, where they are exposed to potential losses.

The hours per week plot indicates a concentration around the 40-hour for both groups, representing a standard full-time job. However, the distribution for those earning over $50 000 shows that individuals in this group are more likely to work both the typical 40 hours and longer hours, potentially contributing to their higher income.

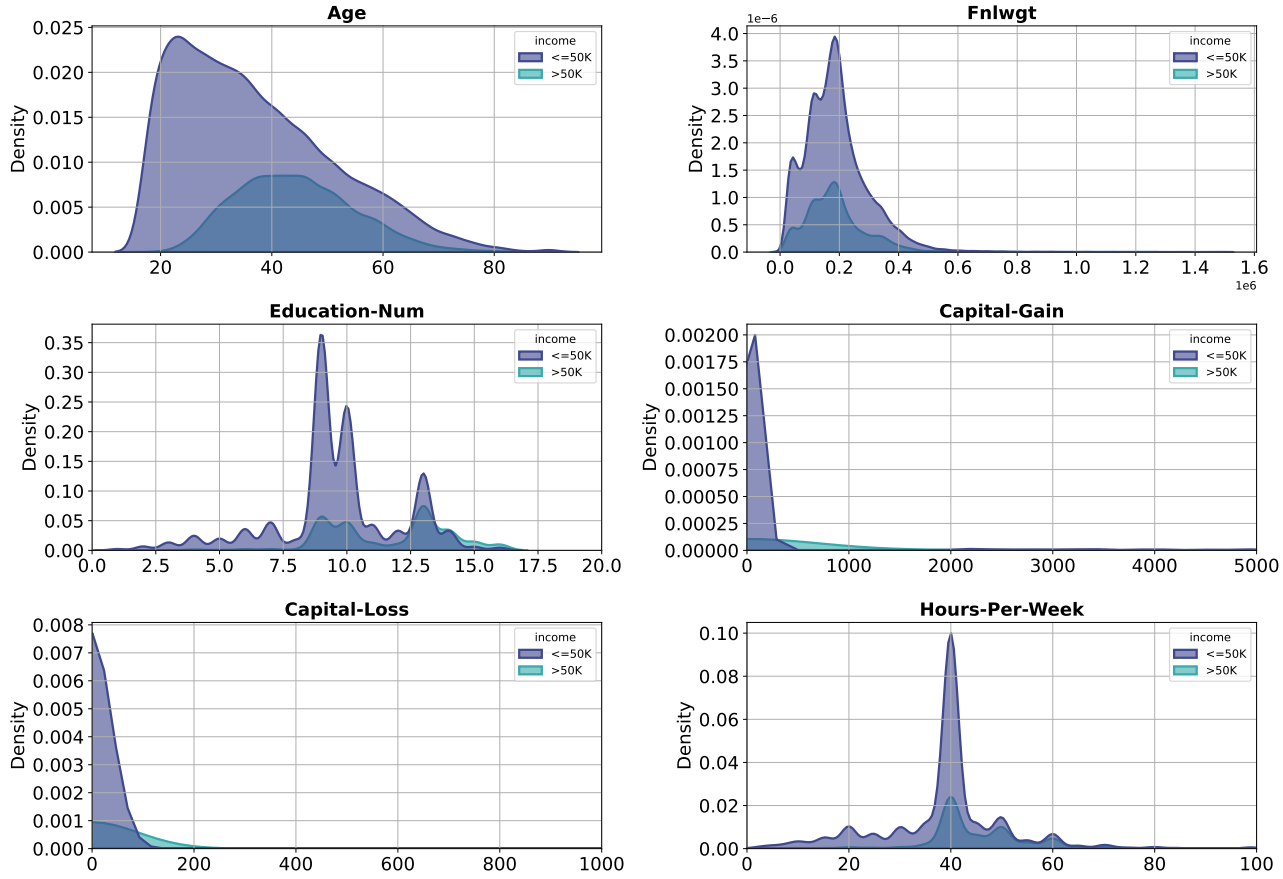

Figure 8: Density distributions of key variables such as Age, Final Weight (Fnlwgt), Number of Years of Education (Education-Num), Capital Gain, Capital Loss, and Hours Per Week, categorized by income levels (below and above $50K per year)

4.3 Hyperparameter Tuning and Classification Results for All Features

Optimal Hyperparameters

Hyperparameter optimization was conducted using `GridSearchCV` coupled with `RepeatedStratifiedKFold`. This approach ensures a thorough exploration of the hyperparameter space under a cross-validation scheme, configured with 10 splits and 3 repetitions. The aim was to ascertain the optimal set of hyperparameters for each classifier used, enhancing the model's accuracy and reliability. The results of this hyperparameter tuning process are presented in Table 7.

Table 7: Optimal Hyperparameters for Each Classifier

Classifier	Optimal Hyperparameters
LDA	`shrinkage: None`
QDA	`reg_param: 0.9`
Logistic Regression	`C: 10, max_iter: 100, penalty: l2, solver: sag`
Decision Tree	`criterion: gini, max_depth: 10, max_features: None, min_samples_leaf: 2, min_samples_split: 20`
K-Neighbors	`n_neighbors: 9, weights: uniform`
Random Forest	`criterion: entropy, max_depth: None, max_features: sqrt, min_samples_leaf: 2, min_samples_split: 10, n_estimators: 300`

Cross-Validation Results

Figure 9 provides a comparison of F1 scores achieved by various classification models, each optimized with the best set of hyperparameters as identified through our tuning process. The models evaluated include Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), Logistic Regression (LR), Decision Tree (DTR), K-Nearest Neighbors (KNN), and Random Forest Classifier (RFC). The F1 score was the measure on which the comparison was made. It was chosen because of the non-balanced nature of the labels in the analysed dataset.

The figure illustrates two scaling techniques: MinMax Scaling and Standard Scaling. Across most models, the choice of scaling technique does not significantly impact the performance. Notably, Standard Scaling appears to generally enhance the performance of QDA and KNN models, as indicated by higher median F1 scores compared to those obtained using MinMax Scaling. This suggests that Standard Scaling is better suited for our models and dataset and in the later analysis only it was used, MinMax Scaling method was discarded. All classifiers achieved comparable F1 score. RFC was the best with a median of 0.68, followed by QDA and DTR.

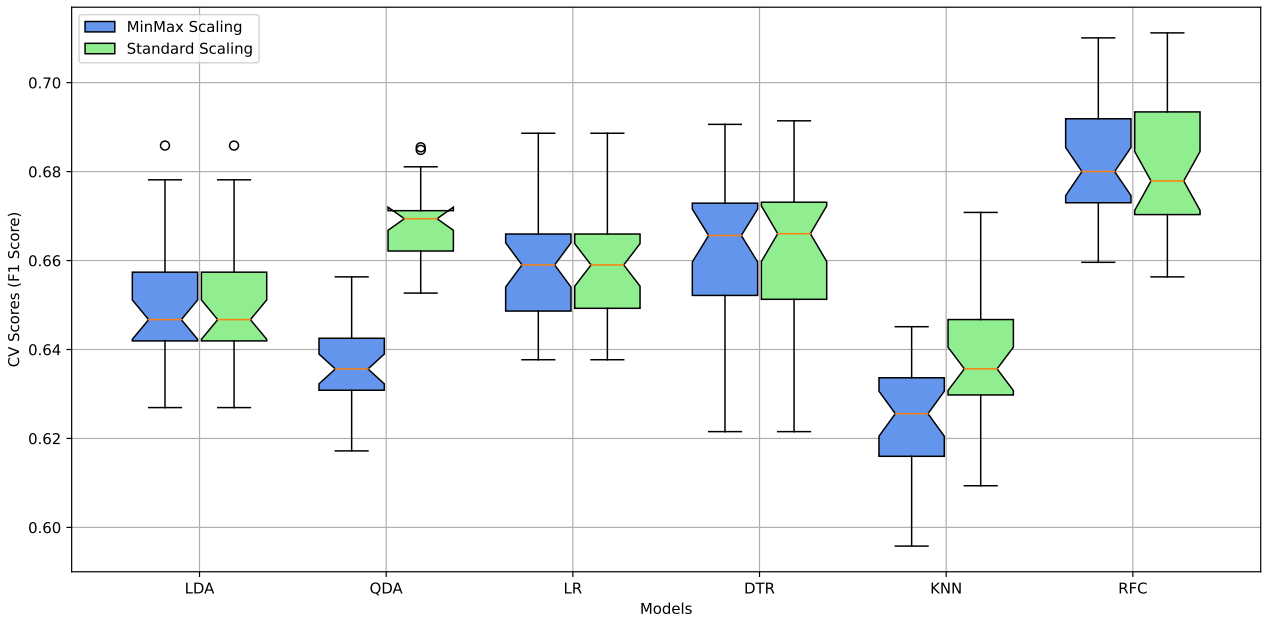

Figure 9: Cross-validation score comparison among different classification models, including Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), Logistic Regression (LR), Decision Tree (DTR), K-Nearest Neighbors (KNN), and Random Forest Classifier (RFC). The box plots illustrate the distribution of F1 scores achieved under two different scaling techniques: MinMax Scaling and Standard Scaling.

Results for Test Set

In the next step, all models were tested with the best hyperparameters and the standard scaling technique on the training and test data. We note that the test data were completely separated and did not participate in the cross-validation process, according to Figure 3. The results are presented in Figure 10, where we compare accuracy, precision, recall and F1 score, and in Figure 11, where we present the confusion matrices for the test set for all classifiers.

In general, the results obtained on the test set are similar to those obtained on the training set. Only for KNN and RFC the measures for the test set are lower by about 6-10 percentage points. The following description focuses on the results for the test set.

The evaluation revealed that most models achieved comparable outcomes, with accuracy rates generally ranging between 82% and 86%. The Random Forest Classifier exhibited the

highest accuracy, while the Quadratic Discriminant Analysis performed slightly less effectively (82%).

A notable variation was observed in the trade-off between precision and recall among the models. In terms of precision, RFC was the most precise with a remarkable 77%, suggesting its efficacy in minimizing false positives. Conversely, recall rates were generally consistent across the board, hovering around 60%, with QDA being a significant exception by achieving a notably higher recall of 76%, which could be crucial in scenarios where reducing false negatives is paramount.

The F1 scores, which balance precision and recall, varied from a low of 64% for the K-Nearest Neighbors to a high of 68% for RFC. Given these metrics, KNN could be considered the least effective model due to its lowest F1 score combined with suboptimal precision and recall rates. In choosing the most suitable model, RFC stands out as a robust option due to its superior F1 score. However, QDA might be preferred in applications where the minimization of false negatives is critical, despite its other limitations. In the following analyses, only these two classifiers were focused on, as the others obtained results similar to or worse than the RFCs.

The matrices shown in Figure 11 confirms the conclusions we have formulated. They show very well QDA's ability to minimise false negatives. It obtained 558 of them, while second was RFC with 880 cases.

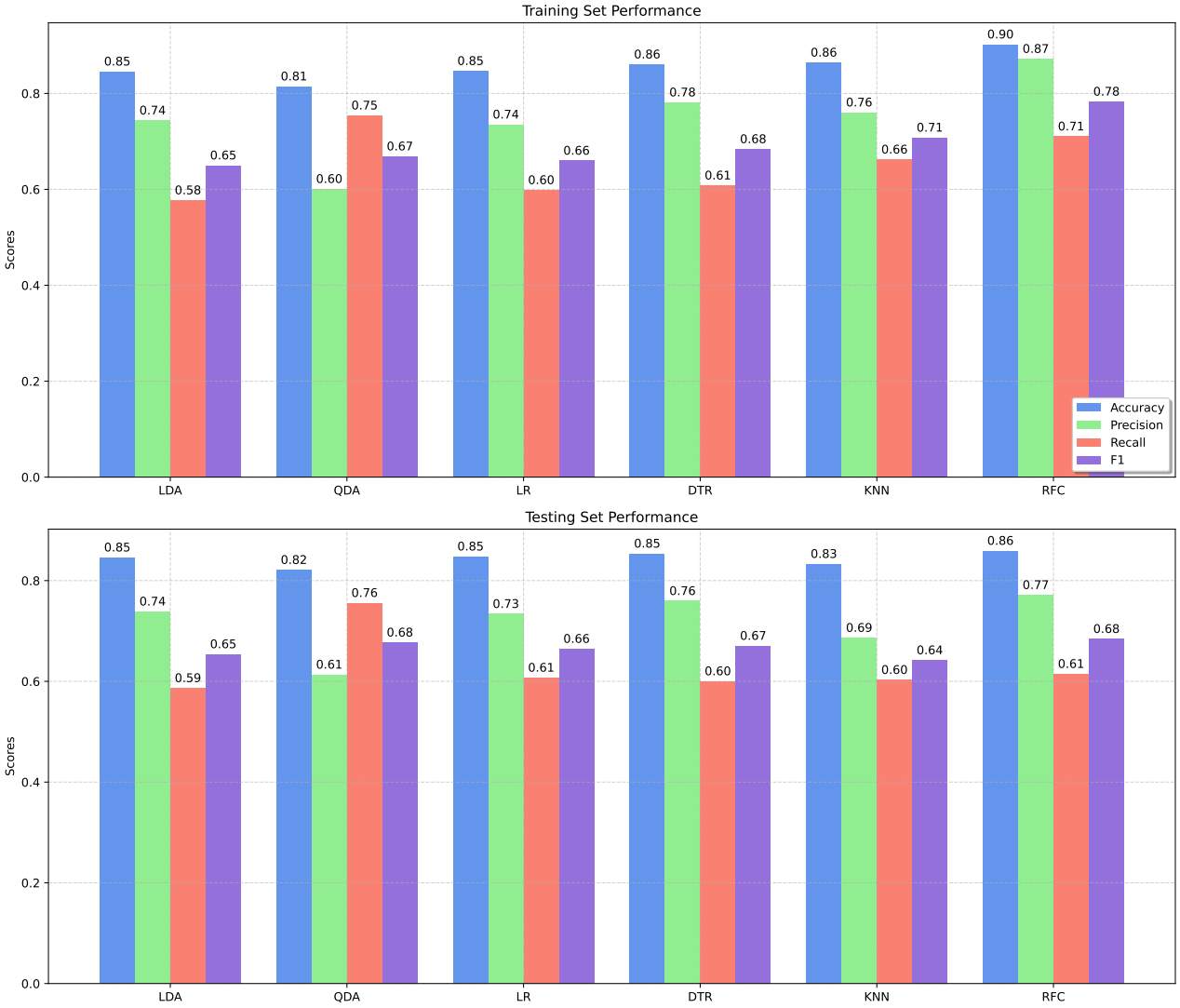

Figure 10: Training and Testing sets performance metrics for various classifiers including Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), Logistic Regression (LR), Decision Tree (DTR), K-Nearest Neighbors (KNN), and Random Forest Classifier (RFC). Each classifier's performance is evaluated based on four metrics: Accuracy, Precision, Recall, and F1 Score.

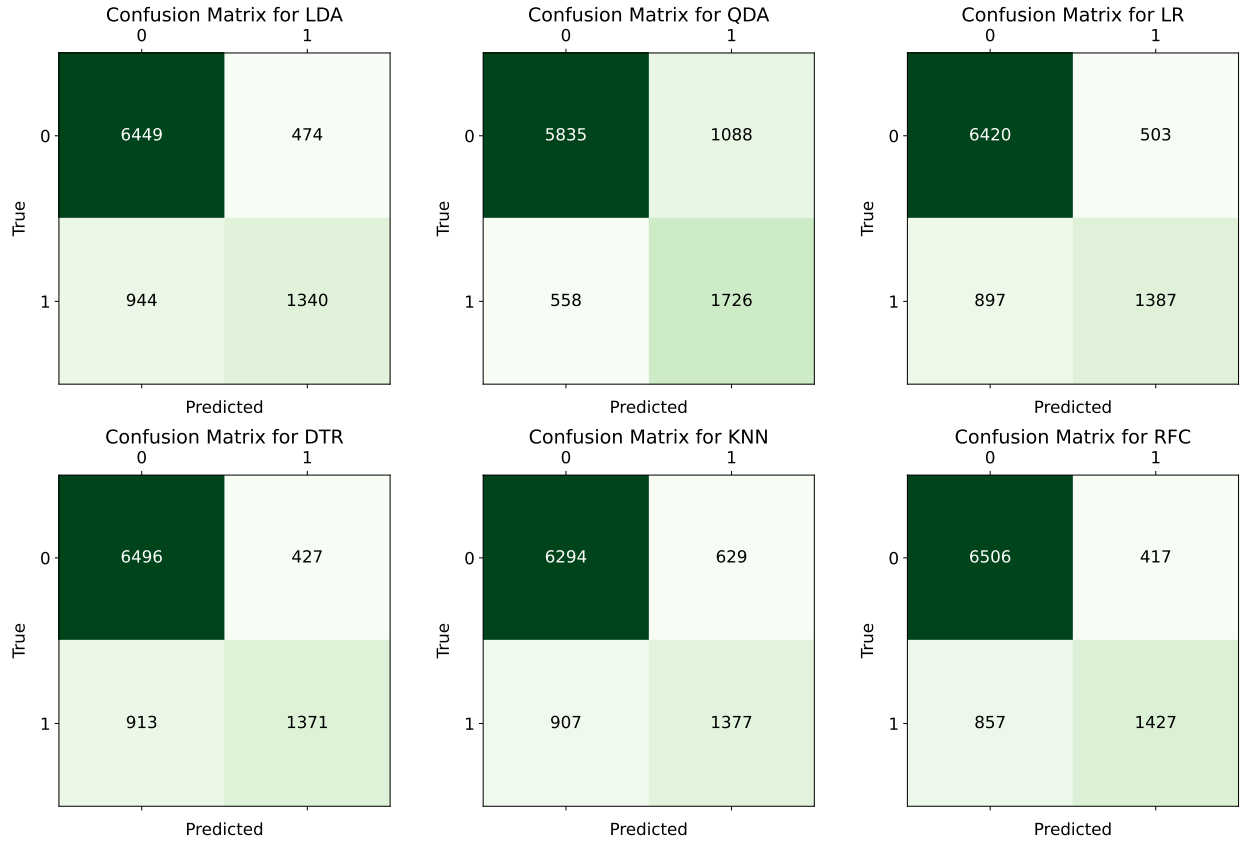

Figure 11: Confusion matrices for various classifiers including Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), Logistic Regression (LR), Decision Tree (DTR), K-Nearest Neighbors (KNN), and Random Forest Classifier (RFC)

4.4 Classification results for feature subsets

The subsequent phase of our analysis involved performing classification tasks on various subgroups of the feature set. To ascertain the importance and influence of each feature on the final classification outcome, the Information Value (IV) (see Section 3.3.5) measure was utilized. The initial findings, which identified the traits "relationship" and "marital-status" as the most influential, and "capital-loss" and "capital-gain" as the least, are documented in Figure 12.

Following this, new datasets were constructed incrementally, starting with the most influential feature as determined by IV. Each additional feature was included in the order of its determined importance. Using the `RepeatedStratifiedKFold` method, with 10 splits and 3 repetitions, we conducted cross-validation to measure the F1 score for each progressively expanded dataset. For instance, the initial dataset comprised solely the "relationship" trait, the subsequent dataset combined "relationship" and "marital-status", and so forth. Of course we point out that before the start of training, each categorical feature was converted into a dummy variable.

The comprehensive results from this incremental feature addition are illustrated in Figure 13. Notably, the bottom graph in Figure 13 excludes the first two columns of results to enhance the visibility of the remaining boxplots.

The analysis revealed specific insights into model performance:

- For the dataset containing only one feature, both the Random Forest Classifier and Quadratic Discriminant Analysis struggled with the classification task, unable to handle it effectively.

- With two features, only QDA managed to perform the classification, achieving an average F1 score of 60%.
- The performance of RFC consistently improved as more features were added, indicating a uniform enhancement in model efficacy.
- For QDA, performance increased with the inclusion of up to seven features; however, adding further features initially led to a decrease in performance, followed by a significant improvement after incorporating the final feature.
- Excluding the results obtained from the full feature set, QDA consistently outperformed RFC.
- None of the results obtained for the checked subgroups of features turned out to be better than the result obtained for Cross validation for the whole set of features.

The inclusion of the last feature, despite its lowest IV score, resulted in a substantial improvement in performance for both models. This suggests that the feature has a significant impact on the correct decision-making ability. Its low IV score might be attributed to the predominance of zero values, with only a minor subset exhibiting different values.

In addition, chi-square tests for qualitative variables and ANOVA tests for quantitative variables were carried out to test the statistical significance of Information Value. The pvalues obtained are shown in Table 8. For all features except "fnlwgt" the difference in income distribution is statistically significant. P-value doesn't give information about the strength of relationship and for this reason Cramer's V for the qualitative variables was also counted. The results obtained for it correlate with the values obtained for Information Value.

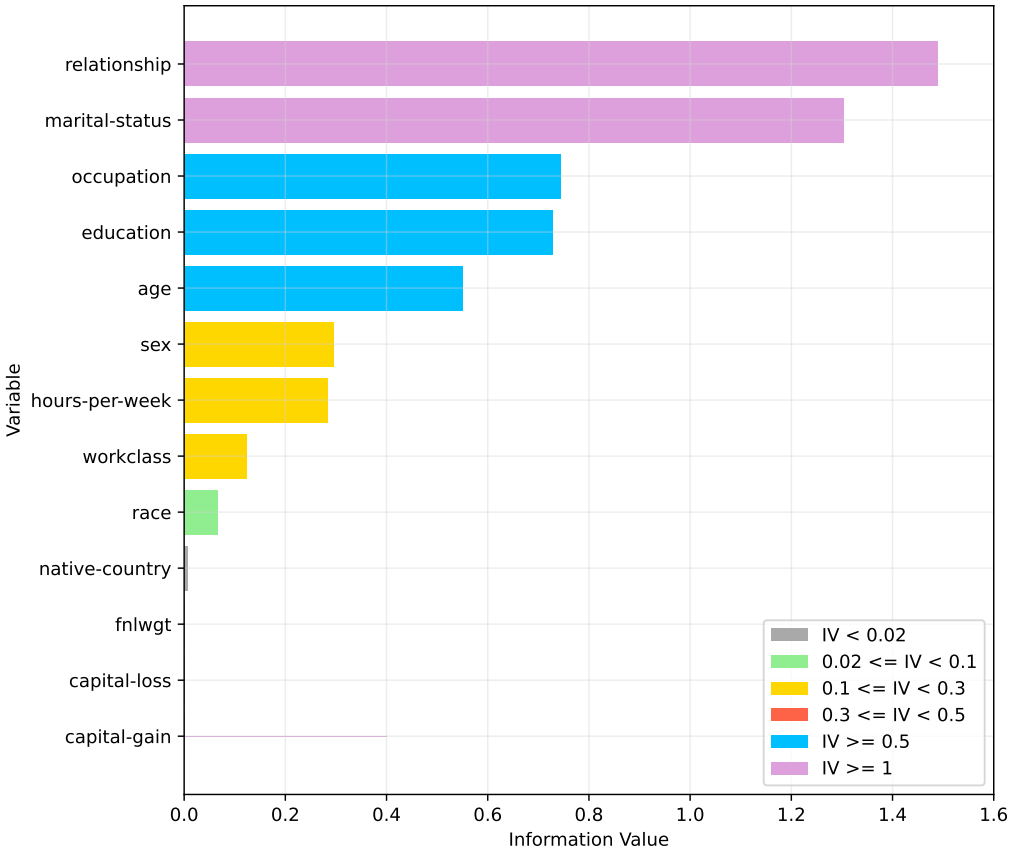

Figure 12: Information Value (IV) for various features within the dataset. Each bar represents a different feature, color-coded to indicate the range of IV.

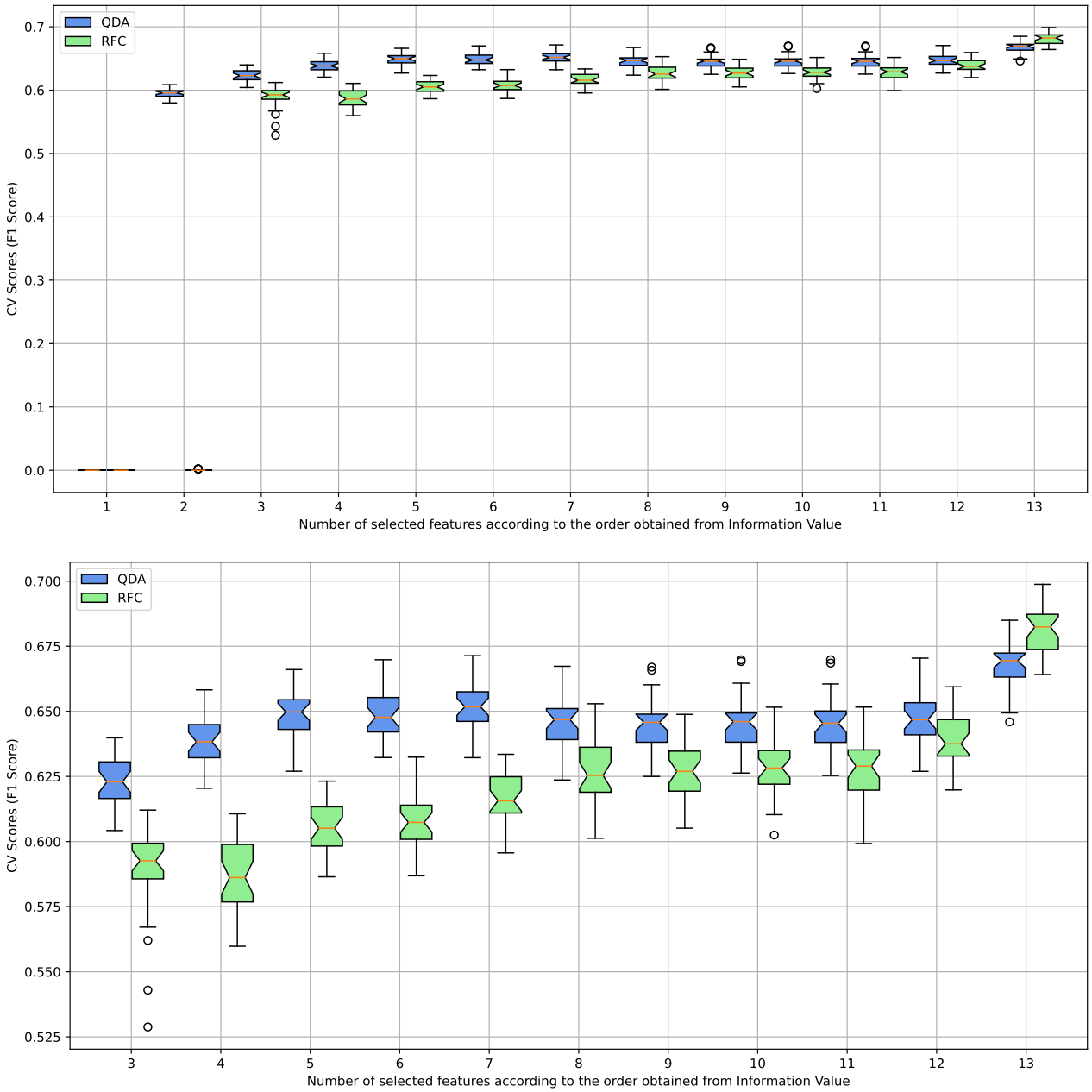

Figure 13: Performance evaluation of QDA and RFC models as more features are incrementally added based on their Information Value. Graphs illustrate the cross-validation F1 scores for both models as each new feature is included, starting from the single most influential feature up to all 13 features. The bottom graph provides a focused view, omitting the first two feature sets to highlight the changes in model performance from the third feature onward.

Table 8: Information Value, p-value and Cramer's V for various features. P-values were obtained from chi-square test for qualitative features and ANOVA test for quantitative features.

Feature	IV	P-Value	Cramer's V
Relationship	1.49	0.00	0.45
Marital-status	1.30	0.00	0.45
Occupation	0.74	0.00	0.34
Education	0.73	0.00	0.36
Age	0.55	0.00	-
Sex	0.30	0.00	0.22
Hours-per-week	0.28	0.00	-
Workclass	0.12	0.00	0.16
Race	0.07	0.00	0.10
Native-country	0.01	0.00	0.04
Fnlwgt	0.00	0.14	-
Capital-gain	0.00	0.00	-
Capital-loss	0.00	0.00	-

4.5 Classification Results for Cost Sensitive Learning

The graph presented in Figure 14 illustrates the distribution of F1 score, precision, and recall across various class weight scenarios, ranging from 1:1 to 1:100, as implemented in a classification model Random Forest in the `RepeatedStratifiedKFold` validation process with the number of splites 10 and three repetitions.

From the boxplots, it is evident that the precision metric generally decreases as the weight assigned to the minority class increases. This trend suggests that while the model becomes more adept at identifying the minority class (increasing recall), it does so at the expense of incorrectly predicting more negative instances as positive, which lowers its precision. This inverse relationship highlights the typical trade-off between recall and precision in scenarios where the class weight is adjusted to favor the minority class.

The F1 scores tend to peak in scenario CW 2, suggesting optimal balance point where the adjustments in class weights effectively enhance the model's ability to identify minority class instances without a substantial drop in precision. Beyond these points, the F1 score begins to decrease slightly, reflecting the growing impact of the compromise on precision despite gains in recall.

Figure 15 and Table 9 show the results on the test set for the best in terms of F1 score of the model created. They are the best results obtained by us.

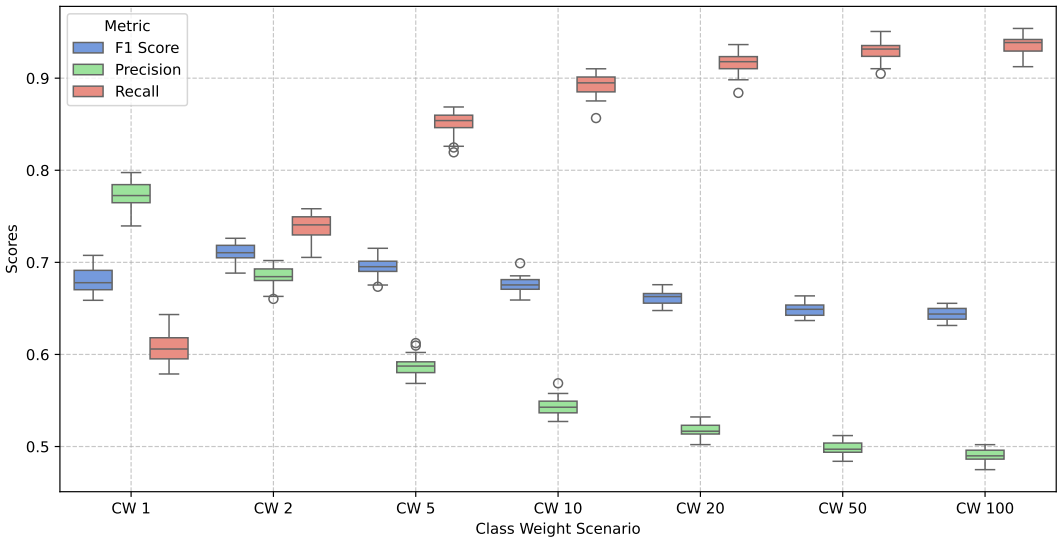

Figure 14: F1 score, precision, and recall across different class weight scenarios (class weights 1:1 to 1:100) for a classification model. Each boxplot shows the distribution of scores over multiple cross validation runs on training dataset/

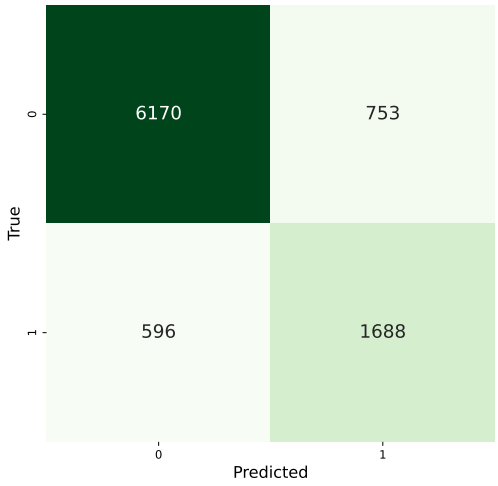

Figure 15: Confusion matrix visualization on the test set for the Random Forest model with adjusted hyperparameters, weights 1:2

Metric	Score
Accuracy	0.85
Precision	0.69
Recall	0.73
F1 Score	0.71

Table 9: Accuracy, precision, recall and F1 values calculated from confusion matrix presented on Figure 15

5 Conslusions

The detailed analysis of the Adults dataset reveals several significant insights that are pivotal for strategic development across various sectors, including policy-making, educational planning, and workforce development.

Firstly, the study conclusively shows that higher levels of education correlate strongly with higher income levels. This indicates that individuals with college degrees or higher education

tend to earn more than those with less education. This finding underscores the importance of investments in educational programs and scholarships, particularly those aimed at increasing access to higher education for economically disadvantaged groups.

Moreover, the analysis highlights that marital status, particularly being married, is associated with higher income levels. This correlation might reflect the economic benefits and stability brought about by dual-income advantages or other socio-economic factors linked to marital status. Programs that provide support for family structures or offer counseling that promotes long-term commitments could be beneficial.

The occupation of individuals also plays a critical role in determining income levels. Occupations in management, specialty fields, and tech are more likely to yield higher incomes compared to manual labor or service-oriented jobs.

Additionally, the feature selection process helped to select the most important features affecting whether or not a person earns more than $50k a year.

Lastly, the study uncovered a gender disparity in income, with men more likely to earn above $50K compared to women. This finding can catalyze initiatives aimed at achieving gender parity in pay.

The decision boundary we have created can identify constrictive cases and classify them into a specific class. The Random Forest model with optimal hyperparameters selected, weights set due to features and standard scaling process achieved an accuracy of 85% and an F1 score of 71% on the test set.

6 Further Research Suggestions

In the context of the current work more strategies can be proposed to extend the findings and enhance the model's performance and applicability.

An immediate extension could involve the detailed assessment of feature importance (`feature_importances_`) using the Random Forest model. By identifying which features most significantly influence the prediction outcomes, researchers can gain deeper understanding of the underlying patterns and relationships within the data.

Adjusting class weights is a straightforward method for implementing cost sensitive learning in Random Forest Classifier, other techniques, such as modifying the decision threshold in models like Quadratic Discriminant Analysis, can offer nuanced control over the trade-offs between different types of classification errors. For instance, lowering the threshold for predicting the minority class can increase the model's sensitivity.

References

[1] ANDERSON R. *The Credit Scoring Toolkit: Theory and Practice for Retail Credit Risk Management and Decision Automation.* Oxford University Press, 2007

[2] www.archive.ics.uci.edu/dataset/2/adult.

[3] BRUCE P., BRUCE A., GEDECK P. *Practical Statistics for Data Scientists: 50+ Essential Concepts Using R and Python, 2nd Edition*, II ed. Helion, 2021.

[4] SINGH M. Understanding Categorical Correlations with Chi-Square Test and Cramer's V. *Medium*, Jun 18, 2023.