

**Yanis Chowdhary**

500901377

Supervisor : PhD Ceni Babaoglu

November 10, 2025

**CIND 820 - D1H**

Initial Results



**Toronto  
Metropolitan  
University**

## Table of Contents

---

Introduction and Data Analysis Section	3
Data Analysis Section Continued	4
Dataset Preparation Section	5
Model Evaluation	6
Model Evaluation and Insights	7
Conclusion	8
References	9

## **Income Classification using Machine Learning and the UC Irvine Census Income Data Set**

### **Introduction:**

In this project I used the UC Irvine Census Income Dataset (Kohavi, 1996) and machine learning tools to predict whether an individual earns more than \$50000 per year based on many different demographic and occupational features. I used supervised learning techniques such as Logistic Regression, Decision Tree, and Random Forest to evaluate the performance of these models and identify specific trends resulting in an income level.

### **Data Analysis Section :**

The census income dataset has information from 1994 consists of 48,842 records with 15 attributes and an income level or target variable which is  $\leq 50K$  or  $> 50K$ .

### **Key Variables :**

Numerical variables such as age, education-num, capital-gain, capital-loss, hours-per-week.

Categorical Variables such as workclass, education, marital-status, occupation, relationship, race, sex, native-country.

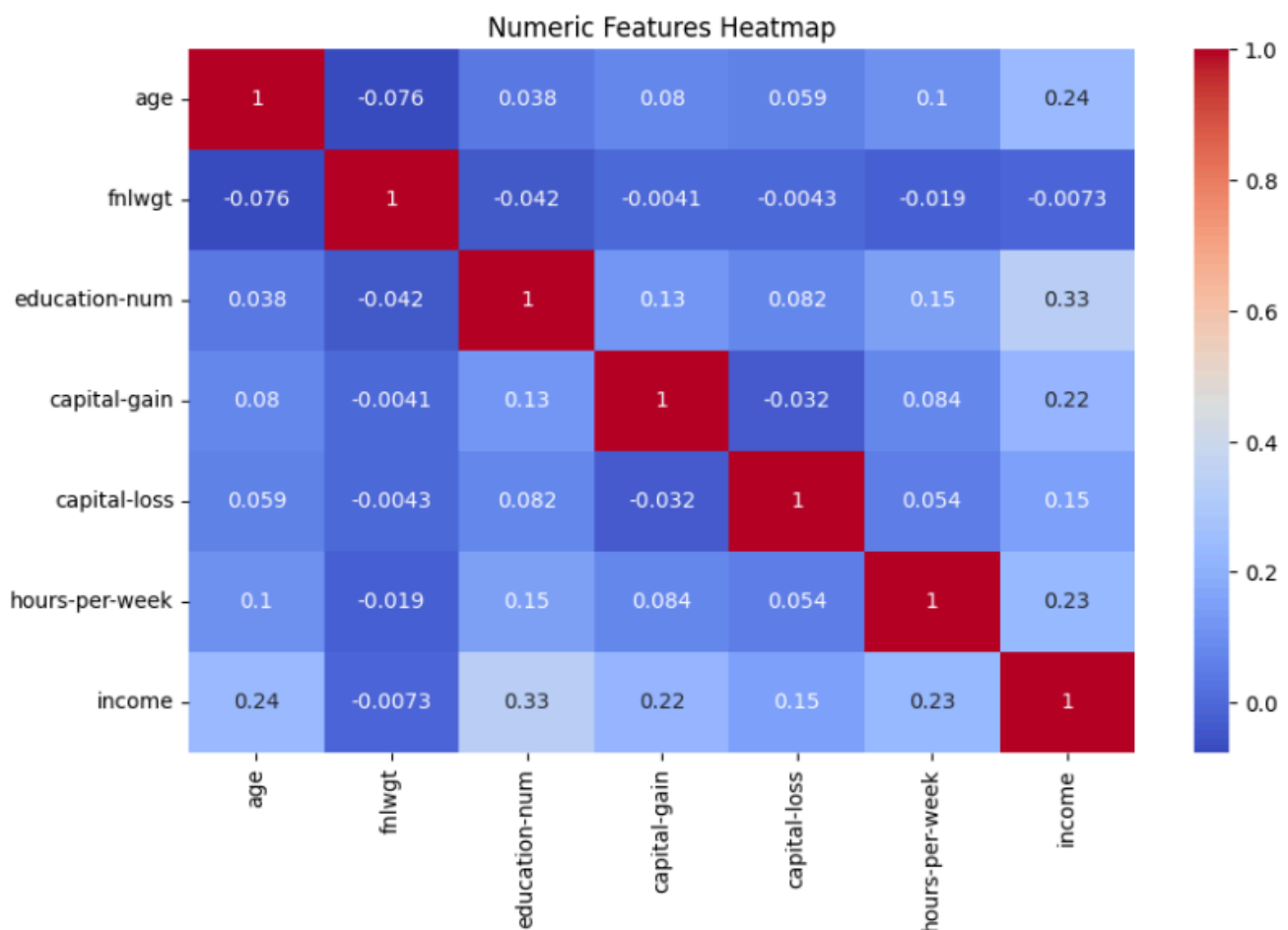
### **Descriptive Statistics:**

The exploratory analysis showed that the mean was 38 years old, with a standard deviation of 13 years. The majority of individuals in this dataset worked 40 hours per week. There was also a significant class imbalance where 34014  $\leq 50K$  whereas only 11208 earned  $> 50K$ .

### **Visual Patterns and Trends**

The income distribution showed significant class imbalance so this would have to be addressed and could be due to oversampling. I used the correlation heatmap to see if there were any strongly correlated features where I noticed education-num and education had a strong correlation along with hours-per-week and income. My main takeaways when initially looking at this dataset was that age, education and hours per week would determine if some made  $\leq 50K$  or  $> 50K$ . Some socioeconomic patterns I also noticed were that marital status plays a large factor

into predicting income as people raising a family are likely to have larger incomes to support their children or grandparents. Education and work hours also positively influence an individual's income as a person who has completed post-secondary education will have a degree and be able to find a job catering to that field. It was also evident that capital gain is a feature in this dataset and this indicated that people with investments are likely to earn above \$50000. These findings align with (Kotsiantis, 2007) who emphasized the role of education, occupation and marital status when determining income.



## **Data Preparation Section:**

### **Preprocessing steps:**

1. Data Cleaning: Replace or remove “?” entries to handle missing values. Also stripped whitespace, and removed punctuations. (Kohavi, 1996)
2. Encoding: Apply One-Hot Encoding for categorical variables to avoid ordinal bias and retain interpretability. (Quinlan, 1996)
3. Scaling: Standardized the numerical features using StandardScaler to normalize features, which was crucial for gradient-based algorithms. (Kotsantis, 2007)
4. Balancing: Apply SMOTE (Synthetic Minority Oversampling Technique) to oversample the minority class (>50K). (Chawla, 2002)
5. Splitting: Divide dataset into training (70%) and testing (30%) subsets used stratified sampling to preserve the class proportions. (Friendman, 2001)

These preprocessing steps allowed for consistency with all variables whether categorical or numerical, fair training of the linear models as Logistic Regression is sensitive to scale. Balanced learning towards the majority income class by removing bias through SMOTE. And compatibility and comparability between the three models.

## Model Evaluation

A supervised modelling approach was followed using Baseline Models such as Logistic Regression and Decision Trees. Logistic Regression provided interpretability for the baseline linear performance and the Decision Tree was able to handle non-linear relationships as well as categorical interactions.

I then moved onto a more advanced model being Random Forest, which is an ensemble of Decision Trees used to reduce overfitting and increase the overall predictiveness of a model.

All the models were trained on the scaled, SMOTE balanced data and evaluated using Accuracy Precision Recall and F1-Score

## Model Performance Summary

Model Comparison:				
	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.761701	0.513102	0.751338	0.609777
Random Forest	0.840127	0.678861	0.673409	0.676124
Decision Tree	0.795460	0.579690	0.635039	0.606104

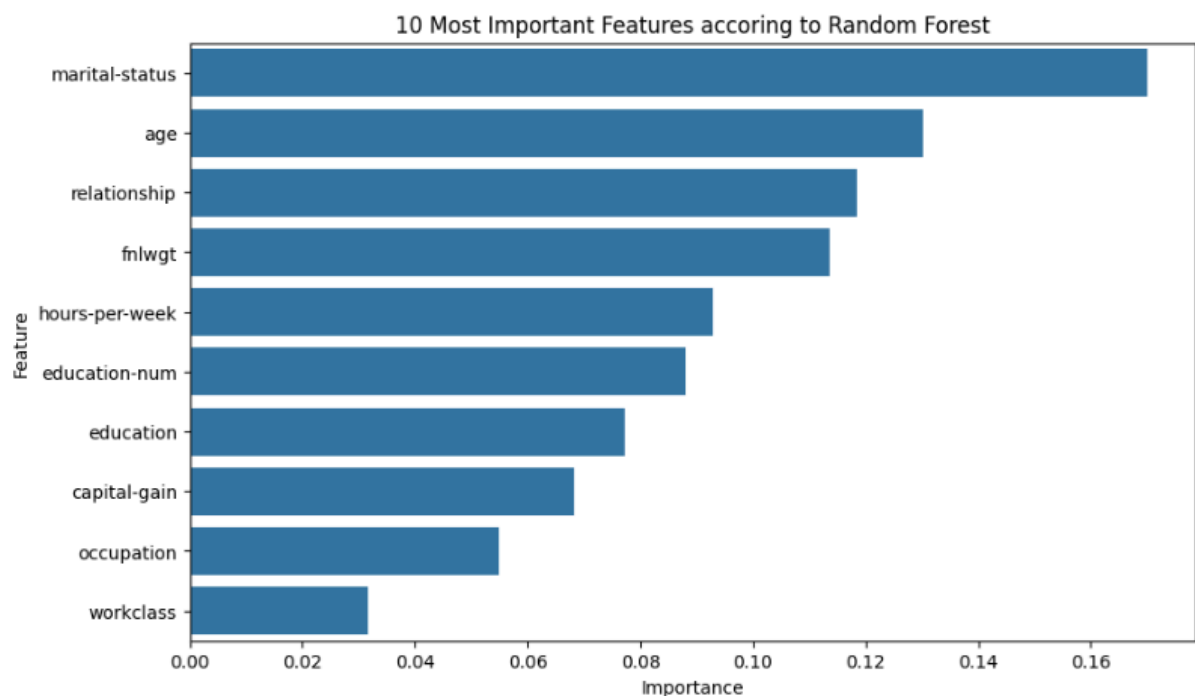
The Random Forest had the highest overall accuracy along with precision and F1-Score which showed that this ensemble method is able to interpret complex non-linear relationships

Logistic Regression had the highest recall indicating it is able to identify high-income individuals but also produced more false positives.

Decision Tree achieved average results, with being able to have marginally better accuracy than the Logistic Regression.

## Feature Importance Analysis

The Random Forest also identified the ten most influential predictors of income:



The feature analysis shows the strong predictive influence of demographic and socioeconomic attributes. Features such as marital-status, age, education level remain consistent even to this day with income prediction theories.

## Insights

Random Forest was superior to both base line models using this allowed for more accurate results of a more robust model. Balancing the dataset using the Synthetic Oversampling technique prevented bias towards the majority class. The importance of marital status, education, age and hours-per-week reflect the sociological and economic determinants of having a higher income.

**Conclusion:**

Through my learnings I was able to successfully apply and evaluate multiple different supervised learning techniques on the UC Irvine Census income dataset. I used proper data preprocessing which included encoding, scaling and class balancing to ensure accurate and interoperable results. My findings confirm that using ensemble methods such as Random Forest will outperform the baseline models especially when dealing with both numerical and categorical features.

## References

Kohavi, R. (1996). *Census Income [Dataset]*. UCI Machine Learning Repository. <https://doi.org/10.24432/C5GP7S>

Kohavi, R. (1996). Scaling up the accuracy of naive-Bayes classifiers: a decision-tree hybrid. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD)*, 202–207.

Dua, D., & Graff, C. (2019). *UCI Machine Learning Repository: Census Income Dataset*. University of California, Irvine. Retrieved from <https://archive.ics.uci.edu/dataset/20/census+income>

Kotsiantis, S. B. (2007). Supervised machine learning: A review of classification techniques. *Informatica*, 31(3), 249–268.

Quinlan, J. R. (1996). Improved use of continuous attributes in C4.5. *Journal of Artificial Intelligence Research*, 4, 77–90.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5), 1189–1232.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16, 321–357.