Income Class Predictive Model

Sayed Mohd Mahdi, Rachel Malzacher, Adam Randall

Purpose: Analyze demographic data from the 1994 Adult Income Census to predict if an individual's income exceeds $50,000 per year.

Objectives:

- Understand demographic factors influencing income levels.
- Build predictive models to classify individuals' income levels.
- Evaluate the performance of different models and identify the most effective one.

Applications:

- Market segmentation
- Public policy making
- Socio-economic studies

Dataset Description

There are 14 categories with 5 numeric and 9 nominal attributes with approximately 42k observations.

age	workclass	education-num	maritalstatus	occupation	relationship	race	sex	capital-gain	capital-los	hours-per-week	nativecountry	income
29	Private	13	Never-married	Prof-specialty	Not-in-family	Asian-Pac-Islander	Male	0	0	1	Japan	<=50K
74	Private	6	Divorced	Other-service	Not-in-family	White	Female	0	0	1	United-States	<=50K
39	Private	7	Divorced	Farming-fishing	Unmarried	White	Male	0	0	1	United-States	<=50K
57	Self-emp-not-inc	9	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	1	United-States	<=50K
27	Private	9	Never-married	Machine-op-inspct	Other-relative	White	Male	0	0	1	United-States	<=50K
22	Private	9	Never-married	Machine-op-inspct	Other-relative	Black	Male	0	0	1	United-States	<=50K
21	Private	9	Never-married	Machine-op-inspct	Own-child	Black	Male	0	0	1	United-States	<=50K
73	Self-emp-not-inc	10	Married-civ-spouse	Prof-specialty	Wife	White	Female	0	0	1	United-States	<=50K

Numeric / Continuous Variables	Categorical Variables
• Age (17-99 years) • Capital Gain ($0 - 99,999) • Capital Loss ($0 - 4,356) • Hours per week (1-99 hrs) • Education Number (duplicative)	• Income (binary) - 1, <= $50k) & 2, ($50K >) • Sex - 2 biological genders • WorkClass - 7 classes • Education Level - 16 levels • Marital Status - 7 statuses • Occupation - 14 categories • Relationship - 6 statuses • Race - 5 categories • Native Country - 41 countries

Sampleset Characteristics

```
 as.factor(nativecountry)
United-States:41292
Mexico       :  903
Philippines  :  283
Germany      :  193
Puerto-Rico  :  175
Canada       :  163
(Other)      : 2213
```

Some categories are more represented in the data and may skew the results.
Care needs to be taken when drawing conclusions based on these items.

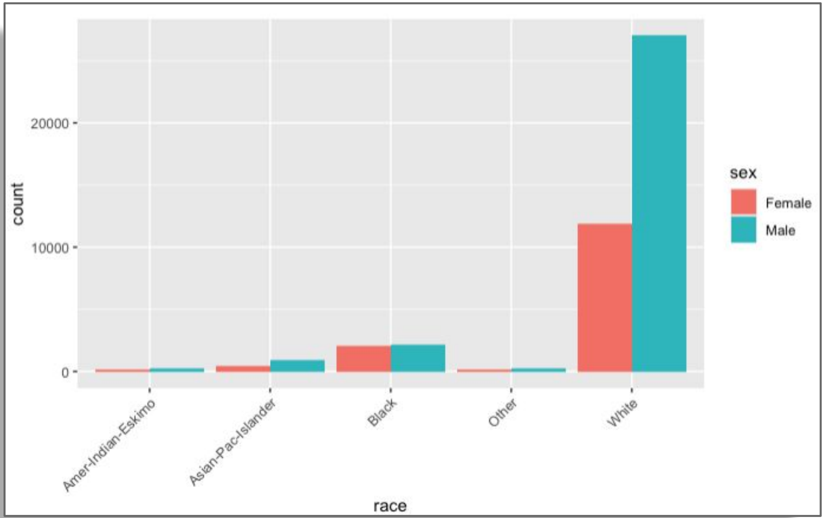

Figure 1a: Histogram of Race by Sex

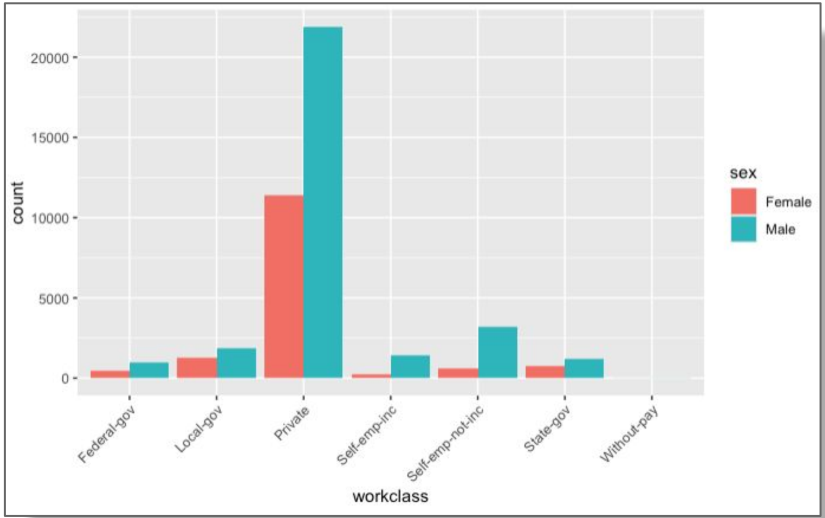

Figure 1b: Histogram of Workclass by Sex

Age & Gender Distribution

Age shows a similar representation over most of the range with less values at higher age groups.

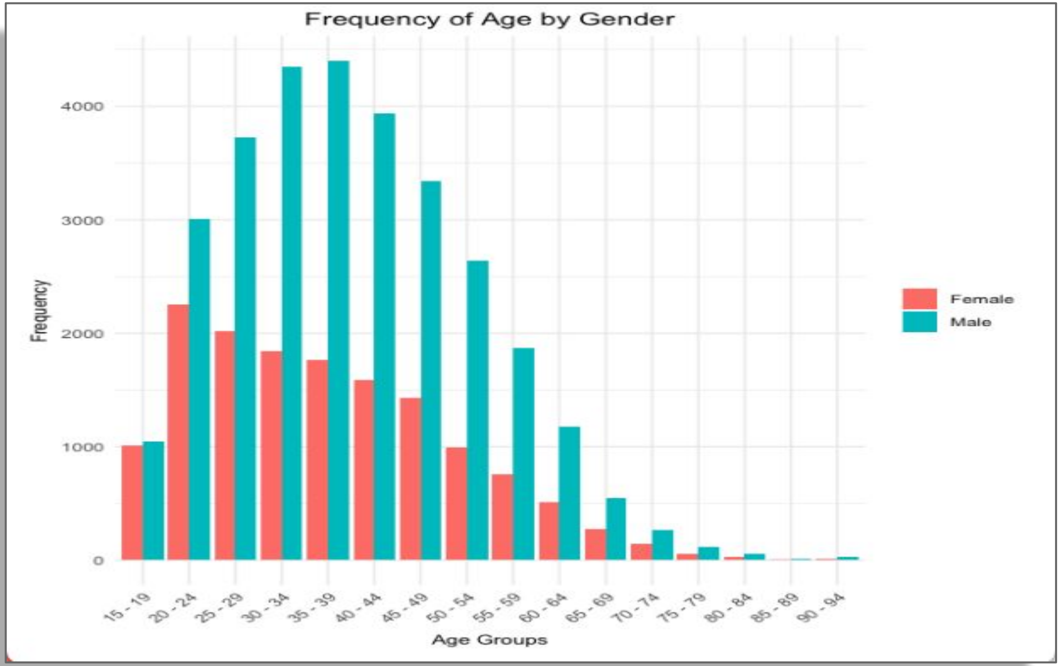

Figure 2: Histogram of Age groups by Sex

Age & Gender Distribution

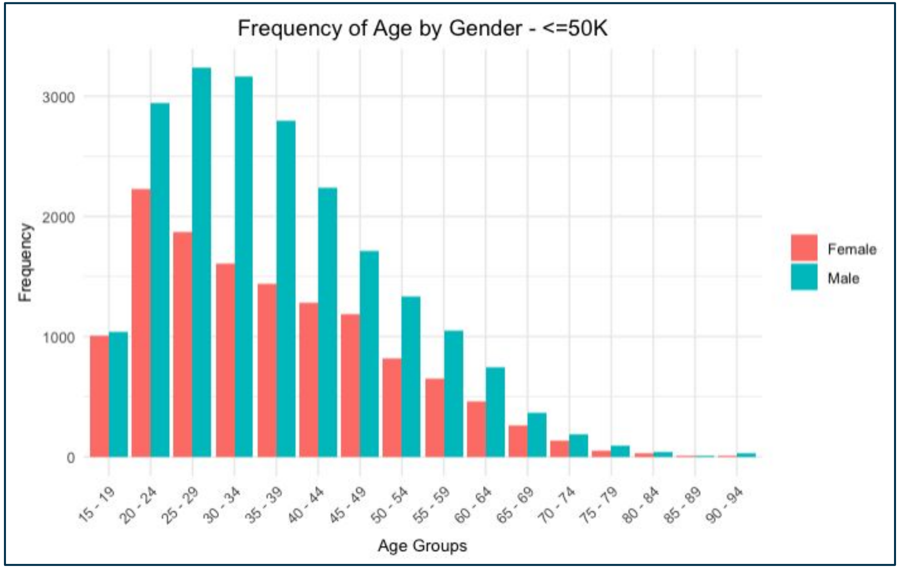

Figure 3a: Frequency of Age by Gender (Income: <=50k)

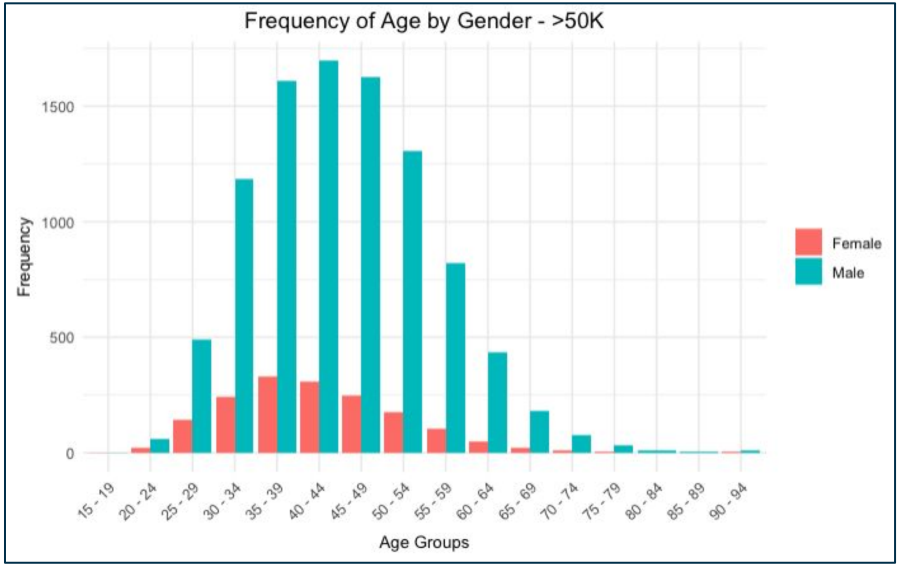

Figure 3b: Frequency of Age by Gender (Income: >50k)

Distributions of Numerical Data

Age is normally distributed with hours-per-week showing some levels stand out.

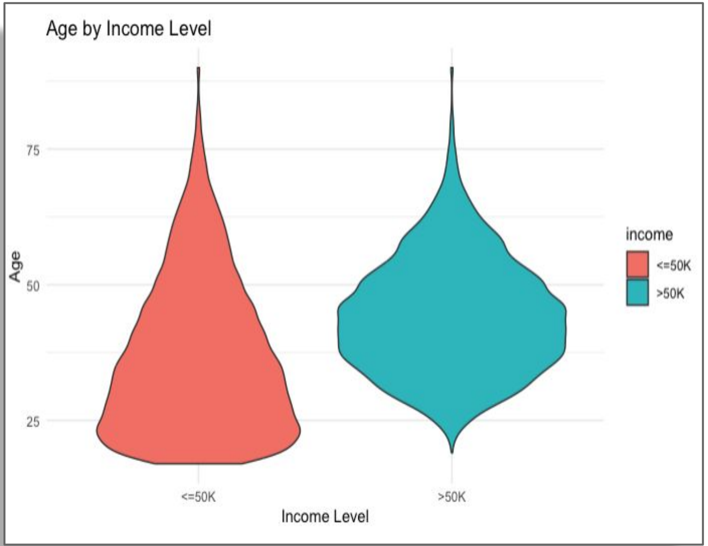

Figure 4a: Distribution of Age by Income level

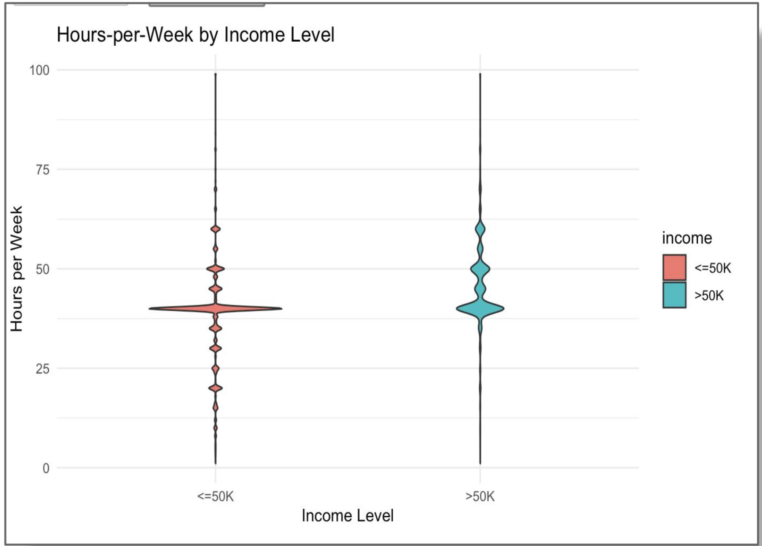

Figure 4b: Distribution of Hours-per-week by Income level

Statistical Techniques Applied

Results:-

1. T-test: There is a significant difference in age between individuals earning `<=50K` and those earning `>50K`.

t = -60.66	df = 25916	p-value < 2.2e-16

2. ANOVA (Education): Education levels significantly affect the number of hours worked per week.

F = 123.2	p-value < 2.2e-16

3. ANOVA (Marital Status): Marital status significantly affects the number of hours worked per week.

F = 538.8,	p-value < 2.2e-16

4. Two-way ANOVA: Both education and gender independently affect hours worked per week, and their interaction also has a significant effect.

Education:	F = 129.829	p-value < 2e-16
Gender:	F = 2573.794	p-value < 2e-16
Interaction (Education * Gender):	F = 5.501	p-value < 2e-16

5. Chi-square Test: There is a significant association between gender and income, indicating that income levels are not independent of gender.

X-squared= 2248.8	df = 1	p-value < 2.2e-16

Correlation Matrix

This correlation plot visually represents the strength and direction of similarities between numerical predictors and the response variable, Income. Both binary factors of Income and Sex are treated as numerical in this test.

- Income shows the most positive correlations with Education, Age, Hours-per-Week, and Capital Gain. Although all correlations are weak, <0.35.
- Nearly all coefficients within the matrix are considered highly significant.

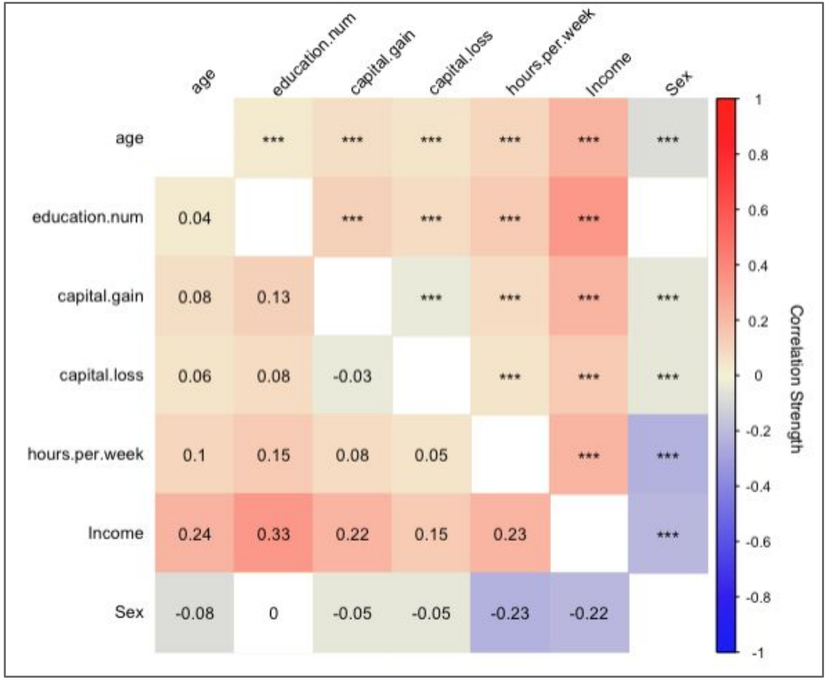

Figure ##: Correlation Matrix with Confidence Levels

Classification Analysis

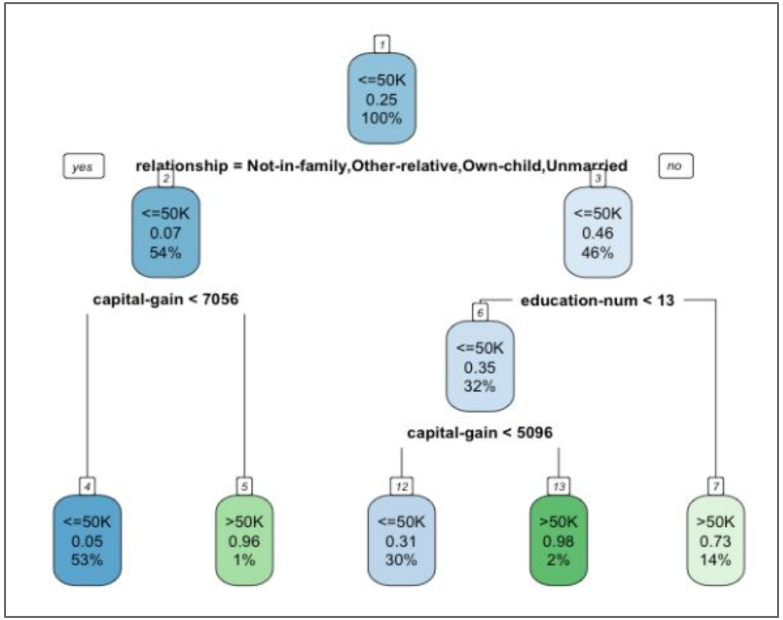

Figure ##: Decision Tree with Factored & Numerical Variables

Confusion Matrix Results			
	Decision Tree	**Random Forest**	**Logistic Regression**
Accuracy	85.02%	86.4%	84.18%
Sensitivity	.863	.932	.924
Specificity	.783	.653	.592
Pos. Pred Value	**.953**	.893	.873
Neg. Pred Value	.528	**.754**	.719
‘Positive Class’	“<= 50k”	“<= 50k”	“<= 50k”

CONCLUSION

This project successfully analyzed demographic factors influencing income levels using the 1994 Adult Income Census database, revealing that age, education, hours per week, and capital gain are significant predictors of higher income. Both decision tree and random forest models performed well, with accuracies of 85.02% and 86.4% respectively, demonstrating strong predictive capabilities. Statistical tests, including ANOVA, T-tests, and Chi-square, confirmed significant differences and relationships between income and various predictors. The correlation matrix underscored the significance of these predictors.Overall, the analysis provides valuable socio-economic insights and demonstrates the efficacy of machine learning models in income prediction.

Acknowledgements:

Thank you to Professor Mondal and Naveen Ramachandra Reddy

[1] Ronny Kohavi and Barry Becker.(1996) Data Mining and Visualization. *Retrieved from* https://www.kaggle.com/datasets/wenruliu/adult-income-dataset/data

Introduction – Original

The objective of this project is to:

- Understand the demographic factors that influence income levels.
- Build predictive models to classify individuals income levels based on their demographic information.
- Evaluate the performance of different models and identify the most effective one.

The purpose of this project is to analyze demographic data from the 1994 Adult Income Census database [1] to predict whether an individual's income exceeds $50,000 per year. Understanding these patterns can help in various applications such as market segmentation, public policy making, and socio-economic studies. This analysis aims to identify key factors influencing income levels and develop predictive models to classify individuals based on their demographic information.

https://www.kaggle.com/datasets/wenruliu/adult-income-dataset/data

Introduction

- Income predictive models provide market segment information
    - Increase value by targeting pricing models
- Drive public policy decision making based on data
- Provide projections based on history

https://www.kaggle.com/datasets/wenruliu/adult-income-dataset/data