CSCI 5502: Final Project Report

Aaptha Boggaram
aaptha.boggaram@colorado.edu
University of Colorado Boulder
Boulder, USA

Arindrajit Paul
arindrajit.paul@colorado.edu
University of Colorado Boulder
Boulder, USA

Nikola Tanovic
nikola.tanovic@colorado.edu
University of Colorado Boulder
Boulder, USA

1 INTRODUCTION

Many individuals these days are attempting to find out how they may earn enough money to live a respectable life, if not more. There is a common consensus in both the past and now regarding how to raise your income above the average in order to live on this ever-changing planet.

Typically, it is said that you must have a higher education degree from a well-known university, and that degree must be in a field that is currently prominent on the market. It is also believed that a specific length of time and experience in the industry in which you work is required to get a level of income that is considered above average. There are also some claims that it is always preferable to create a private business in order to earn a bigger salary than to work for someone else. Finally, some claim that an employee's gender or ethnicity has a significant influence on their pay.

We use the "Census Income" dataset [1] from UC Irvine's (UCI) Machine Learning Repository in this study. We are attempting to get knowledge regarding which method was employed to offer the greatest probability of having income above average by mining data from the 1996 census in the United States. With this approach, we can decide how accurate the claims above are. Also, by comparing the trends extracted from this dataset to the data summary from 2022, it is possible to determine how and if those patterns have evolved over the course of 25 years. By analyzing these patterns, we can forecast how they will change in the future and, as a result, give statistics on the best methods to earn above-average incomes in the future. For example, if a high school senior student is unsure about which major to pursue to earn a higher income without pursuing graduate studies after graduation, knowledge mined from this dataset could provide him with information about the occupation with the highest number of above-average income while holding a bachelor's degree.

Furthermore, if the knowledge extracted from this dataset reveals information about the amount of income that is influenced by characteristics such as gender or ethnicity, the government might take that into account and strive to make the situation better and fairer. For example, if the information extracted through mining reveals that being male greatly boosts the likelihood of earning more than $50,000, in addition to possessing other characteristics similar to those of women, it should be taken into account by the entire population.

To summarize, the major goal of this research is to discover previous income patterns and then compare them with a summary of current income data in order to gain the greatest information about the traits a person has to have in order to have greater chances for a higher income.

From the perspective of the authors, this knowledge can be very helpful since it could provide important information about possible career paths they could choose in order to earn a higher income, given their education level, age, and other traits.

Stated that, the following is the structure of this document. First, we will examine a summary of techniques employed with UCI's "Census Income" dataset. Following that, we will look at numerous research studies that used this dataset and the insights they derived from it. Next, we will elucidate the exploratory data analysis and modeling techniques used on the census income dataset. Finally, we will mention the results and conclude the project discussion.

2 LITERATURE SURVEY

UC Irvine's (UCI) "Census Income" [1] dataset was officially made open-source in 1996. This was 27 years ago. Since then, much work has been done using this dataset. Below is a description of work done using the census income dataset, as well as a few research articles that used this dataset.

2.1 Summary of Present work

Presently, the census income dataset has about 14 attributes and approximately 48842 instances. With this amount of data, a lot useful information has been mined. A few of them are listed below:

- Correlation between features.
- Visualization of trends in certain features.
- Identification of patterns between some features and income.
- Classification tasks for finding income with a set of features, and so on...

There is indeed a lot of work in the data mining aspect of this problem. But the actual challenge is incorporating data mining, analysis and predictive analysis of the data to gain more detailed patterns/trends. That said, several authors have developed very useful algorithms and tested them with the census income dataset. A brief description of insights found from their work is given in the following subsections.

2.2 Article 1

Wexler et al. [2] explored how interactive data visualization tools help people understand and analyze datasets and machine learning models better. In fact, they developed an interactive tool called the "What-If" tool. This application is a simulation platform that displays several exploratory data analysis methods, such as:

- Confusion matrices with different feature combinations
- Histograms, bar charts, and column charts
- Scatter plots of two-dimensional data
- Basic predictive models using pairs of features

While their study was confined to developing this tool, they tested and analyzed its outcomes using UCI's census income dataset. Without a doubt, the What-If tool performs well in small data mining/predictive tasks, but fails in larger data mining scenarios that demand analysis with more than a pair of features.

2.3 Article 2

Horesh et al.'s [3] work mainly focused on fairness regularization in machine learning using a paired-consistency approach. The paired-consistency approach tests how close the predicted output is to several protected pairs of information, such as age, race, and gender in terms of fairness among the classes.

Moreover, they used UCI's census income dataset to test their framework. There wasn't any explicit data mining that was performed on this dataset. However, feature importance, correlation of features, and several small data analysis techniques were used and plotted. In brief, their work was mainly focused on creating and testing an algorithm they developed on UCI's census income dataset with little emphasis on data analysis.

2.4 Article 3

Chung et al. [4] introduced a data-slicing technique for large quantities of data that helps users identifying the subsets of data where their predictive models might perform poorly. Further, this technique is a combination of three main techniques. It includes: Decision trees, clustering, and Lattice search algorithms. Incidentally, they too used UCI's census income dataset for testing their data-slicing technique. And yet, data mining wasn't carried out using this dataset.

While this is true, they did mention preprocessing and how they cleaned their data. In summary, they performed basic data cleaning, preprocessing, and creating an algorithm that suits their research objective, i.e., helping users retrieve data where their models might fall short in terms of performance.

2.5 Updated Literature Survey

This segment encompasses an updated literature review within the specified domain of investigation. Initially, Mohammed Temraz [5] employed artificial neural networks, decision trees, and the random forest algorithm in his research. However, it is imperative to note that the effectiveness of these models can be improved. On a parallel note, Jinglin Wang [6] and Rehman et al. [7] adopted diverse classical supervised and unsupervised learning methodologies for income classification. Despite the prevalence of these models in contemporary research, critical scrutiny reveals potential areas for enhancement, particularly in the optimization of performance metrics. The existing literature suggests that current models may fall short of achieving the desired level of efficacy.

Notwithstanding, in the study conducted by Sharath et al. [8], visually compelling and easily interpretable graphs were employed for data mining tasks related to income classification. Nevertheless, it is noteworthy that the absence of essential features in their dataset hindered the execution of predictive analysis and thorough critical examination of the data. As a consequence, the study's analytical depth and predictive capabilities were compromised, accentuating the significance of comprehensive feature inclusion in datasets for robust analysis.

Furthermore, these authors utilized a variety of metrics to measure the efficacy of their predictive analysis models. They are listed in table 1.

Table 1: Metrics used for evaluation of model performances

Metrics	Count
Accuracy	2
Precision	2
F1 - Score	1
Support	1
Recall	2
Gini Importance	1
Error	1
Other	1

The aforementioned paragraphs explain that there is precedent research conducted in this domain. It is noteworthy that various models have been applied in this context. However, this study aims to differentiate itself by undertaking a comparative analysis of less widely utilized models, thereby contributing novel insights to the field. Table 2 shows the models used for income prediction and the number of times each model has been used in other people's work.

Table 2: Predictive Models Used in Previous Work

Model	# of uses
Logistic Regression	12
Random Forest	7
Decision Tree	5
K Neighbors Classifier	5
Support Vector Machine	3
Gaussian NB	2
Boosted Gradient Descent	1
Linear Discriminant Analysis	1
Stacked Model	1
Bernoulli NB	1
K Folds	1
XG Boost	1
Gradient Boosting	1
Naive Bayes	1

This tabular data is used for this work to compare the models which have been extensively used with those that are not. And exploratory data analysis is performed on the dataset to compare the census insights of the past with those of the present.

3 PROPOSED WORK

Figure 1 depicts the workflow of the project. We obtain the dataset titled "Census Income" from the UCI Machine Learning Repository during the data acquisition stage. The dataset, which has the file format comma-separated value (CSV), has 15 features and more than 48,000 instances. The descriptions of the features are detailed in Table 3.

Table 3: Description of Data Attributes

Attribute Name	Data Type	Description
age	Continuous	Age
workclass	Categorical	Type of work class
fnlwgt	Continuous	Final weight
education	Categorical	Education level
education.num	Continuous	Encoded education level
marital.status	Categorical	Marital status
occupation	Categorical	Occupation
relationship	Categorical	Relationship status
race	Categorical	Race of the individual
sex	Categorical	Gender
capital.gain	Continuous	Amount gained in capital
capital.loss	Continuous	Amount lost in capital
hours.per.week	Continuous	Number of hours worked per week
native.country	Categorical	Country of birth
income	Categorical	Annual income

To clean and prepare the data for analysis and prediction, several preprocessing techniques are then used in the following stage. Preprocessing techniques to be used include denoising, imputation, feature selection, normalization, and transformation, among others. Following the data preprocessing step, we move on to exploratory data analysis (EDA), where we create data visuals to better understand the structure and distribution of the features.

For the predictive analysis step, we employ selected classification models to predict the target feature, that is, if a person makes over $50,000 a year. Lastly, we assess the performance of our models with relevant evaluation metrics and compare our interpretations from EDA with income census data from 2022-2023.

4 METHODOLOGY

This section delineates the procedural framework of the envisioned research project. Initially, in Subsection 4.1, a comprehensive examination unfolds pertaining to the dataset, elucidating the intricacies entailed in the cleansing process. Subsequently, Section 4.2 expounds upon the intricacies of Exploratory Data Analysis (EDA), delving into the methodology applied and meticulously elucidating the resultant findings. The analysis presented underscores the interpretative nuances, providing insights into the discernible patterns and implications extracted from the dataset.

4.1 Data Cleaning

Before proceeding with cleaning, it was necessary to have a general overview of the whole data and make observations of the discrepancies. Upon the first examination of the dataset, the following observations were made:

- The names of the features are in an improper format which could be problematic while accessing and working with individual features. Additionally, some feature names were not easily interpreted.

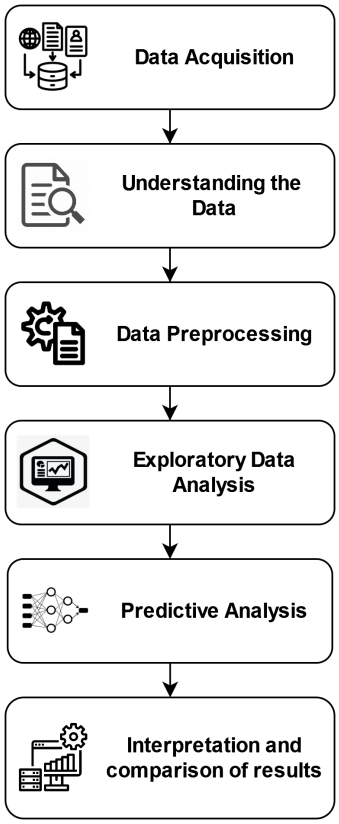

Figure 1: Intended Strategy of Work

- The range of data for several numerical features are inconsistent. For instance, the feature 'age' and 'fnlwgt' are not in the same range.
- There are several observations with missing data. Additionally, these data are represented as '?' instead of true null values.
- Some features have duplicates. For instance, the features 'education' and 'education.num' convey the same information - 'education.num' is the encoded version of 'education'.

The process of cleaning the data was carried out step-by-step by the following steps.

4.1.1 Step 1 - Imports. : The necessary modules from Python were imported. For this work, only Pandas and NumPy sufficed.

4.1.2 Step 2 - Data Transformation. : The names of the features were changed to a consistent format where the words were separated by an underscore rather than a period. They were also renamed to be more interpretable. Additionally, the type of the missing value was changed to a true null from '?'. Moreover, the categorical features were encoded using label encoder.

4.1.3 Step 3 - Data Imputation/Reduction. : The imputation technique used here was hot-deck imputation, where the missing values were replaced with the data from the previous observation. However, this technique might not have been the best one to use as the data was not ordered. Another option was to delete the observations that contained null values. This could have been viable as only about 7% of the whole data contained observations with missing data, and deleting those was very unlikely to have had a significant impact on the predictive analysis.

4.1.4 Step 4 - Data Normalization. : The data was then normalized using the min-max normalization method, where the numerical data were scaled between 0 and 1. The numerical-ordinal data were not altered.

4.1.5 Step 5 - Feature Selection. : While inspecting the distribution of the features it was seen that the features "capital.gain" and "capital.loss" were heavily skewed (as shown in the illustrations in Section 4.2 Exploratory Data Analysis). Heavily skewed data introduce bias in the whole dataset and degrade a model's predictive performance, especially if the model is sensitive to outliers. As a result, the features "capital.gain" and "capital.loss" were dropped from the dataset.

4.2 Exploratory Data Analysis

The first step of EDA begins by having a visual overview of each of the features, and how they are distributed. These can be easily achieved using boxplots and histograms. The following observations are made:

- The range of ages for the individuals is between 1 and 90, where the median age is about 37 years. The data is very slightly skewed to the right.

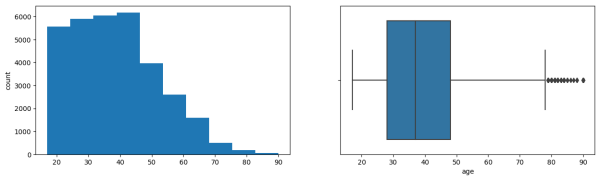

Figure 2: Bar Chart and Boxplot for distribution of 'age'

- The feature 'fnlwgt' is skewed to the right, where the median is at around 0.17 (after scaling). Additionally, a lot of outliers are also present.

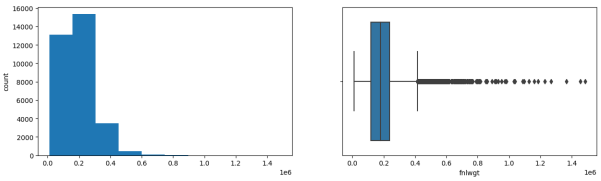

Figure 3: Count vs fnlwgt Bar Chart and Boxplot

- 'capital.gain' and 'capital.lost' are heavily right-skewed distributions where most of the data falls in the outlier range. This finding is consistent with the feature engineering step in data cleaning where these features are recommended to be dropped.

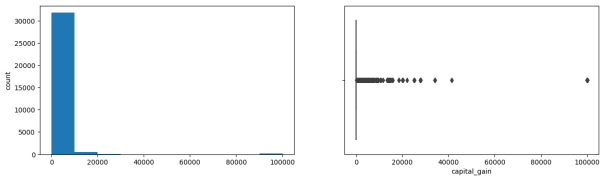

Figure 4: Count vs capital-gain Bar Chart and Boxplot

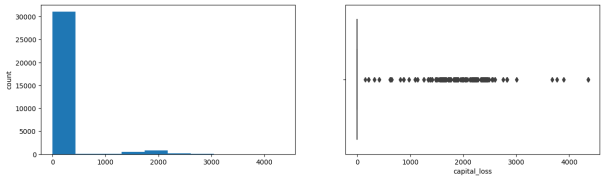

Figure 5: Count vs capital-loss Bar Chart and Boxplot

- The distributions of 'education.number' and 'hours.per.week' are quite similar. Most of the individuals belong to the education category of 10, which corresponds to a high-school graduate. Additionally, most of the individual worked for around 40 hours per week at the private sector.

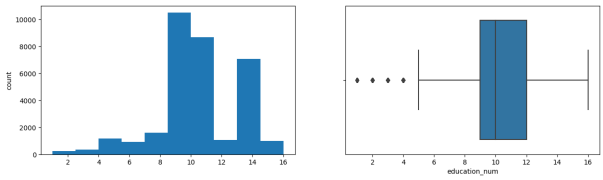

Figure 6: Count vs Education Bar Chart and Boxplot

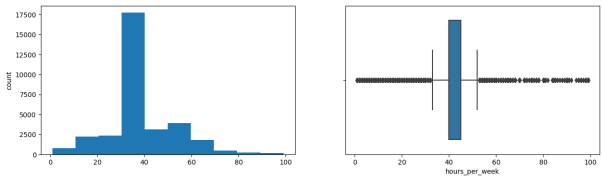

Figure 7: Count vs hours per week Bar Chart and Boxplot

- Most of the individuals in the observations are from the United States. Majority of these observations are ethnically white and males. Additionally, 76% of the total observations earn less than 50K a year, which indicates a significant class imbalance in the data.

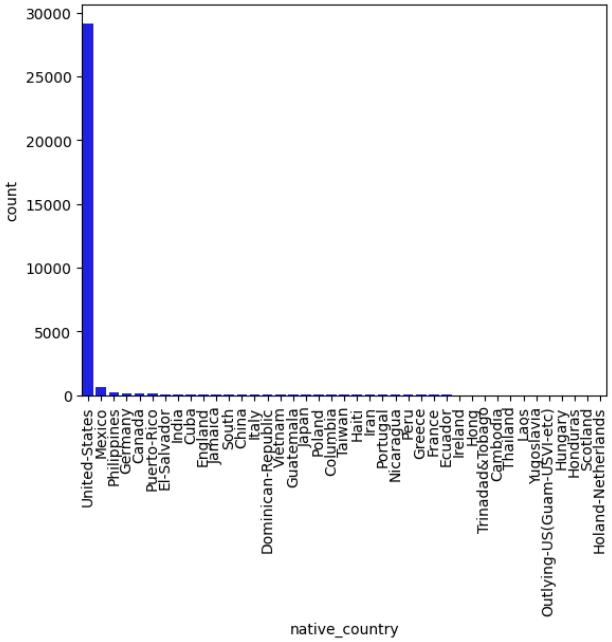

Figure 8: Count vs Country Bar Chart and Boxplot

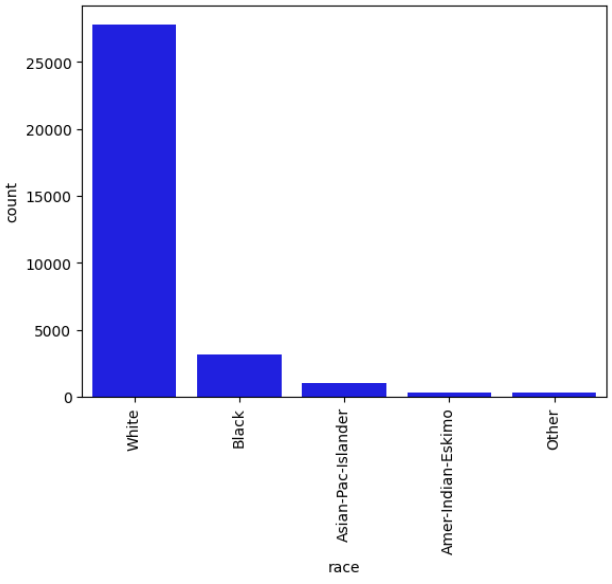

Figure 9: Count vs Race Bar Chart and Boxplot

Furthermore, the aforementioned graphs and results represent the fundamental exploratory data analysis (EDA) that can be performed to draw inferences from the data. However, our task is to delve deeper and gain broader insights, which are presented in the following paragraphs of text.

First, we plotted a jointplot to identify the relationship between the hours worked per week and the age of individuals recorded in the census income dataset. The hues are varied to differentiate people based on their sex (as mentioned in the legend). As observed from the plot, we notice that most people, irrespective of their gender and age, work 40 hours a week. However, it is noticeable that people aged 60-80 work fewer than 40 hours a week, and many people aged 20-50 have work hours of >60 hours per week. This suggests that younger people often tend to work longer hours than older people, but the median is a 40-hour work week. The same can be visualized in figure 10.

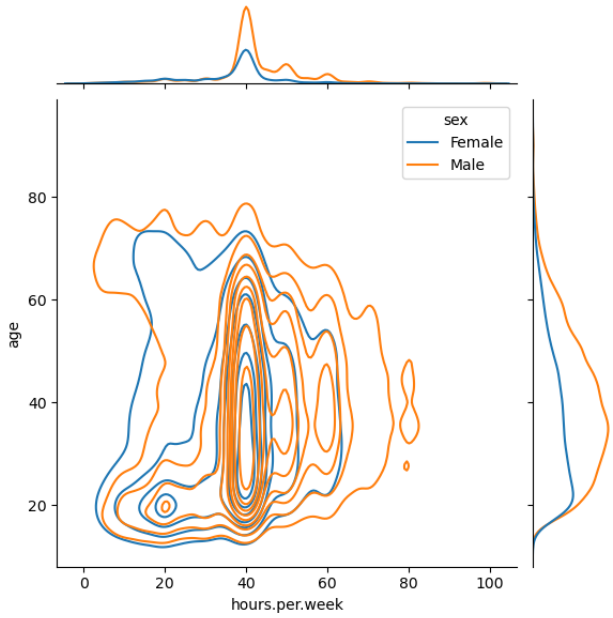

Figure 10: Relationship between age, hours per week & sex

Additionally, we have also plotted multiple line plots to identify any trends and relationships between the age, marital status and hours worked per week. Figure 11 suggests that most women over the age of 50 are widowed and those between the ages of 20-30 are never married. Moreover, the number of hours these women work varies significantly. And, as visualized in figure 11.

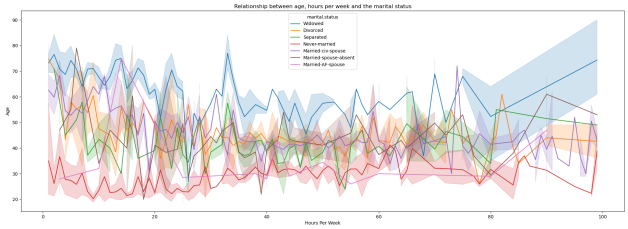

Figure 11: Relationship between age, hours worked per week & marital status

Next, we created a stacked bar plot to identify the relationship between the various working classes concerning gender. The count

serves as the indicative measure plotted and compared for both genders. As seen in the stacked plot (figure 12), most people work in the private sector, and there are fewer females in every working class compared to males. This illustrates a significant difference in the gender ratio during that period.

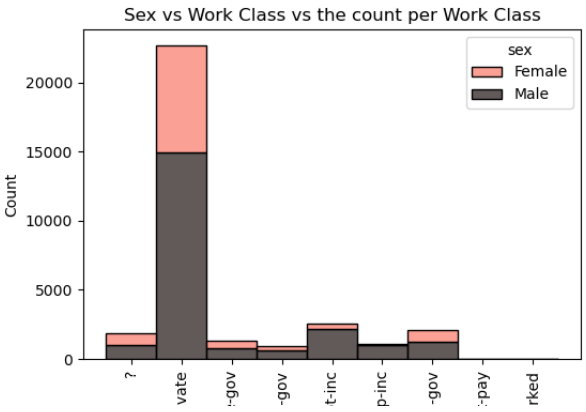

Figure 12: Working class distribution by gender

Equally important, we have plotted a feature correlation graph between the 'occupation' and 'education' features. Interestingly, several inferences can be drawn from this plot (see figure 13).

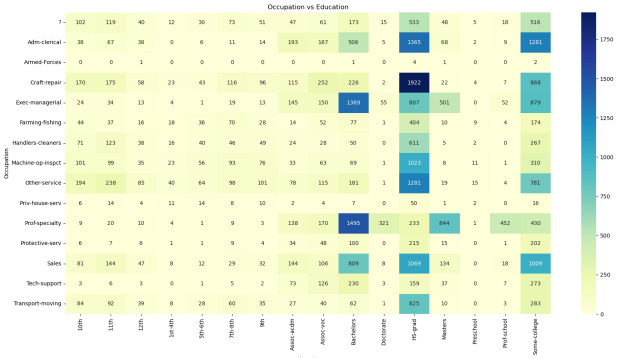

Figure 13: Occupation-Education Correlation Analysis

As seen in figure 13, the following statements can be made:

- Approximately 12 occupations have individuals who have cleared only high school ('HS-Grad' feature) as their education.
- However, individuals in managerial and professional specialty occupations graduated with a bachelor's degree. This indicates that specialized professions often require a higher level of education.
- There are, however, some outliers, with hundreds of people only graduating from 10th or 11th grade, and very few individuals even graduating with a doctorate.
- The Professional Specialty occupation has the highest number of individuals with doctorate degrees, whereas blue-collar occupations have zero to none with doctorate degrees.
- Another interesting observation is that many occupations, such as 'Armed-Forces', 'Priv-house-serv', and others, have very few observations, indicating an imbalance in the collection of data regarding individuals' occupations.

Violin Plots are very informative as they illustrate the numerical distributions with respect to a categorical variable. For instance, we plotted the distribution that shows the hours worked per week for every occupation and distinguished them by their income i.e., '<=50K' and '>50K'. This can be seen in figure 14. Here, if we look closely, you can see that majority of the people who earn more than 50K a year work more than 40 hours a week. Surprisingly, 'Armed-forces' and 'Priv-house-serv' occupations do not earn more than 50K a year.

Also, we noted down the relationship between the income of individuals and their education standing. The findings are shown in table 4. Majority of the distribution are high school grads, followed by some-college grads closely followed by those who have completed their bachelors.

Table 4: Income Distribution by Education Level

Education	<=50K	>50K
10th	871	62
11th	1115	60
12th	400	33
1st-4th	162	6
5th-6th	317	16
7th-8th	606	40
9th	487	27
Assoc-acdm	802	265
Assoc-voc	1021	361
Bachelors	3134	2221
Doctorate	107	306
HS-grad	8826	1675
Masters	764	959
Preschool	51	0
Prof-school	153	423
Some-college	5904	1387

Furthermore, we extracted the relationship between the native country and the income of the people from those countries, highlighting the countries where fewer than 30% of the people have an income greater than 50K. These highlighted countries are presented in Table 5.

Notably, we plotted a bar plot to observe the differences in the counts of each relationship category. Consequently, the hues differentiate the income in every relationship type. It is observed that the number of individuals in the 'Wife' relationship category is significantly lower compared to that of the 'Husband' category. However, the proportion of wives who earn less than and greater than 50K in annual income is very similar, unlike those of husbands. It is also noteworthy to acknowledge that the rest of the relationship categories have a significantly larger proportion of people who earn less than 50K compared to those who earn more than that. The graph can be seen in Figure 15.

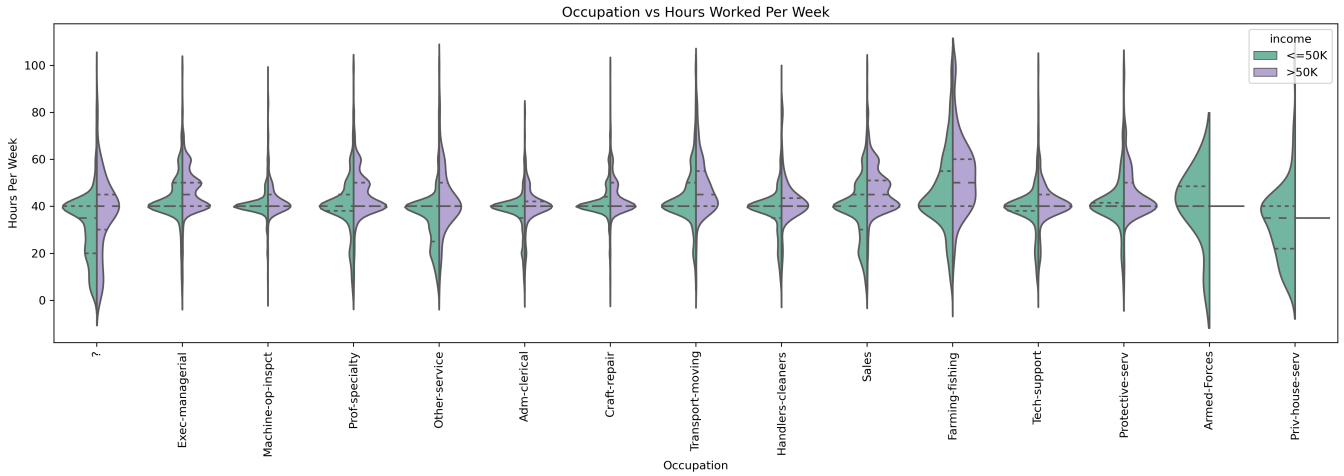

Figure 14: Occupation vs hours per week with respect to Income

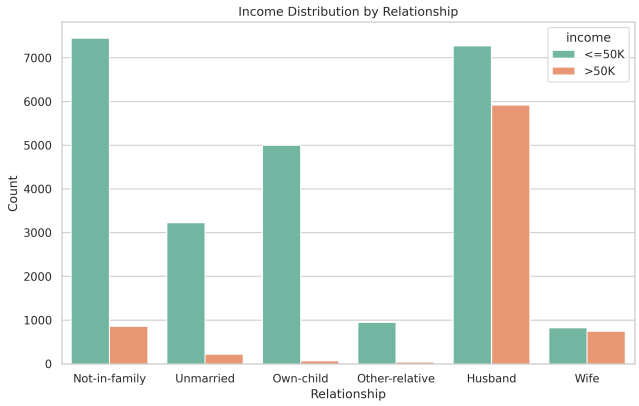

Figure 15: Income Distribution by Relationship

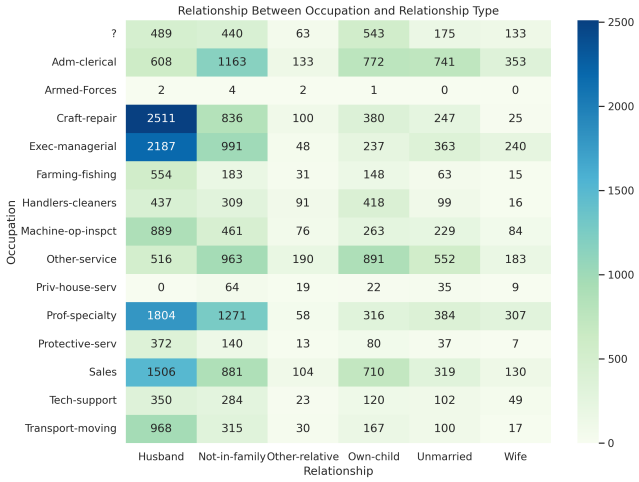

Figure 16: Relationship between Occupation and Relationship

Exploring further, we plotted a heatmap to find the correlation between the 'relationship' and 'occupation' features (Figure 16). The results obtained were as follows:

- The plot suggests a class imbalance, with the number of husbands being larger than that of all the other classes.
- There are no husbands working as private house servants; however, husbands are the most predominant protective servants.
- Neither wives nor unmarried individuals work in the armed forces. It is worth mentioning that the total number of people working in the armed forces is very low to begin with.

We halted our exploratory data analysis at this juncture, as these findings encapsulate the most crucial, pertinent, and concrete results we uncovered. While there is a wealth of additional data to explore, these highlighted outcomes form the cornerstone of our analysis. The following section explains in detail, the modeling techniques used and inferences are made based on that as well.

4.3 Modeling

As previously observed in the literature survey (Table 2), logistic regression and several other models were more prominently utilized compared to others. Nonetheless, a few models have been employed by a limited number of researchers. In our study, we compare the majority of traditionally used models with a couple of non-traditional ones and draw inferences concerning census income prediction. Additionally, we employ these models to gain insights into the distributions of other features.

4.4 Decision Tree

A Decision Tree, a versatile tool in supervised learning, serves both as a predictor and classifier. Its hierarchical tree structure culminates in leaf nodes representing the utmost purity. The determination of each split in the tree involves an impurity measure

Table 5: Countries with <30% population having >50K income

income	<=50K	>50K	perc_>_50K
native.country			
?	437	146	25.042882
Cambodia	12	7	36.842105
Canada	82	39	32.231405
China	55	20	26.666667
Columbia	57	2	3.389831
Cuba	70	25	26.315789
Dominican-Republic	68	2	2.857143
Ecuador	24	4	14.285714
El-Salvador	97	9	8.490566
England	60	30	33.333333
France	17	12	41.379310
Germany	93	44	32.116788
Greece	21	8	27.586207
Guatemala	61	3	4.687500
Haiti	40	4	9.090909
Holand-Netherlands	1	0	0.000000
Honduras	12	1	7.692308
Hong	14	6	30.000000
Hungary	10	3	23.076923
India	60	40	40.000000
Iran	25	18	41.860465
Ireland	19	5	20.833333
Italy	48	25	34.246575
Jamaica	71	10	12.345679
Japan	38	24	38.709677
Laos	16	2	11.111111
Mexico	610	33	5.132193
Nicaragua	32	2	5.882353
Outlying-US	14	0	0.000000
Peru	29	2	6.451613
Philippines	137	61	30.808081
Poland	48	12	20.000000
Portugal	33	4	10.810811
Puerto-Rico	102	12	10.526316
Scotland	9	3	25.000000
South	64	16	20.000000
Taiwan	31	20	39.215686
Thailand	15	3	16.666667
Trinadad&Tobago	17	2	10.526316
United-States	21999	7171	24.583476
Vietnam	62	5	7.462687
Yugoslavia	10	6	37.500000

or information gain. Various measures, such as the Gini index or entropy, can guide these decisions.

In this work, we opted for the Gini impurity measure. The formula for the same is: $Gini(D) = 1 - \sum_{i=1}^{c}(p_i)^2$.

In this equation:

- D represents the dataset being evaluated.
- c is the number of classes in the dataset.
- p_i is the probability of choosing a datapoint of class i from the dataset.

4.5 Random Forest Classifier

The Random Forest Classifier, an ensemble learning paradigm, operates concurrently by harnessing the power of multiple decision trees. Employing a parallel approach, it integrates diverse decision trees trained on distinct sub-datasets to enhance the learning process. In our work, a Random Forest Classifier was instantiated with the amalgamation of 100 decision trees, each constructed with the "**gini**" impurity criterion, contributing to the robustness and collective predictive strength of the model.

4.6 Logistic Regression

Logistic regression is a supervised learning algorithm mainly used for binary classification tasks. In our case, as the target label is binary (0 for "<=50K" and 1 for ">50K"), we achieved good results with it.

4.7 k-nearest Neighbors Classifier

The k-nearest neighbors (k-NN) classifier, another supervised learning algorithm, categorizes data based on the "k" closest labeled instances. Selecting the optimal k is pivotal; a too large k may lead to misclassifications, while a too small k might overlook valuable labeled data. In our study, we employed the elbow method to discern the optimal value for k, ultimately determining it to be 8.

4.8 Gradient Boosting Classifier

Gradient Boosting Classifier is an ensemble learning approach based on the idea that the best model, when merged with prior models, minimizes error the best.

In summary, we utilized the following models using the parameters specified: Following that, the sections that follow examine the performance of the models used and undertake a comparison study, highlighting both the strengths and limitations of these models in regard to our dataset.

5 RESULTS

5.1 Prerequisite Knowledge

To assess the performance of the models, the following evaluation metrics will be used:

- Accuracy
- Precision
- Recall
- Receiver Operating Characteristic (ROC) curve - Area Under Curve (AUC)

To calculate the values for the evaluation metrics, the following terms need to be taken into consideration:

- **True Positive (TP):** Outcome when a model correctly predicts the positive class.
- **True Positive (TN):** Outcome when a model correctly predicts the negative class.

- **False Positive (FP):** Outcome when a model incorrectly predicts a positive class; also known as Type 1 Error.
- **False Negative (FN):** Outcome when a model incorrectly predicts a negative class; also known as Type 2 Error.

5.1.1 Accuracy. : It is defined as the ratio of the number of correct predictions to the total number of predictions.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

5.1.2 Precision. : It is defined as the ratio of true positives to the number of predicted positive outcomes.

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

5.1.3 Recall. : It is defined as the ratio of true positives to the number of actual positive outcomes. This is also known as the True Positive Rate (TPR).

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

5.1.4 ROC Curve. : It is used to measure the performance of a classification model at various thresholds. It is a plot of TPR vs. False Positive Rate (FPR). FPR is defined as the ratio of false positives to the total number of negative outcomes.

$$FPR = \frac{FP}{TN + FP} \tag{4}$$

In general, ROC is a probability curve and AUC represents the measure of separability. A higher value of AUC is an indication of good classification performance. An AUC value of 0.5 suggests no separability, 0.7 to 0.8 is considered acceptable and 0.9 to 1 is considered excellent.

The following section clarifies the results of all the predictive analysis methods used.

5.2 Model Results

The performance metrics of the 6 chosen models have been tabulated in table 6. In the table below, % Acc - Accuracy; P - Precision; R - Recall; AUC - Area Under Curve.

Table 6: Performance Metrics

Model	% Acc	P	R	AUC
Decision Tree	74.4%	75.3%	74.4%	0.67
RandomForest Classifier	80.4%	79.5%	80.4%	0.71
Logistic Regression	81.2%	79.8%	81.2%	0.66
Gaussian Naive Bayes	78%	80.5%	78%	0.75
k-Neighbors Classifier	82.1%	81.4%	82.1%	0.73
Gradient Boosting Classifier	82.9%	82%	82.9%	0.73

With the exception of k-Neighbors Classifier, all the models were used in their default settings. All the models had a similar range of values for their performance metrics, but overall, RandomForest Classifier, Gradient Boosting Classifier and k-Neighbors (with 13 neighbors) performed best.

6 CONCLUSION

Generally, a score of 80% and above for all the performance metrics mentioned denote a model's good performance. Out of the 6 models that were used, the ensemble models, that is, RandomForest Classifier and Gradient Boosting Classifier, performed the best. This is to be expected because the final predictions of such models are the result of aggregation from multiple other models, hence the name ensemble. However, this robustness of the ensemble models should not imply that such models be used with skewed data, which is the case for our dataset.

In reality, skewed data can induce biased learning as this can misrepresent the true distribution of the data, which decreases the models' performance (as evident by the performance metrics). Notably, for the case of k-Neighbors Classifier, such performance is unexpected given that the algorithm is highly sensitive to skewed data. However, the general trend that we saw is that, as the value of k (number of neighbors) increases, the accuracy and the subsequent metrics go up as well, until around $k = 15$, where the numbers plateau. In conclusion, this dataset is more suitable for data trend analysis than predictive modelling, but if the latter is to be conducted, more preprocessing and model tuning is necessary to achieve better performance. Crucially, a greater volume of data is required to ensure ample resources for preprocessing, while simultaneously retaining sufficient data for predictive analysis.

In our analysis, we took in consideration the fact that $1 in the 1990s has the same purchase power of $2.30 today. This means $50k in the 1990s is roughly $115k in today's currency. In the data, significantly more people earn less than $50k a year, which is true for the case of US census data from 2021-2022. More specifically, we have made the following observations:

- The level of income is directly proportional to the education level of an individual. We saw an increase in the number of people with a bachelor's degree or higher earn more than $50k a year, suggesting that a college degree is a significant contributor in the level of income of an individual. This trend, however, is not consistent with other education levels, such as high school diploma or some college degree.

- A significant proportion of individuals younger than 35 years earned less than $50k a year and individuals older than 45 years earned more than $50k a year. This trend is similar for the census data of 2021-2022. Additionally, in both cases, the proportion of individuals in higher level occupational positions, such as executive managers and professional specialist, earning more than $50k were higher compared to other occupations.

- More than 90% of the individuals in the 1990s census data were native to US. Due to this, they had more opportunities at jobs that earned them more than $50k a year, in comparison to the immigrants who were not born in the US. However, the current census data shows that both the natives and the immigrants have had equal opportunities at jobs that pay more than $115k (more than $50k) a year.

REFERENCES

[1] Ron Kohavi. Census Income. UCI Machine Learning Repository, 1996. DOI: https://doi.org/10.24432/C5GP7S.

[2] James Wexler, Mahima Pushkarna, Tolga Bolukbasi, Martin Wattenberg, Fernanda B. Viégas, and Jimbo Wilson. The what-if tool: Interactive probing of machine learning models. *IEEE Transactions on Visualization and Computer Graphics*, 26:56–65, 2019.

[3] Yair Horesh, Noa Haas, Elhanan Mishraky, Yehezkel S. Resheff, and Shir Meir Lador. Paired-consistency: An example-based model-agnostic approach to fairness regularization in machine learning. In *PKDD/ECML Workshops*, 2019.

[4] Yeounoh Chung, Tim Kraska, Neoklis Polyzotis, Ki Hyun Tae, and Steven Euijong Whang. Automated data slicing for model validation: A big data - ai integration approach. *IEEE Transactions on Knowledge and Data Engineering*, 32:2284–2296, 2018.

[5] Mohammed Temraz. A comparison of supervised learning algorithms for the income classification. *International Journal of Computer Applications*, 182(38):19–25, Jan 2019.

[6] Jinglin Wang. Research on income forecasting based on machine learning methods and the importance of features. EAI, 10 2022.

[7] Abd Ur Rehman, Rana M. Saleem, Zeshan Shafi, Muhammad Imran, Manas Pradhan, and Haitham M. Alzoubi. Analysis of income on the basis of occupation using data mining. In *2022 International Conference on Business Analytics for Technology and Security (ICBATS)*, pages 1–4, 2022.

[8] Sharath R, Krishna Chaitanya S, Nirupam K N, Sowmya B J, and K G Srinivasa. Data analytics to predict the income and economic hierarchy on census data. In *2016 International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS)*, pages 249–254, 2016.