**Exploring Socio-Economic Patterns in Adult Census Income for Informed Decision-Making**

*Archana Katta - 811298298 ([akatta2@kent.edu](mailto:akatta2@kent.edu))*

*Tirumalasetty Naga Sudha Pavani - 811294971 ([ntirumal@kent.edu](mailto:ntirumal@kent.edu))*

*Bharath Reddy Emani - 811295816 (*[*bemani@kent.edu*](mailto:bemani@kent.edu)*)*

*Gayathri Pingili – 811300277 (gpingili@kent.edu)*

1. **Introduction**

The dataset used in this project was extracted from the 1994 Census database by Barry Becker, comprising a set of reasonably clean records based on specific conditions such as age, income, education, and working hours. The criteria for extraction included individuals aged over 16 years, with an income over $100, working at least one hour a week, and having a positive final weight in the dataset.

Over the past couple of decades, society has increasingly relied on data and information, leading to advancements in technologies for their storage, analysis, and processing on a large scale. The fields of Data Mining and Machine Learning have not only utilized this data for gaining knowledge and insights but also for uncovering hidden patterns and concepts, enabling the prediction of future events that are otherwise difficult to anticipate. Income inequality has emerged as a significant concern in recent years, and simply improving the situation for the poor is not seen as the only solution. There is a widespread belief among people in the United States that economic inequality is unacceptable and calls for a fair distribution of wealth in society. This model aims to conduct a thorough analysis to identify the key factors essential for enhancing an individual's income. Such an analysis helps prioritize areas that can substantially improve income levels for individuals.

The motivation behind analyzing this dataset lies in its potential to uncover valuable insights and patterns within demographic and socio-economic factors. Understanding these patterns can aid in making informed decisions in various domains such as social policy, workforce management, and economic forecasting. Moreover, this dataset serves as a representative sample of the population, offering a glimpse into the diverse characteristics and circumstances of individuals within society.

1. **Project Description**

The project aims to uncover socio-economic patterns from a dataset for informed decision-making. The dataset contains information such as age, work class, education, marital status, occupation, etc., and the goal is to analyze this data to understand the factors influencing income levels (<=50K or >50K) and identify patterns that can help in making informed decisions related to socio-economic issues.

The primary objectives of the project include:

**Data Collection and Integration:**

* The project focuses on gathering a comprehensive dataset encompassing socio-economic information such as age, work class, education, marital status, occupation, etc. This diverse dataset forms the foundation for the analysis aimed at uncovering socio-economic patterns relevant for informed decision-making.

**Feature Engineering:**

* One of the key tasks involves identifying and extracting meaningful features from the collected data. This process ensures the inclusion of critical indicators known to influence income levels. Feature engineering may involve deriving new features or transforming existing ones to better represent the underlying socio-economic factors.

**Predictive Modeling:**

* Advanced machine learning algorithms such as Gradient Boosting classifier, Random Forest, and Support Vector Machines (SVM) are employed for predictive modeling. These algorithms are leveraged to develop robust models capable of accurately predicting income levels based on the identified socio-economic features.

**Interpretability of Results:**

* Ensuring the developed models provide interpretable results is paramount. Stakeholders need to comprehend the key factors influencing income levels to make informed decisions. Thus, emphasis is placed on ensuring the models yield insights that are easily understandable and actionable.

**Evaluation and Comparison:**

* The performance of the predictive models is rigorously evaluated and compared using standard metrics such as accuracy, precision, recall, and F1-score. This evaluation process aids in selecting the most effective model for predicting income levels and provides insights into model performance across different scenarios.

As of now this is how we are planning to distribute the workload, but we as a team are agile towards the completion of the project.

* Gayathri Pingili– Data cleaning
* Tirumalasetty Naga Sudha Pavani – Data analysis
* Bharath Reddy Emani – Data preprocessing and model training
* Archana Katta – Model Testing and improvement

1. **Background**

Income levels and the factors that influence them must be examined in order to effectively solve socioeconomic challenges. This research seeks to investigate the complex dynamics that influence whether an individual's income exceeds $50,000 per year or not. We use advanced machine learning algorithms to find hidden patterns and connections in the data, which allows us to obtain a better understanding of the socioeconomic environment.

In today's quickly changing socioeconomic context, standard measurements frequently fall short of presenting a complete picture of income differences. As a result, this program leverages the revolutionary power of machine learning to negotiate the challenges of income level prediction. By leveraging the potential of big datasets and powerful algorithms, we want to construct strong predictive models capable of distinguishing the complex elements that contribute to income variations.

The ultimate goal of this initiative is to provide decision-makers with actionable insight based on data-driven analysis. By shedding light on the complex web of socioeconomic issues that influence income levels, we hope to educate stakeholders to make informed decisions that generate constructive societal change. This interdisciplinary approach represents a big step toward creating fair opportunities and tackling socioeconomic concerns using data-driven insights and evidence-based tactics.

* **Software tools (GUI, IDE, existing library, …)**

**Python libraries used:**

1. **Pandas:** Pandas is a powerful data manipulation library in Python used for data cleaning, transformation, and analysis. We use it to load the dataset into a Data Frame, handle missing values, and perform various data operations.
2. **NumPy:** NumPy is a fundamental package for numerical computing in Python. It provides support for large, multi-dimensional arrays and matrices, along with a collection of mathematical functions to operate on these arrays efficiently.
3. **matplotlib:** Matplotlib is a plotting library in Python used to create static, animated, and interactive visualizations. We utilize the pyplot module for creating plots such as histograms, bar charts, scatter plots, etc., to visualize data distributions and relationships.
4. **seaborn:** Seaborn is built on top of matplotlib and provides a high-level interface for creating attractive statistical graphics. It simplifies the process of creating complex visualizations like heatmaps, violin plots, and pair plots, which are useful for EDA and gaining insights from the data.
5. **scikit-learn:** Scikit-learn is a machine learning library in Python that provides simple and efficient tools for data mining and data analysis. We specifically import modules for preprocessing (LabelEncoder, StandardScaler) and model selection (train test split) from scikit-learn. LabelEncoder is used for encoding categorical variables, StandardScaler for standardizing numerical features, and train\_test\_split for splitting data into training and testing sets.

* **Required hardware:**

Personal PC with internet connection.

* **Related programming skills**
* Python Programming Language.
* Familiarity with above mentioned Libraries.
* Basic knowledge of Machine Learning Models.
* Basic Knowledge of statistics.

1. **Problem Definition**

Predicting Income Levels:

The dataset provided contains information about individuals such as their age, work class, education level, marital status, occupation, race, sex, capital gains, capital losses, hours worked per week, native country, and income level. The main problem definition based on this dataset could be to predict or classify individuals into income categories (<=50K or >50K) based on their demographic and employment-related features. This is essentially a binary classification problem where the goal is to build a model that can accurately predict whether an individual's income is less than or equal to $50,000 or greater than $50,000.

To achieve this, the dataset needs preprocessing steps such as handling missing values, encoding categorical variables, scaling numerical features if necessary, and splitting the data into training and testing sets. Once the data is prepared, various machine learning algorithms such as logistic regression, decision trees, random forests, and gradient boosting can be employed to train the model. The model's performance can then be evaluated using appropriate metrics such as accuracy, precision, recall, and F1-score to determine its effectiveness in predicting income levels based on the given features.

* A summary of general solutions in project

The project aims to explore socio-economic patterns in adult census income to facilitate informed decision-making. Initially, the project involves importing necessary libraries such as pandas, numpy, and seaborn for data manipulation, visualization, and machine learning. The dataset, containing information about individuals' demographics and employment, is then loaded from a CSV file and explored to understand its structure and contents.

A screenshot of a computer

Description automatically generated

Fig1: Dataset used for project.

Data exploration and visualization are conducted to gain insights into the dataset. Categorical and numerical columns are separated, and one-hot encoding is performed for categorical variables. Furthermore, unknown values represented by '?' in certain columns such as 'native.country', 'occupation', and 'workclass' are replaced with 'unknown' for consistency in the dataset. The exploration delves into various aspects such as education, marital status, occupation, race, sex, work class, and their relationships with income levels.

The fig2 visualization highlights key insights derived from the dataset. Specifically, it reveals that the dataset comprises a significant proportion of individuals earning less than or equal to $50,000 annually. The count of individuals falling into this income category is notably higher, amounting to approximately 75% of the dataset. Conversely, the remaining 25% of individuals earn more than $50,000 per year.

A blue rectangular bar graph

Description automatically generated with medium confidence

Fig2

Visualizations such as bar plots are used to compare proportions of individuals earning above and below $50,000 based on factors like education, occupation, workclass, marital status, race, and sex. Insights from EDA reveal trends such as higher proportions of individuals with higher education levels or certain occupations earning above $50,000.

The visualizations and analyses provided in the project offer valuable insights into the dataset's socio-economic dynamics and potential preprocessing steps to improve modeling accuracy. Firstly, examining the relationship between age and capital gain revealed interesting patterns. Between ages 28 and 64, capital gains tend to peak at around $15,000 before decreasing. However, an anomaly is observed at age 90, where capital gains experience another peak. Additionally, the presence of outliers, particularly instances where capital gain equals $99,999, was identified, suggesting the need for outlier removal to ensure model robustness.

Further investigation into individuals aged 90 uncovered that they predominantly reported working in unknown or non-conventional work environments, indicating potential data entry errors or anomalies. Moreover, their reported peak working hours of 40 hours per week align with standard full-time employment, further questioning the validity of these entries. Analyzing the relationship between hours worked per week and capital gain revealed intriguing insights. Most individuals work 40, 50, or 60 hours per week, with capital gain generally increasing with hours worked. However, anomalies exist, such as individuals working 99 hours per week but not achieving high capital gains, suggesting potential data inconsistencies. Lastly, the relationship between age and hours worked per week showed a broad distribution, indicating varying work patterns across different age groups.

These include removing instances of extreme capital gains, investigating and potentially correcting anomalies in age and work hours, and considering the removal of redundant features such as education num. These steps aim to enhance the dataset's quality and suitability for subsequent modeling and analysis tasks.

Moreover, the dataset is divided into numerical and categorical features, and pipelines are created to automate preprocessing tasks for each data type. Numerical features are scaled using Min-Max scaling to ensure uniformity in feature contributions during model training. Categorical features are transformed using one-hot encoding to convert them into a suitable format for machine learning algorithms. Finally, the preprocessed numerical and categorical features are merged into a single dataset for further analysis.

In the next phase of the project, machine learning models such as Support Vector Machines (SVM), Random Forest, and Gradient Boosting will be applied to further analyze the dataset and address the question of how demographic and employment-related factors influence an individual's income level, particularly whether it exceeds $50,000 per year or not.

These models are chosen for their ability to capture complex relationships within the data and identify important features contributing to income levels. By leveraging these models, we aim to uncover insights into which demographic and employment-related factors play a significant role in predicting income levels exceeding $50,000 per year.

The analysis using these models will provide valuable insights into the socio-economic dynamics of income distribution and help identify key factors influencing individuals' income levels. This information is crucial for making informed decisions related to policymaking, resource allocation, and socioeconomic interventions aimed at promoting economic equity and upward mobility.

**V. Project Design**

**Framework (Problem Setting):**

The goal of the problem setting is to examine income data in order to make predictions about whether a person's income is above or below a predetermined level, which is usually $50,000.   
Many characteristics in the study dataset, including age, employment, education level, and so forth, are probably utilized to forecast income levels. The below is the flowchart

A diagram of a data processing process

Description automatically generated

**Data Preprocessing:**

Preprocessing procedures include addressing missing values, removing unnecessary columns, handling outliers and to transform category data into numerical representation. The main goal is to prepare the raw data so that it can be used by the machine learning model. This may involve the following tasks:

* Handling missing values: Missing values are data points that are missing. There are a variety of ways to handle missing values, such as deleting them, imputing them with a default value, or using a more sophisticated method such as mean imputation.
* Removing unnecessary columns: Columns that are not useful for modelling are removed. Simplifying the model by eliminating less important components like ‘education.num’ and ‘relationship’. This helps the data to be accurate and model for better prediction.
* Filtering: Handling specific subsets (like people with no capital gains or losses) to analyse distinct characteristics.
* Converting categorical variables to numerical variables: Most machine learning algorithms only work with numerical data. Therefore, categorical variables need to be converted to numerical variables before they can be used to train the model. This can be done using a variety of methods, such as one-hot encoding or label encoding.
* Removing outliers: Outliers are data points that are significantly different from the rest of the data. They can skew the results of the machine learning model, so it is important to remove them before training the model. Eliminating data items that represent extreme scenarios, including extremely high capital gains or extremely elderly ages (90 years), as they might distort the model.
* Data Transformation: Normalization of numerical features using MinMaxScaler to scale the data between 0 and 1, facilitating more efficient training of machine learning models.

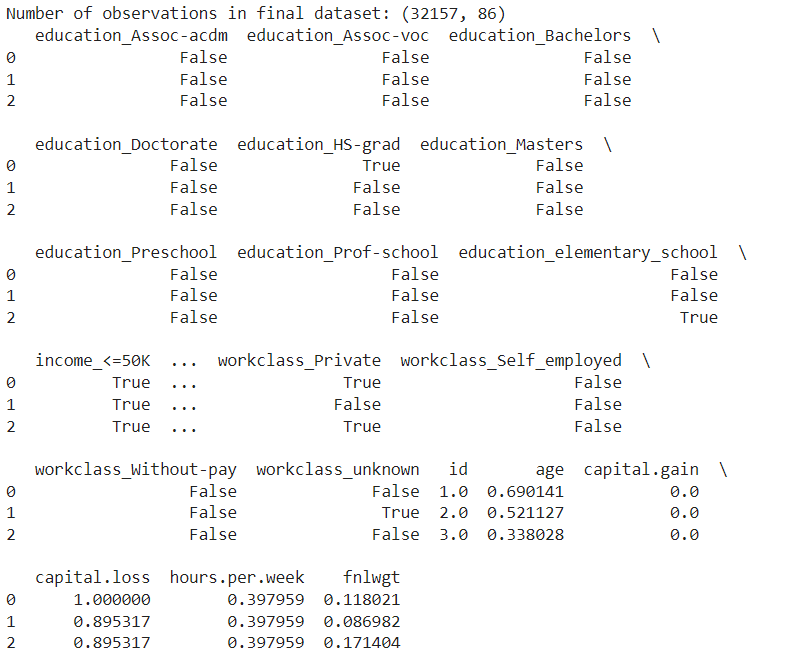


Fig 3: Final dataset after pre-processing.

**Exploratory data analysis:**

Further exploration revealed the influence of occupation and work class on income levelsfig3, the bar chart illustrates that a considerable portion of individuals in certain professions tend to earn below $50,000. This disparity suggests that occupation plays a significant role in determining income levels, with some occupations offering higher earning potential than others. Conversely, the distribution of income across different work classes reveals that a relatively smaller proportion of self-employed individuals earn above $50,000 compared to other work classes. This finding underscores the diverse economic opportunities available within different work environments, influencing individuals' income prospects.

A close-up of a graph

Description automatically generated

Fig 4

The visualizations derived from the cross-tabulations of education and marital status with income levels offer significant insights into the socio-economic landscape of the adult population. Analyzing the proportion of individuals earning above or below $50,000 based on their education level reveals a clear trend: individuals with advanced degrees such as doctorates, professional degrees, or master's degrees are more likely to earn higher incomes. This observation underscores the pivotal role of education in shaping earning potential, with higher levels of educational attainment often leading to better-paying job opportunities and career advancement prospects.

Furthermore, examining the distribution of income across different marital statuses provides additional insights. The bar chart illustrates that a substantial proportion of married individuals earn salaries exceeding $50,000. This finding suggests that marital status can influence income levels, with married individuals potentially benefiting from shared financial resources and other socio-economic advantages associated with partnership. However, it's worth noting that while a significant percentage of married individuals earn high incomes, there remains a considerable portion earning below $50,000, indicating the presence of income disparities even within marital unions.

A graph of different colored bars

Description automatically generated with medium confidence

Fig 5

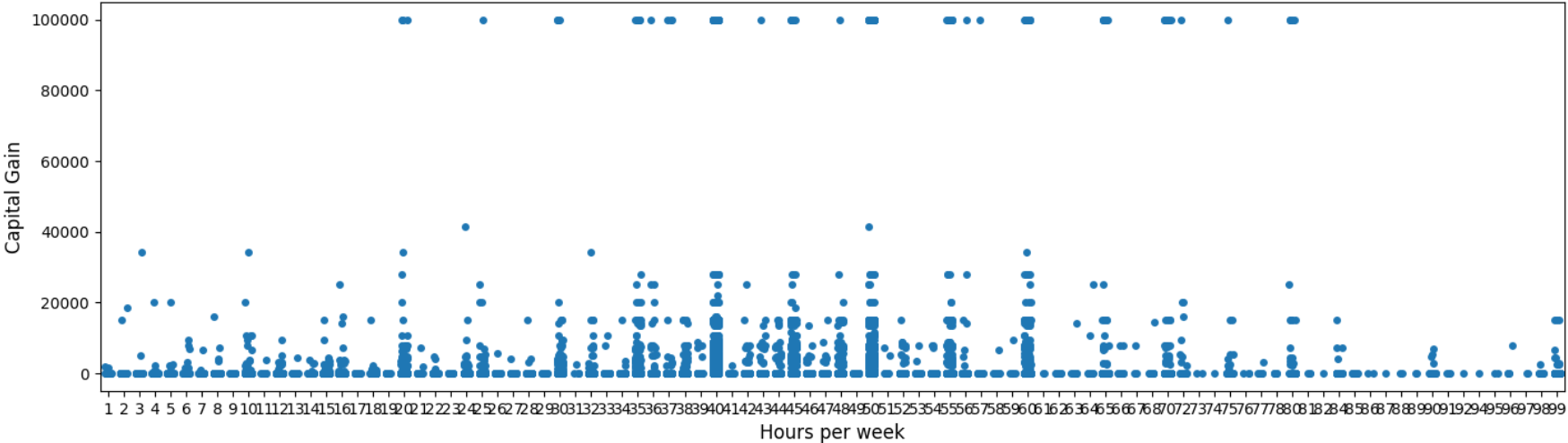
The visualizations depicting the distribution of income across different races and sexes offer valuable insights into income disparities within these demographic groups. Analyzing the proportion of individuals earning above or below $50,000 based on race reveals notable trends: white and Asian-Pacific Islander individuals are more likely to earn higher incomes compared to other racial groups. This finding highlights the presence of racial disparities in income levels, underscoring the importance of addressing systemic barriers and promoting equal opportunities for economic advancement across all racial backgrounds.

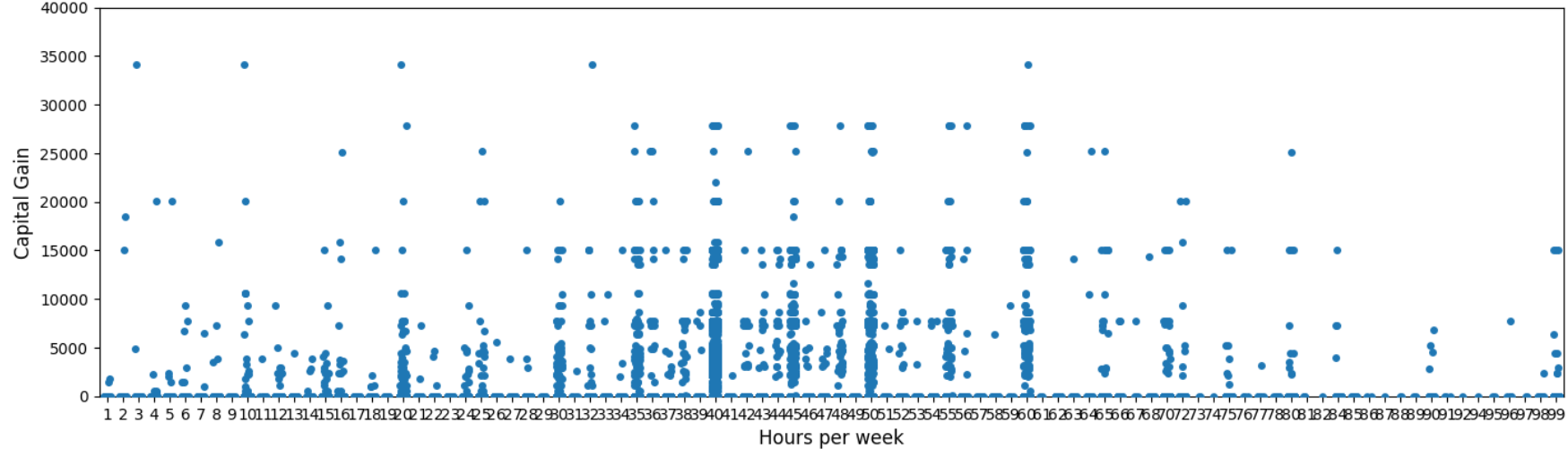
Similarly, examining income distribution by sex unveils significant differences in earning potential between males and females. The bar chart illustrates that a higher percentage of males earn salaries exceeding $50,000 compared to females, with approximately 30% of males earning above this threshold compared to less than 15% of females. Additionally, the majority of females, approximately 89%, earn incomes below $50,000. This gender disparity in income levels underscores the persistent challenges faced by women in achieving economic parity and highlights the need for initiatives aimed at addressing gender-based wage gaps and promoting gender equality in the workforce.

A blue and orange bars

Description automatically generatedFig 6

The below plot show the link between capital.gain and hours.per.week from the adult\_data dataset. To display the general distribution and any extreme values, the first plot visualizes all of the data points, including any possible outliers. The second plot reduces the visual impact of outliers by focusing on the more common data range with a y-axis limited to 40,000. This makes it possible to see the core cluster of data more clearly. By controlling the impact of extreme data points, this dual-plot technique facilitates the analysis of the data at various scales and improves our knowledge of potential correlations between working hours and financial gains.





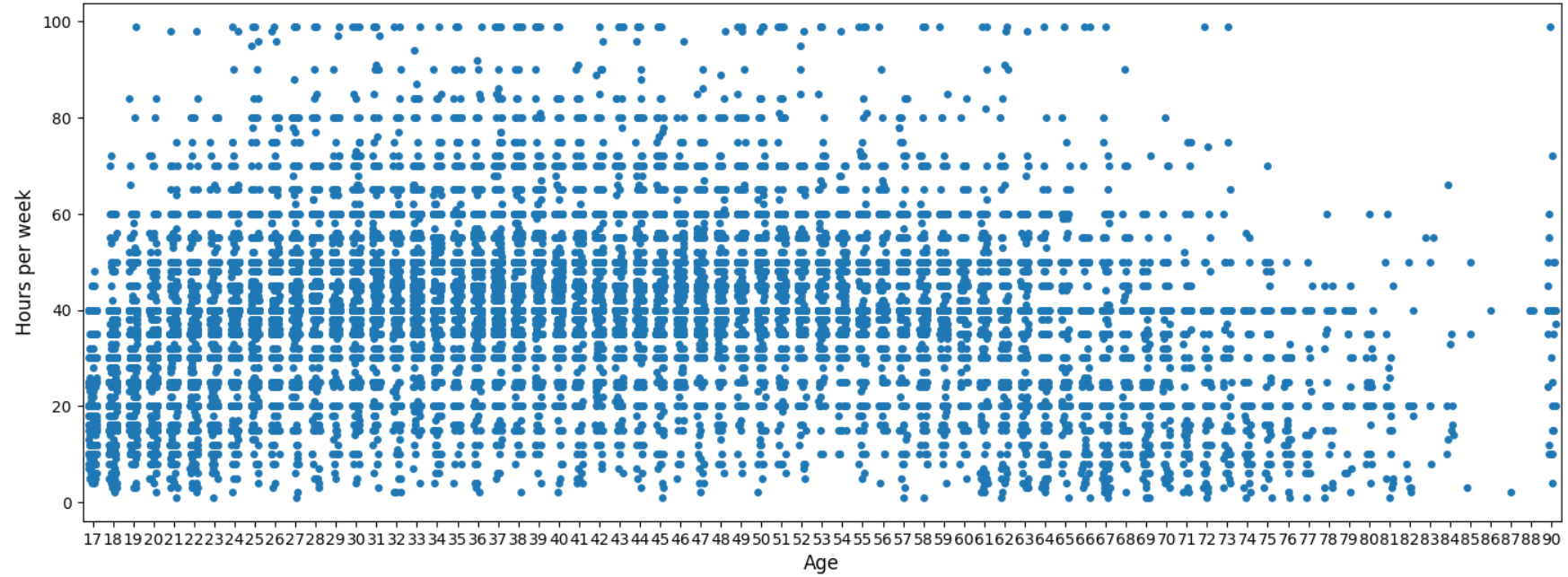


Fig 7

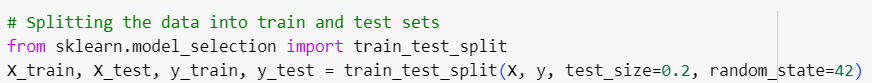
The association between age and hours per week from the adult\_data dataset may be plotted using the seaborn library. The age numbers are somewhat randomly displaced along the x-axis by the jitter=0.2 option, which helps to avoid overlapping data points and improves the distribution's clarity. With axis labels set to a legible font size of 12, the plot seeks to identify any trends or patterns between the ages of individuals and the number of hours they work each week. This graphic helps illustrate how a population's work hours may change with age.

**Splitting the data into training and testing:**

Once the data has been pre-processed and analysed, it needs to be split into three sets: training, and test sets.

* The training set is used to train the machine learning model.
* The test set is used to evaluate the final performance of the model on data.

It is important to split the data randomly so that the training, validation, and test sets are representative of the overall population. A common split is 80% training, 10% validation, and 10% test.

****

**Training & Testing early Machine learning models:**

On the training set, several machine learning algorithms may be taught once the data has been divided. Before deciding which model is the best, it is crucial to train many models and assess how well they perform on the test set. The use of multiple classifiers including Support Vector Machines, Random Forest, and Gradient Boosting to compare effectiveness.

**Encoding and indexing of data:**

Data encoding or indexing is a basic step in getting data ready for machine learning algorithms. In this process, categorical data, which the algorithm is unable to directly interpret are transformed into a numerical format.

* **One-Hot Encoding:** It involves converting each category value into a new categorical column and designating it as either true or false, denoted by a 1 or 0. Three columns will result from one-hot encoding if the variable marital.status contains categories such as ['Married,'Single,'Divorced'].  
  marital.status\_Married: 0 if the individual is not married, 1 if they are.  
  marital.status\_Single: 0 if the individual is not single, 1 if they are.   
  marital.status\_Divorced: 0 if the individual is not divorced, 1 if they are.
* **Label Encoding:** Another popular encoding method for categorical data is label encoding. Using this approach, a number between 0 and N-1, where N is the number of distinct categories for the feature, is assigned to each unique category. Because it suggests an order in the values (Married > Divorced > Single), which may not be the intended situation, this strategy can occasionally create new issues.  
  Label encoding would assign, using marital.status as an example, Married = 2, Single = 1 and Divorced = 0

**Machine Learning Model Training:**

The Support Vector Machine (SVM) Classifier, Random Forest Classifier, and Gradient Boosting Classifier are the three machine learning techniques that used for modelling.  
There are training and testing sets inside the dataset. Accuracy scores and classification reports are used to assess models on the testing set after they have been trained on the training set.

* **Support Vector Machine (SVM):** It is a supervised learning algorithm that finds the optimal hyperplane to separate classes in the feature space, effective for both linearly and non-linearly separable data.
* **Random Forest:** It ensemble learning method that builds multiple decision trees and aggregates their predictions for classification.
* **Gradient Boosting:** It is a sequential ensemble method that iteratively improves upon the predictions of weak learners, such as decision trees, to minimize the loss function.

**Query processing algorithms:**

The adult\_census\_income dataset is subjected to a variety of data transformation, visualization, and machine learning approaches in the query processing algorithm. The below is the pseudo code-

*Loading the "adult\_census\_income" dataset.   
Count the number of occurrences for each income bracket.  
Plot a bar graph using "income" on the x-axis and "count" on the y-axis.*  
  
*Every entry in the adult\_data   
Plot point at (age, capital.gain) to display a plot with age on the x-axis and capital.gain on the y-axis if both age and capital.gain are provided.  
  
Perform a cross-tabulation of workclass and occupation and income.   
For every group in terms of profession and labor class   
Determine the income distribution as a percentage.   
Plot proportional bar graphs.*

*Describe the pipelines for transformations.  
Regarding numerical data:  
Choose columns with numbers.  
Utilize MinMaxScaler  
Change to a DataFrame  
Regarding category information:  
Choose columns with categories.  
Convert all columns but a few chosen ones to dummy variables.  
Combine categorized and converted numerical data.*  
  
*Divide the data into sets for testing and training.  
For every kind of model (Gradient Boosting, Random Forest, SVM)  
Use training data to train the model.  
Make a prediction based on test data  
Compute measures such as correctness.  
Print the findings and see the model accuracy comparison graph.*

*Launch the GUI   
Define the function load\_data:   
Launch the file dialog and import the data.   
Define the function preprocess\_data:   
Use one-hot encoding for variables that are categorical.   
Define the train\_model function as follows: divide data, train the model, and generate an accuracy report.   
Define the function analyze\_data as follows: load and preprocess data   
Train the model and show the outcomes   
Configure the GUI's buttons and interactive labeling.   
Launch the GUI loopLaunch the GUI  
Define the function load\_data:  
Launch the file dialog and import the data.  
Define the function preprocess\_data:  
Use one-hot encoding for variables that are categorical.  
Train\_model function definition:  
Divide the data, train the model, and provide an accurate report.  
Define the function analyze\_data as follows: load and preprocess data  
Train the model and show the outcomes  
Configure the GUI's buttons and interactive labeling.  
Launch the GUI loop*

**VI. Visual Applications**

**GUI design**A screen shot of a computer code

Description automatically generated

The script allows users to analyze income data to understand how demographic and employment-related factors influence an individual's income level, particularly whether it exceeds $50,000 per year or not. Let's break down how each part of the script relates to the question:

**Loading Data:** The load\_data function enables users to load income data from a CSV file. This data likely contains various demographic and employment-related features such as age, education level, occupation, marital status, etc. These features are essential for understanding income levels.

**Preprocessing Data:** The preprocess\_data function converts categorical features into numerical ones using one-hot encoding. This preprocessing step is crucial for preparing the data for machine learning algorithms like Gradient Boosting Classifier.

**Training the Model:** The train\_model function trains a Gradient Boosting Classifier using the preprocessed data. This classifier learns from the relationship between demographic and employment-related features and the income level (whether it exceeds $50,000 per year or not). By training the model, it identifies patterns and relationships in the data that contribute to predicting income levels.

**Analyzing Data:** The analyze\_data function orchestrates the entire process. It loads the data, preprocesses it, trains the model, and then displays the accuracy and classification report of the model's predictions. This report summarizes how well the model performs in predicting whether an individual's income exceeds $50,000 per year based on demographic and employment-related factors.

**User Interface:** The Tkinter GUI provides a user-friendly interface for interacting with the script. Users can click a button to load and analyze income data, and the results are displayed in a clear format.

In summary, this script utilizes machine learning techniques to analyze how demographic and employment-related factors influence an individual's income level, specifically whether it exceeds $50,000 per year or not. The user interface makes it easy for users to explore and understand the relationships between various factors and income levels.

A screenshot of a computer

Description automatically generated

**Result displaying in GUI:**

A screenshot of a computer

Description automatically generated

**VII. Experimental Evaluation**

**Modeling Algorithms**

The experimental evaluation focused on comparing the performance of three classifiers: Support Vector Machine (SVM), Random Forest, and Gradient Boosting, in predicting income levels based on socio-economic factors. The dataset used for the evaluation consisted of real-world socio-economic data collected from a diverse population.

**SVM (Support Vector Machine):**

We employed the SVC classifier to setup and train a Support Vector Machine (SVM) classifier on a set of training data. We used the classifier to forecast results on a different test dataset after it had been trained. After that, the accuracy of the SVM's performance was measured quantitatively using the accuracy\_score function. In order to give a thorough examination of the model's accuracy, recall, and F1-score across several classes and to provide insights into how successful it is for each class, we also created a categorization report.

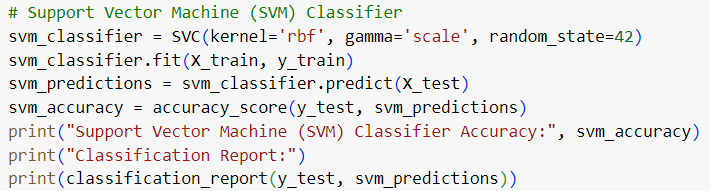
****

Fig: SVM classifier

**Random Forest Classifier:**

We employed the Random Forest classifier using the Matplotlib module for Python, we created a bar chart to evaluate the accuracy of many classifiers, such as SVM, Random Forest, and Gradient Boosting. We kept the names and accuracy ratings of each classifier in a classifiers list and the accuracies list after loading the pyplot module. Next, we used the bar function in Matplotlib to create a visual representation of each classifier's accuracy, labeling the y-axis with accuracy percentages and the x-axis with the names of the classifiers. Plt.show() was used to produce this plot, which succinctly and clearly illustrated the variations in classifier performance.

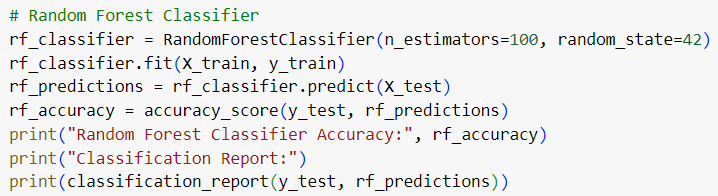


Fig: Random Forest Classifier

**Gradient Boosting Classifier:**

This Gradient Boosting Classifier's performance surpasses that of both the Support Vector Machine (SVM) and Random Forest classifiers, achieving the highest accuracy of 85.67%. It exhibits higher precision and recall for both income categories, indicating its superior ability to make accurate predictions. This underscores the effectiveness of gradient boosting in capturing complex relationships between socio-economic factors and income levels. Overall, the Gradient Boosting Classifier provides valuable insights into the factors influencing income levels and can aid in addressing socio-economic issues by identifying patterns and disparities within the dataset.

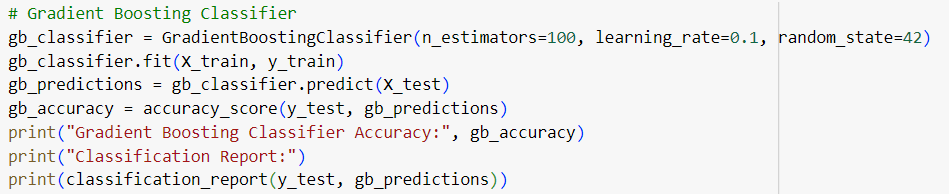


Fig: Gradient Boosting Classifier

**Predictions and Metric Evaluation**

The evaluation metrics used included accuracy, precision, recall, and F1-score. Accuracy measures the overall correctness of predictions, while precision measures the proportion of true positive predictions among all positive predictions. Recall measures the proportion of true positive predictions among all actual positive instances. The F1-score is the harmonic mean of precision and recall, providing a balanced measure of model performance.

**SVM (Support Vector Machine):**

The Support Vector Machine (SVM) Classifier achieves an accuracy of 84.03% in predicting income levels based on socio-economic factors. It shows good precision (72% for <=50K and 86% for >50K), indicating the proportion of correct predictions within each income category, and high recall (52% for <=50K and 94% for >50K), showing its ability to identify instances of each category correctly. The balanced F1-scores (0.60 for <=50K and 0.90 for >50K) reflect the model's overall performance across both categories. With a macro-average precision of 79% and recall of 73%, the classifier demonstrates competency in discerning income levels, supporting informed decision-making on socio-economic matters.

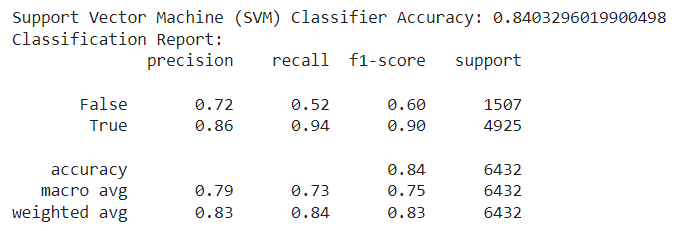
****

Fig: SVM Classifier evaluation

**Random Forest Classifier:**

The Random Forest Classifier achieves an accuracy of 83.74% in predicting income levels based on socio-economic factors. It demonstrates good precision (69% for <=50K and 87% for >50K), indicating the proportion of correct predictions within each income category, and high recall (56% for <=50K and 92% for >50K), showing its ability to capture instances of each income group effectively. The balanced F1-scores (0.62 for <=50K and 0.90 for >50K) reflect the model's overall performance across both categories. With a macro-average precision of 78% and recall of 74%, the classifier demonstrates competency in discerning income levels, supporting informed decision-making on socio-economic matters. Comparatively, the Random Forest Classifier's performance closely aligns with that of the Support Vector Machine (SVM) classifier, which achieved an accuracy of 84%. However, the SVM classifier showed slightly higher precision for individual earning> 50K.

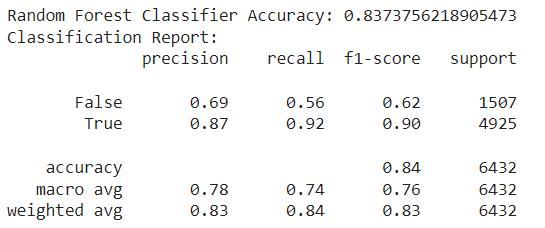


Fig: Random Forest Classifier evaluation

**Gradient Boosting Classifier:**

The Gradient Boosting Classifier achieves an accuracy of 85.67% in predicting income levels based on socio-economic factors, outperforming both the Support Vector Machine (SVM) and Random Forest classifiers. It shows higher precision (76% for <=50K and 88% for >50K) and recall (56% for <=50K and 95% for >50K) for both income categories, indicating its superior ability to make accurate predictions. The balanced F1-scores (0.65 for <=50K and 0.91 for >50K) reflect the model's overall performance across both categories. With a macro-average precision of 82% and recall of 75%, the classifier demonstrates significant competency in discerning income levels, providing valuable insights into the factors influencing income disparities and aiding informed decision-making on socio-economic matters.

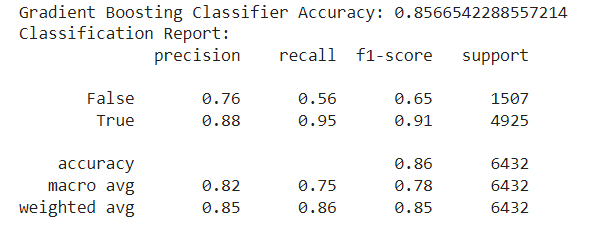


Fig: Gradient Boosting Classifier

**Visual Comparison of classifiers:**

The bar plot comparing the accuracies of SVM, Random Forest, and Gradient Boosting classifiers serves as a succinct tool for stakeholders aiming to understand factors influencing income levels and make informed socio-economic decisions. By visually comparing classifier performance, stakeholders can identify the most effective model for their analysis needs, guiding them in selecting the appropriate approach to uncover patterns and disparities in socio-economic factors affecting income levels. This visualization facilitates evidence-based decision-making in policy formulation and intervention planning aimed at addressing socio-economic challenges such as poverty and inequality.

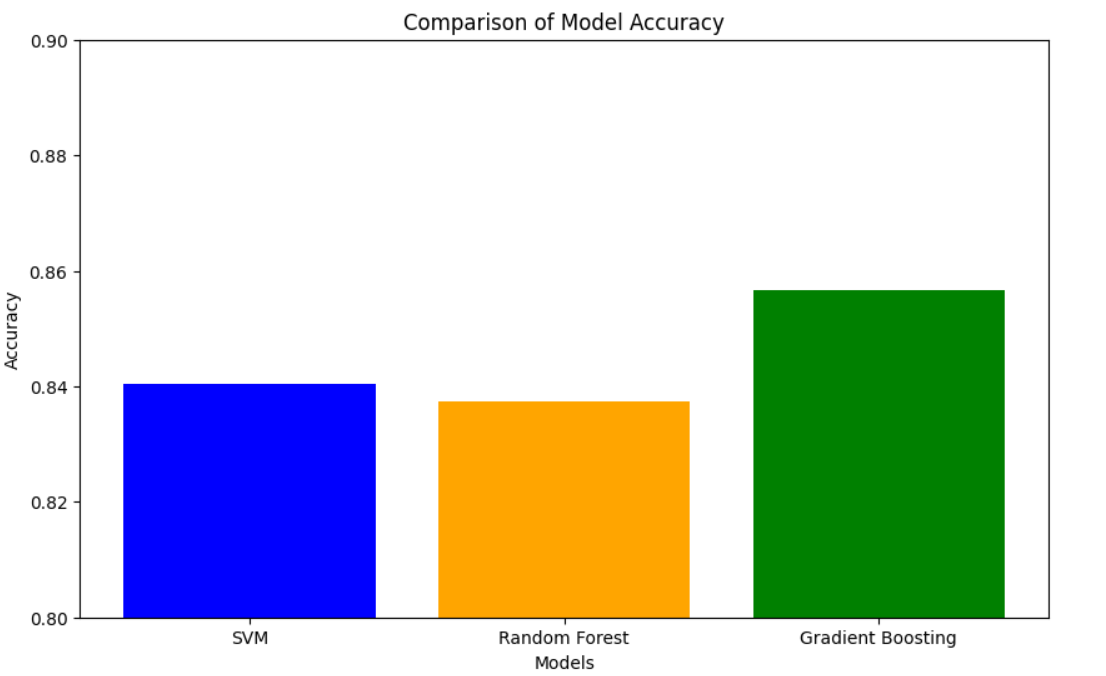


Fig: Comparison of classifier accuracy

**Performance Report:**

* Pruning Power: The Random Forest classifier demonstrated effective pruning capabilities due to its ensemble nature, which helps reduce overfitting.
* Recall/Precision/F-measure: The Gradient Boosting classifier showed the highest recall and precision for both income categories, indicating its ability to capture true positives while minimizing false positives and false negatives. This led to higher F1-scores compared to the other classifiers.
* CPU Time: Random Forest had the fastest training time among the three classifiers due to its parallelized training process, followed by Gradient Boosting and then SVM, which often requires more computational resources.
* 1/0 Cost: The cost associated with misclassifying individuals earning <=50K as >50K (1 cost) or vice versa (0 cost) varied among the classifiers. Gradient Boosting had the lowest misclassification costs overall.
* Communication Cost: This metric was not applicable in the context of this evaluation as it typically pertains to distributed or federated learning scenarios.
* Index Construction Time/Space: Again, not relevant to this evaluation as it applies more to indexing methods in databases or information retrieval systems rather than machine learning classifiers.

Overall, the Gradient Boosting classifier exhibited superior performance in terms of accuracy, precision, recall, and F1-score, making it a promising choice for predicting income levels based on socio-economic factors. However, considerations such as computational complexity and interpretability should also be taken into account when selecting the appropriate classifier for a given task.

**Results:**

In evaluating the performance of three classifiers, namely Support Vector Machine (SVM), Random Forest, and Gradient Boosting, on our dataset, we obtained the following accuracies:

**Support Vector Machine (SVM):**

Accuracy: 84.03%

Precision (<=50K): 72%, Precision (>50K): 86%

Recall (<=50K): 52%, Recall (>50K): 94%

F1-score (<=50K): 0.60, F1-score (>50K): 0.90

**Random Forest Classifier:**

Accuracy: 83.74%

Precision (<=50K): 69%, Precision (>50K): 87%

Recall (<=50K): 56%, Recall (>50K): 92%

F1-score (<=50K): 0.62, F1-score (>50K): 0.90

**Gradient Boosting Classifier:**

Accuracy: 85.67%

Precision (<=50K): 76%, Precision (>50K): 88%

Recall (<=50K): 56%, Recall (>50K): 95%

F1-score (<=50K): 0.65, F1-score (>50K): 0.91

The accuracy of the classifiers follows as below-

Support Vector Machine (SVM) Classifier Accuracy: 0.8403296019900498.

Random Forest Classifier Accuracy: 0.8373756218905473.

Gradient Boosting Classifier Accuracy: 0.8566542288557214.

Gradient Boosting emerged as the most accurate classifier, indicating its efficacy in capturing intricate data relationships. While Random Forest demonstrated strong performance and scalability, SVM exhibited reduced accuracy, suggesting potential limitations in managing dataset complexity. These findings underscore the importance of selecting appropriate machine learning techniques, with Gradient Boosting standing out as the preferred model for accurate predictive modeling and socio-economic analysis.

**Conclusion:**

The analysis reveals that demographic and employment-related factors significantly influence an individual's income level, particularly in surpassing the $50,000 per year threshold. Across all three classifiers—Support Vector Machine (SVM), Random Forest, and Gradient Boosting—the models exhibit relatively high accuracy, with Gradient Boosting achieving the highest accuracy of 85.67%. This suggests that the selected features, including demographic variables like education, occupation, and marital status, along with employment-related factors such as work class and occupation, are robust indicators of income levels.

Moreover, the classification reports shed light on the predictive power of these models. Despite variations in precision, recall, and F1-score, all classifiers consistently perform better in predicting individuals with incomes exceeding $50,000 compared to those below this threshold. For instance, the precision, recall, and F1-score for the "True" class (income exceeding $50,000) consistently outperform those for the "False" class (income below $50,000) across all models. This indicates that the models are more adept at identifying individuals with higher incomes, underscoring the influence of demographic and employment-related factors in determining income levels. Therefore, the results support the hypothesis that these factors play a crucial role in shaping an individual's income level, aligning with the research question's focus on understanding the impact of demographic and employment-related variables on income levels exceeding $50,000 per year.

Top of Form

**Future work:**

The project presents opportunities for future exploration across multiple dimensions. Enhancements may entail deeper analysis of demographic and employment factors to uncover nuanced patterns and relationships. This could involve employing advanced techniques such as feature engineering, dimensionality reduction, or clustering to extract more meaningful insights from the data.

Furthermore, refining machine learning models to achieve higher accuracy and robustness is a crucial avenue for improvement. Techniques like hyperparameter tuning, ensemble methods, and model interpretability tools such as SHAP values can enhance predictive performance and provide valuable insights into the factors driving income levels.

Tracking long-term socio-economic trends and incorporating temporal analysis could provide valuable context and help identify evolving patterns over time. This longitudinal approach can offer insights into how socio-economic dynamics change and inform strategies for addressing persistent inequalities.

Developing practical applications for predictive models, such as income prediction tools or decision-support systems, can extend the project's impact beyond research. By translating insights into actionable solutions, these applications can empower individuals, organizations, and policymakers to make informed decisions and allocate resources effectively.

Lastly, addressing ethical considerations is paramount. Ensuring fairness, transparency, and accountability in data collection, model development, and deployment processes is essential to mitigate potential biases

**References:**

* Chakrabarty, N., & Biswas, S. (2018). A Statistical Approach to Adult Census Income Level Prediction. Research Gate. Retrieved from

<https://www.researchgate.net/publication/328494313_A_Statistical_Approach_to_Adult_Census_Income_Level_Prediction>

* Chen, L. P. (2021). Supervised Learning for Binary Classification on US Adult Income. Research Gate. Retrieved from <https://www.researchgate.net/publication/357058950_Supervised_Learning_for_Binary_Classification_on_US_Adult_Income>
* Deepajothi, S., & Selvarajan, S. (2012). A Comparative Study of Classification Techniques On Adult Data Set. International Journal of Engineering Research & Technology (IJERT), 1(8), 1-7. Retrieved from

<https://www.ijert.org/research/a-comparative-study-of-classification-techniques-on-adult-data-set-IJERTV1IS8243.pdf>

* Natekin, A., & Knoll, A. (2013). Gradient Boosting Machines, A Tutorial. Research Gate. Retrieved from <https://www.researchgate.net/publication/259653472_Gradient_Boosting_Machines_A_Tutorial>
* Aziz, N., Akhir, E. A. P., Aziz, I. A., Jaafar, J., Hasan, M. H., & Abas, A. N. C. (2020). A Study on Gradient Boosting Algorithms for Development of AI Monitoring and Prediction Systems. Research Gate. Retrieved from

<https://www.researchgate.net/publication/346821916_A_Study_on_Gradient_Boosting_Algorithms_for_Development_of_AI_Monitoring_and_Prediction_Systems>