Census Income Prediction

A Machine Learning Approach

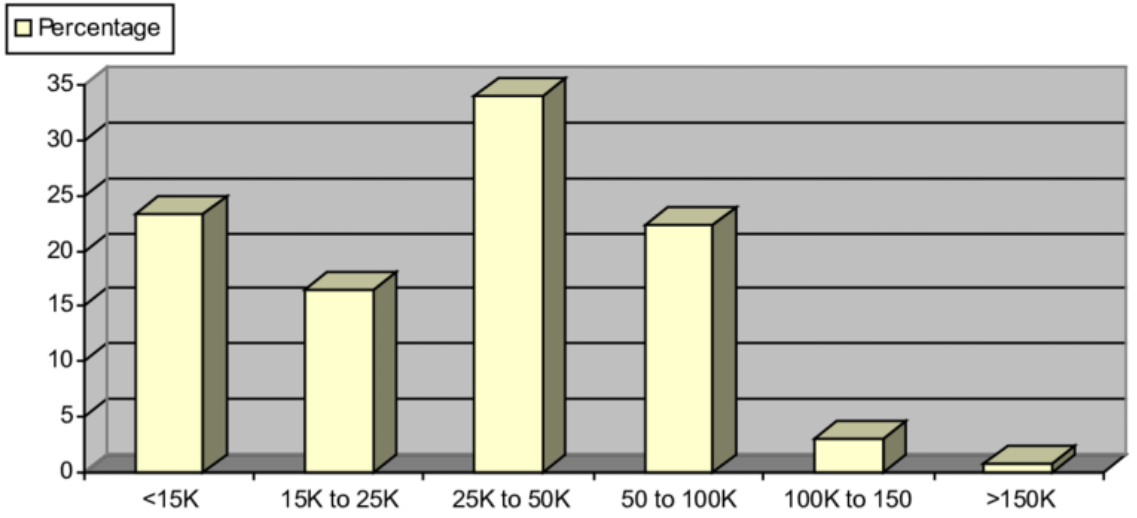
INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY
DELHI

Introduction to Problem Statement

Predicting income levels based on demographic and socioeconomic factors is complex.

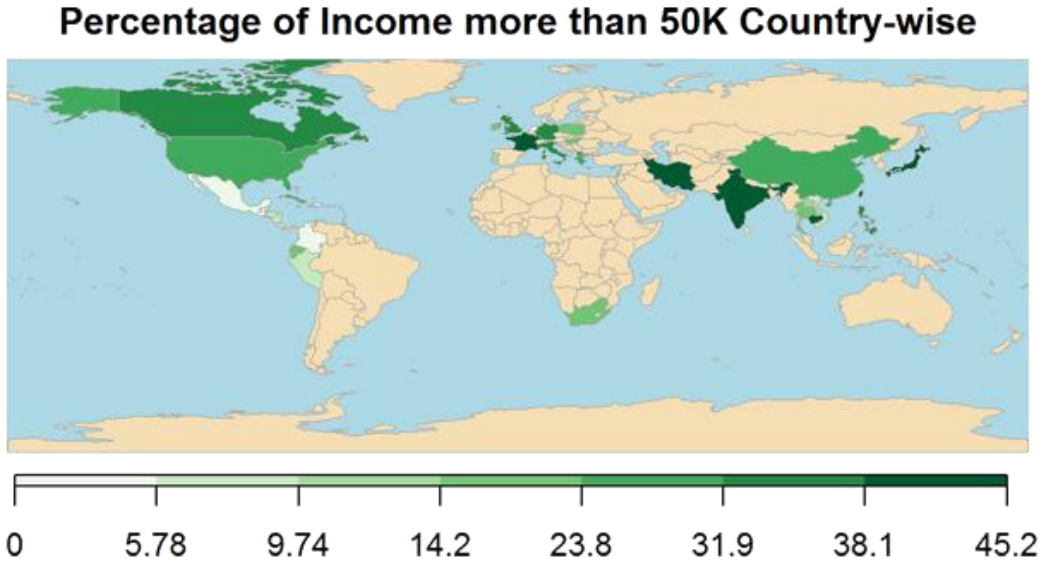
Binary Classification: Identify if an individual's income is above $50,000 or not.

Motivation: Develop multiple machine learning models to determine the best approach for accurate predictions.



Why This Problem Matters?

- Practical applications in **policy-making**, **targeted marketing**, and **financial planning**. (E.g. targeted financial services using income predictions)
- Comparative analysis helps in understanding the **strengths** and **weaknesses** of the algorithm in a real-world scenario.
- **Goal:** Find the most effective methodology to predict income levels accurately and efficiently.



Literature Review

Paper 1

Adult Income Classification using Machine Learning Techniques

- **Objective:** Applied different machine learning algorithms to the adult income dataset.
- **Key Findings:** Logistic Regression and Decision Trees showed good performance, with Random Forest outperforming them.
- **Relevance**: Provides baseline models and insights for comparison in this project.
- **Link:** IEEE Xplore - Adult Income Classification

Paper 2

A Comparative Analysis of Machine Learning Algorithms for Classification Purpose

- **Objective:** Analyzed various classification algorithms across multiple domains.
- **Key Findings:** Identified trade-offs between accuracy and interpretability (e.g. Neural Networks vs. Decision Trees).
- **Relevance:** Informs our model selection strategy and highlights the need for balancing accuracy and interpretability.
- **Link:** IEEE Xplore - Comparative Analysis

Dataset Description

1. Overview of the Dataset
2. Visualization of Key Attributes
3. Preprocessing Details

Overview of the Dataset

- **Dataset Name:** Census Income Dataset
- **Target Variable:** Income (Binary Classification: >$50K or <= $50K)
- **Attributes:** age, work-class, education, marital-status, occupation, relationship, race, sex, hours-per-week, capital-gain, capital-loss, native-country, education-num, fnlwgt (final weight) .

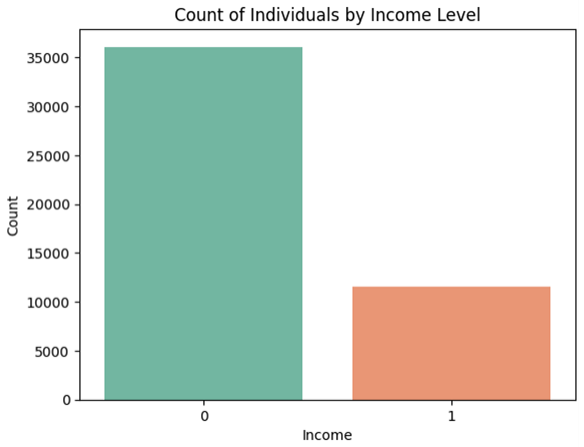
Name	Type	Values	Description
age	Continuous	From 17 to 90	The age of an individual
workclass	Nominal	Private, Federal-Government, etc	A general term to represent the employment status of an individual
fnlwgt	Continuous	Integer greater than 0	Final weight: is the number of people the census believes the entry represents
education	Ordinal	Some-college, Prof-school, etc	The highest level of education achieved by an individual
education-num	Discrete	From 1 to 16	The highest level of education achieved in numerical form
marital-status	Nominal	Married, Divorced, etc.	Marital status of an individual
occupation	Nominal	Transport-Moving, Craft-Repair, etc	The general type of occupation of an individual
relationship	Nominal	Unmarried, not in the family, etc	Represents what this individual is relative to others
race	Nominal	White, Black, Hispanic, etc.	Descriptions of an individual's race
sex	Nominal	Male, Female	The biological sex of the individual
capital-gain	Continous	Integer greater than or equal to 0	Capital gains for an individual
capital-loss	Continous	Integer greater than or equal to 0	Capital loss for an individual
hours-per-week	Continous	From 1 to 99	The hours an individual has reported to work per week
native-country	Nominal	United-States, Cambodia, England, Puerto-Rico, Canada and more	Country of origin for an individual
income	Discrete	(≤50k USD, >50k USD)	The label whether or not an individual

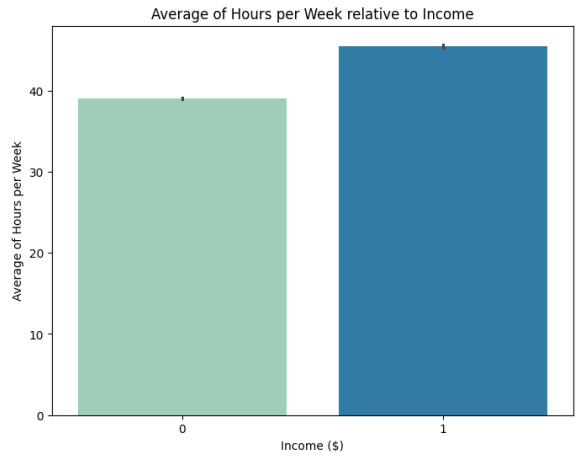
Visualization of Key Attributes

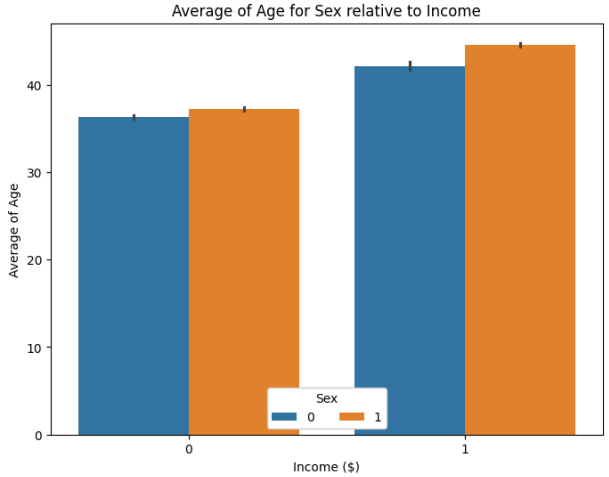
- **Bar Plot:** Income, Avg of hours per week, Avg of age for sex
- **Heatmap:** Income vs Education
- **Histogram:** Numerical Features
- **Box and Violin Plot:** Box plot for numerical features and Violin plot for categorical features
- **Pair Plot:** All features

Bar Plots

1. **Income Distribution:** Bar plot showing the number of people in different income classes (i.e. > $50K or <= $50K).
2. **Average of Hours per Week vs Income:** Bar plot depicting relationship of income growing directly with the average of work hours per week.
3. **Average of Age for Sex vs Income:** Bar plot showing that the people with income 50K have a higher average age. With male average age being greater in both cases.

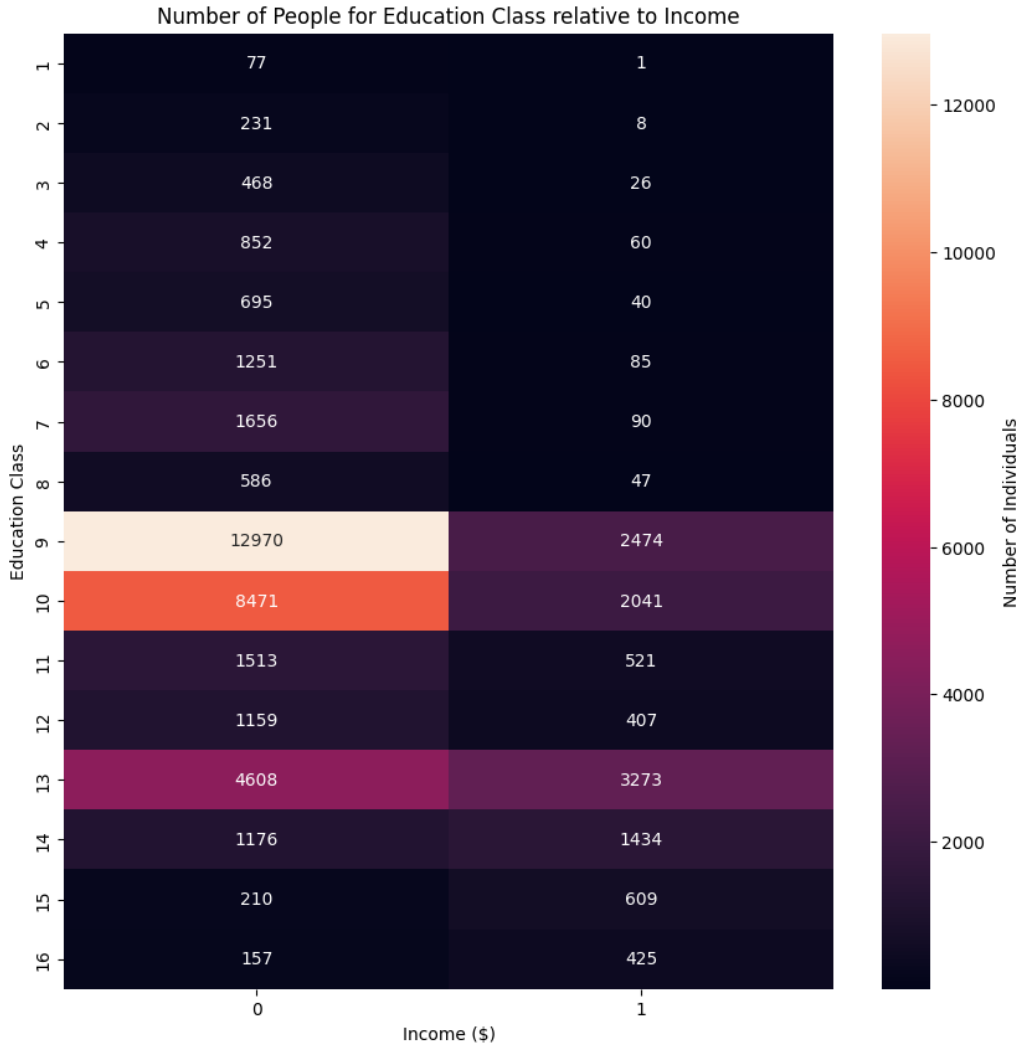






Heatmap

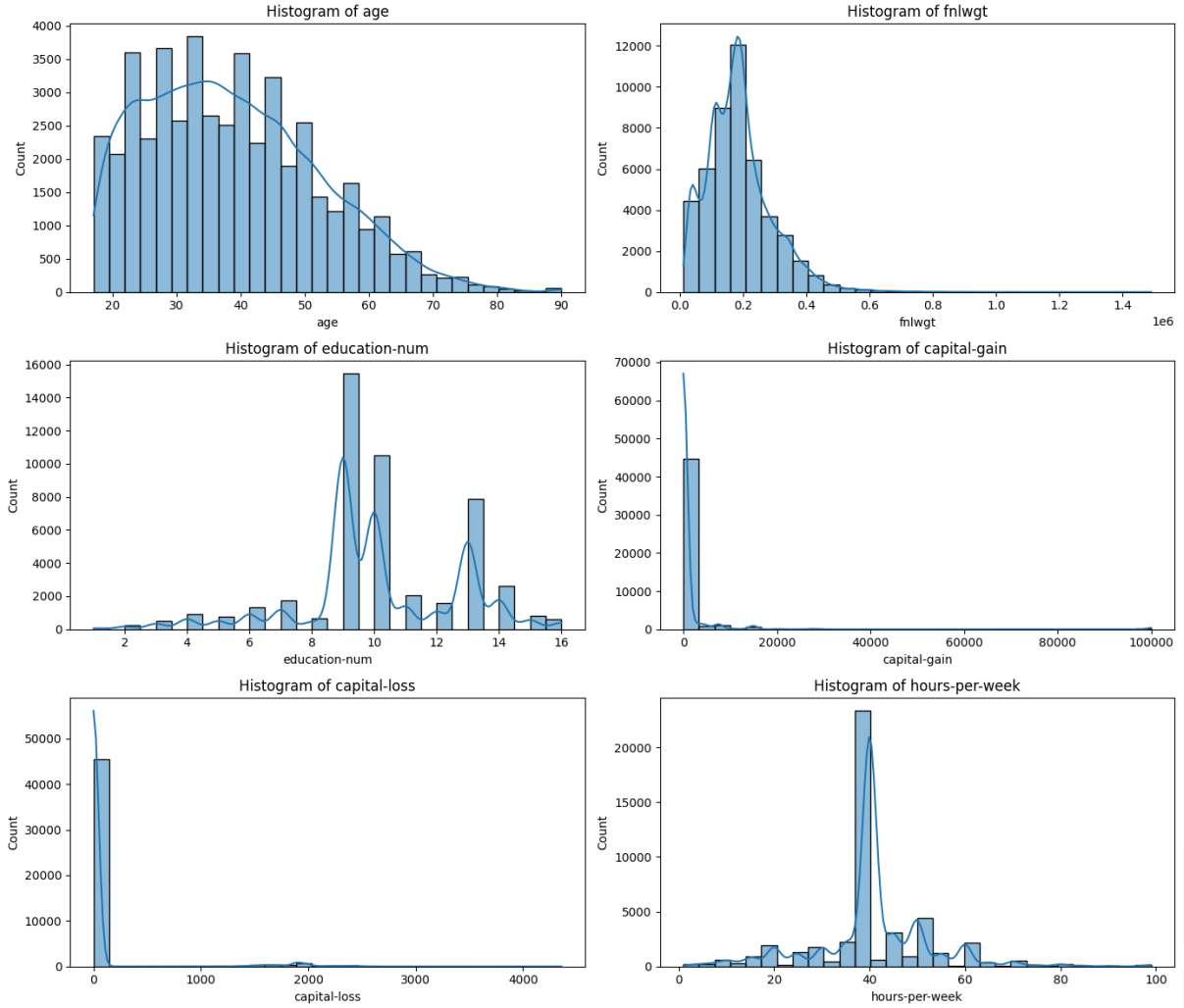
- The Income-Education heatmap suggests a strong relationship between education and income.
- Notably, the 9th and 10th education levels have the highest counts overall lower income (≤$50k).
- While level 13 shows significant representation in both income categories.



Histogram

Histograms of numerical features:

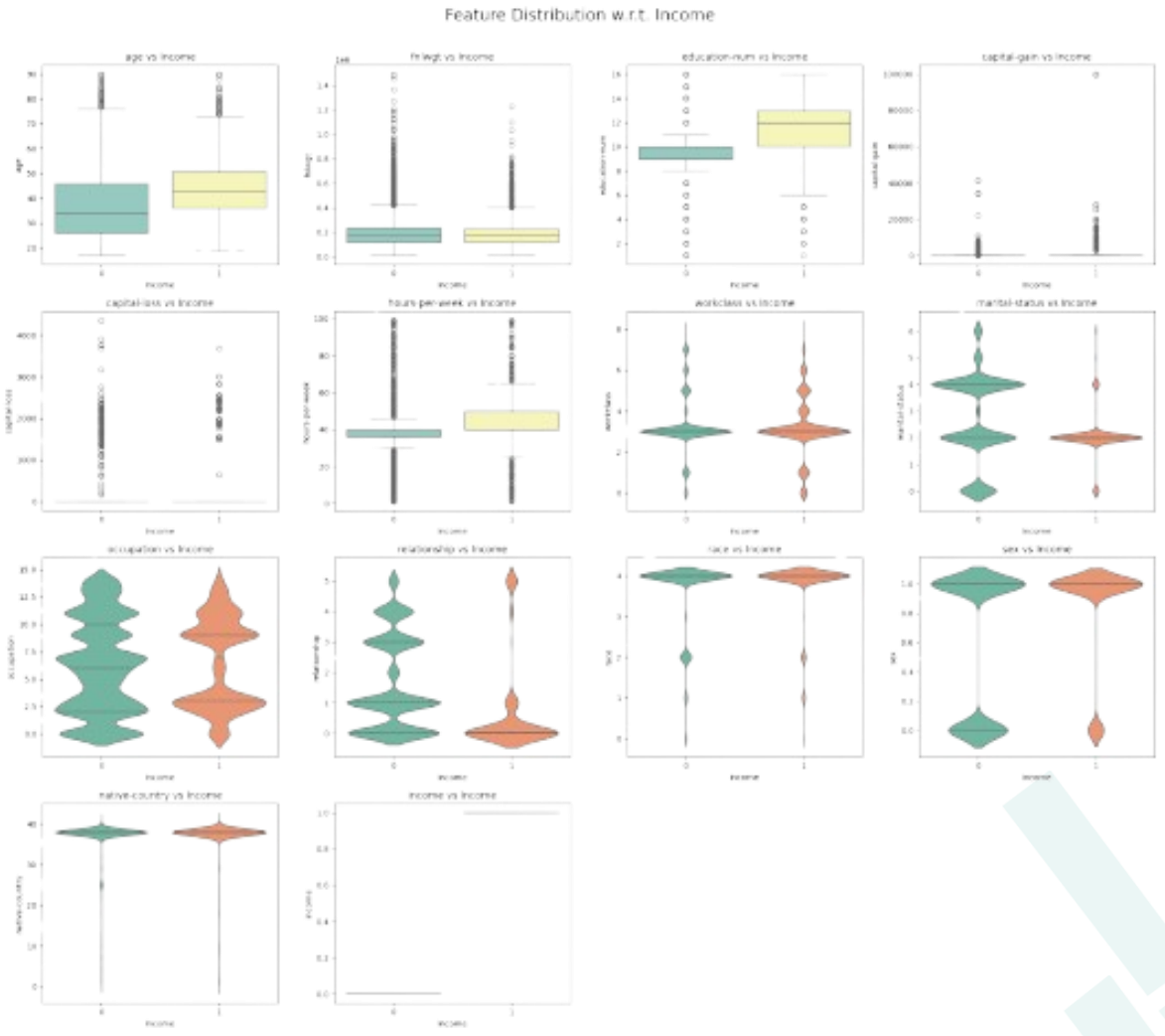
1. **age:** The distribution is right-skewed with most people between 20 and 50 years old.
2. **fnlwgt (final weight):** Strongly right-skewed, with the majority of values concentrated below 0.2 million.
3. **education-num:** Bimodal distribution, peaking at 9 and 10 years of education.
4. **capital-gain:** Highly right-skewed, with most people having zero capital gains.
5. **capital-loss:** Most people have zero capital losses, with very few reporting large losses.
6. **hours-per-week:** Peaks at around 40 hours, with a concentration of people working between 35-45 hours.



Box and Violin Plot

Key observations w.r.t income:

- **capital-gain**: Strongly correlated with high income, with significant gains in individuals earning >50K.
- **capital-loss**: Higher capital losses also occur more frequently in the >50K income group.
- **marital-status**: Married individuals tend to have higher income, indicating marital status may influence earnings.
- **occupation**: Certain occupations are more prevalent among higher-income individuals.
- **relationship**: Married individuals and specific relationship types are more associated with higher income.



Preprocessing Details

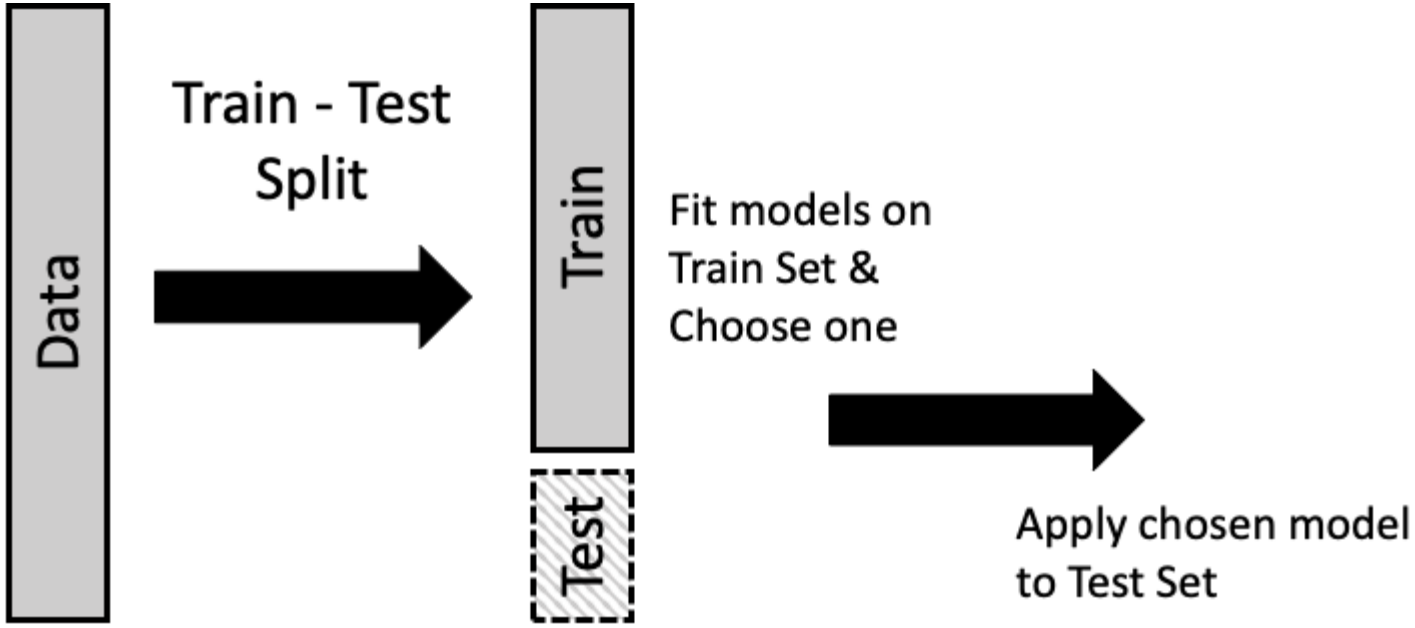
1. Label Modification:
   - Fetched Income labels = {'<=50K.', '<=50K', '>50K.', '>50K'}
   - Final modified Income labels = {'<=50K', '>50K'}
2. Dropped Duplicate Features:
   - `education` column was duplicate of `education-num`, so we dropped it.
3. Handling Missing Data:
   - Dropped all rows with missing value (around 1200 data points from 48k data points).
4. Encoding Categorical Features:
   - Applied Label encoding for features like sex, occupation, work-class, etc.
5. Normalizing Numerical Features:
   - Applied standard scaling technique to continuous numerical features.

Methodology

1. Train-Test Split
2. Model Selection and Training
3. Evaluation Metrics

Train-Test Split

- Split the data in 80:20 ratio.
  - 80% - Training Set to train the model
  - 20% - Testing set to measure model's generalization
- This split helps avoid overfitting and ensures the model is robust when facing new data.



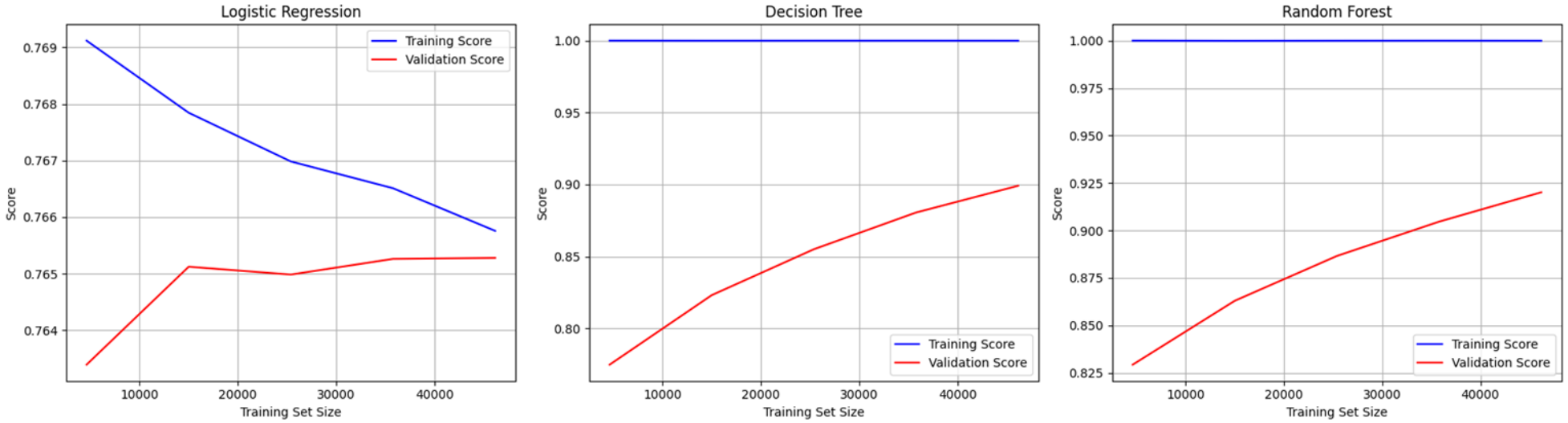
Model Selection and Training

Models:

- Logistic Regression
- Decision Trees
- Random Forest

Random Oversampling: To address the class imbalance in the dataset with individuals earning more than $50k representing 75% of the data, Random Oversampling method was used.

Bias and Variance



- Random Forest shows the best overall performance
- Decision Tree and Random Forest both suffer from overfitting but show promising validation improvements
- Logistic Regression shows underfitting, suggesting it might be too simple for this problem

Evaluation Metrics

- **Accuracy:** Random Forest achieved the highest accuracy at 93%, outperforming both Logistic Regression at 76% and Decision Trees at 91%.
- **Precision and Recall:** Random Forest again performed best, striking a good balance between precision (90%) and recall (97%).
- **F1-Score:** Random Forest had the highest F1-Score at 0.93.
- **AUC-ROC Curve:** Random Forest achieved the highest AUC (= 0.99), indicating strong predictive capabilities across varying thresholds.

Conclusion

Preliminary Conclusion: Random Forest seems to be a strong candidate for further evaluation.

Next Steps:

- Feature Selection to drop unnecessary features and simplify model.
- Hyperparameter tuning to enhance performance.
- SVMs: Exploring more complex models that may offer better accuracy

Timeline

Completed Task:

- Data collection and preprocessing.
- Exploratory Data Analysis (EDA)
- Initial Model Selection and training

Future Timeline:

- **Week 7-8:** Hyperparameter tuning and advanced model comparison.
- **Week 9-10:** Final evaluation and report preparation.
- **Week 11-12:** Presentation and Submission.

Individual Team Member's Contribution

- **Sameer Singh Godara (2022439)**: Visualization, report writing, and presentation preparation.
- **Sanyam Barwar (2022447):** Preprocessing, Exploratory Data Analysis (EDA), project management and finalization.
- **Sanyam Garg (2022448)**: Data collection, feature engineering, and model comparison.
- **Vivan Rangra (2022581)**: Model selection, initial training, and evaluation of traditional ML models.