

# AGENDA

1

Introduction

2

Methodology

3

Exploratory Analysis

4

Preprocessing

5

Modelling

6

Results

# INTRODUCTION

## Adult Census Dataset

The Adult census data is a dataset containing information on individuals in the United States, including their demographic and socio-economic characteristics, and whether their income is above or below \$50,000 per year.

## Problem Statement

Can we predict an individual's income based on their characteristics?

## Objectives

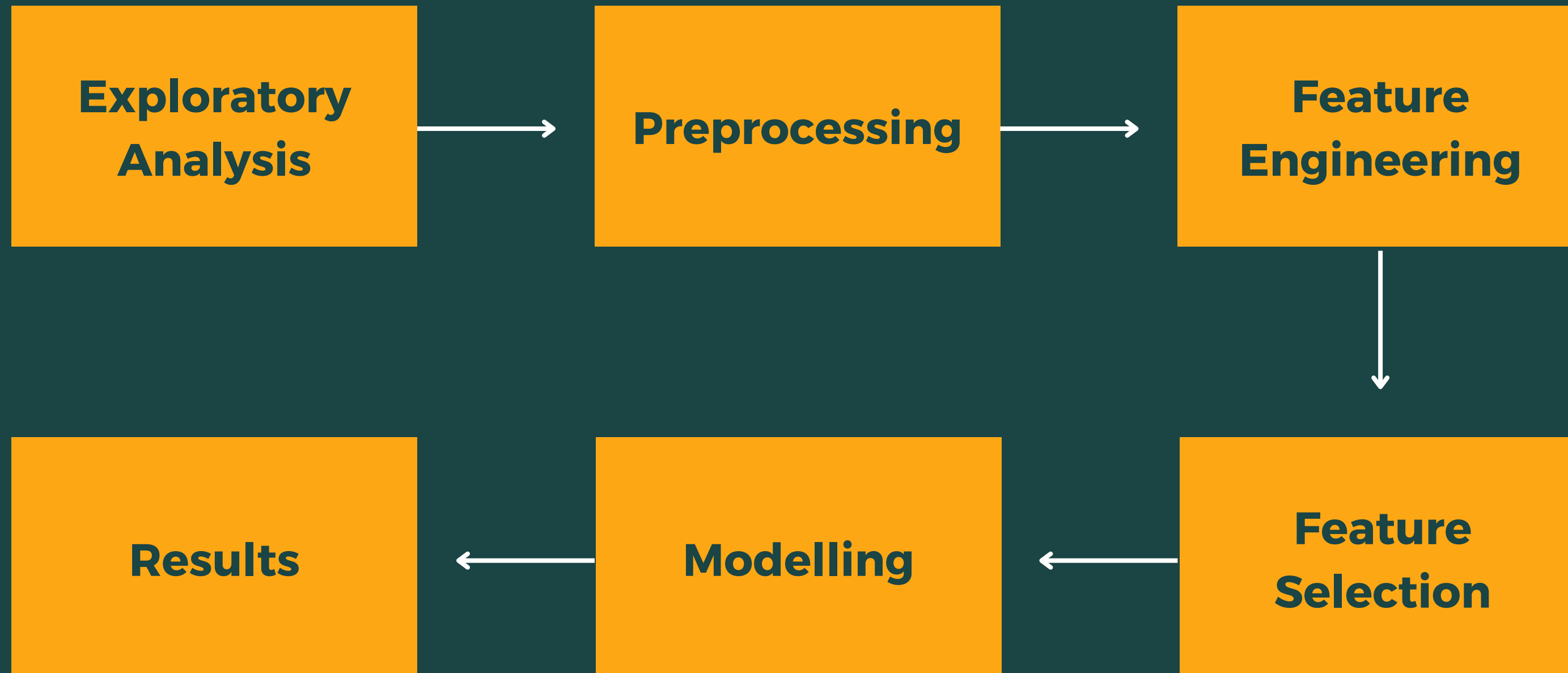
It is significant because it can help researchers better understand the factors that contribute to income inequality in Society.

## Solution

Demonstrate an effective methodology for predicting income from the Adult census data and provide insights into contributing factors.

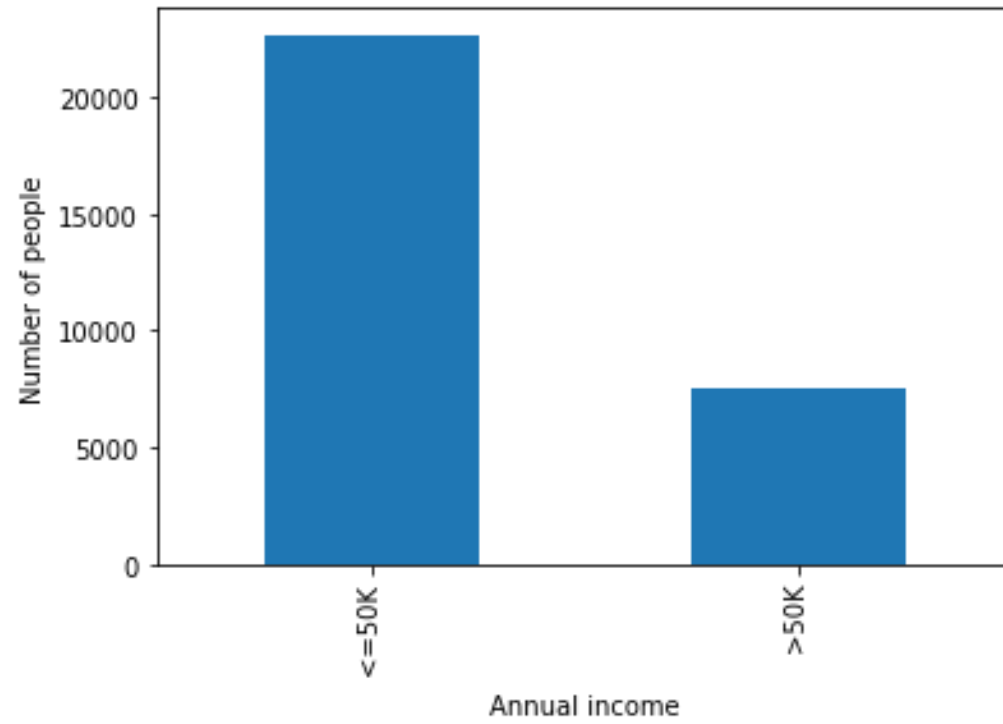


# METHODOLOGY



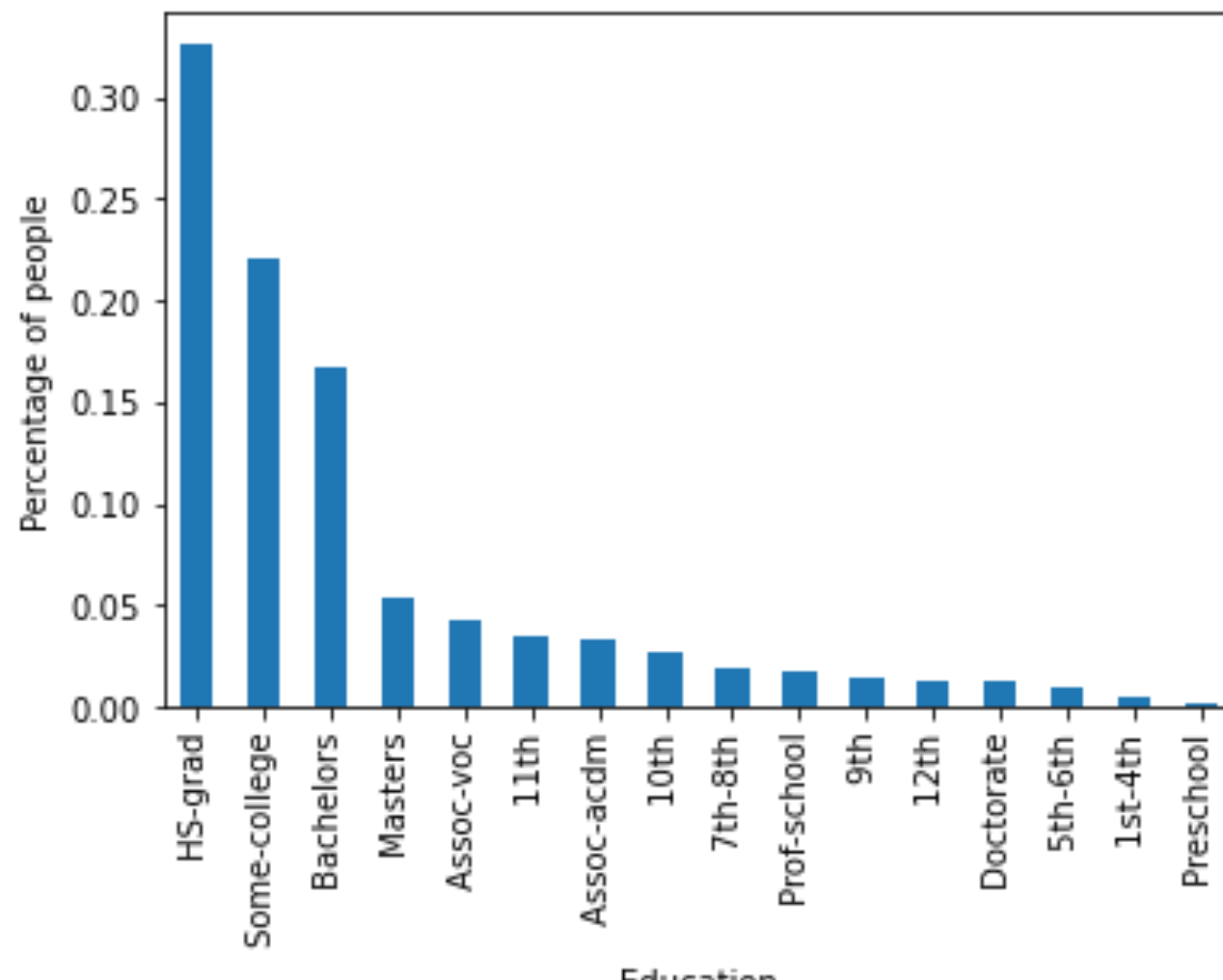
# EXPLORATORY ANALYSIS

Data distribution of feature class

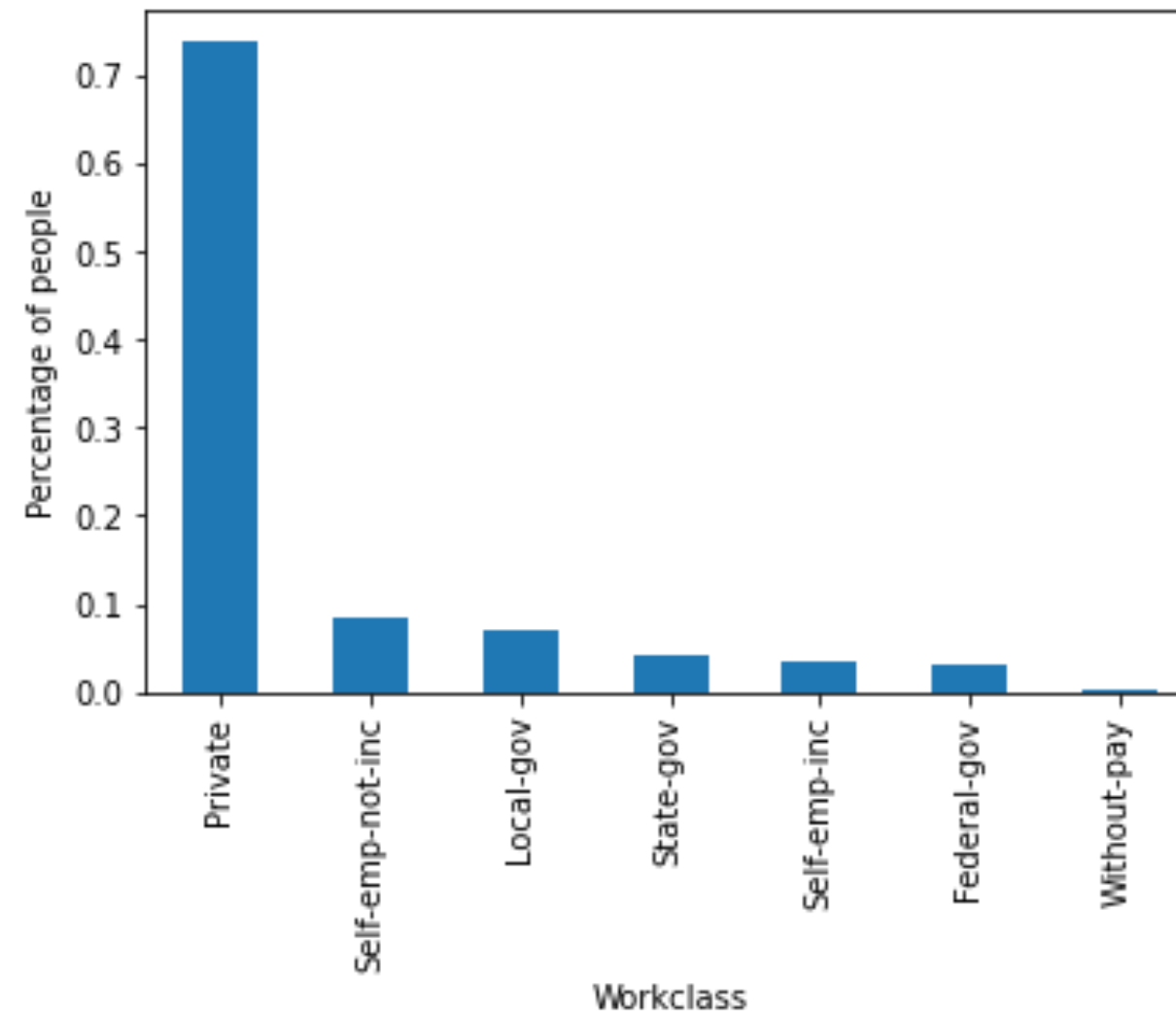


## IMBALANCED CLASSES IN ADULT CENSUS DATA

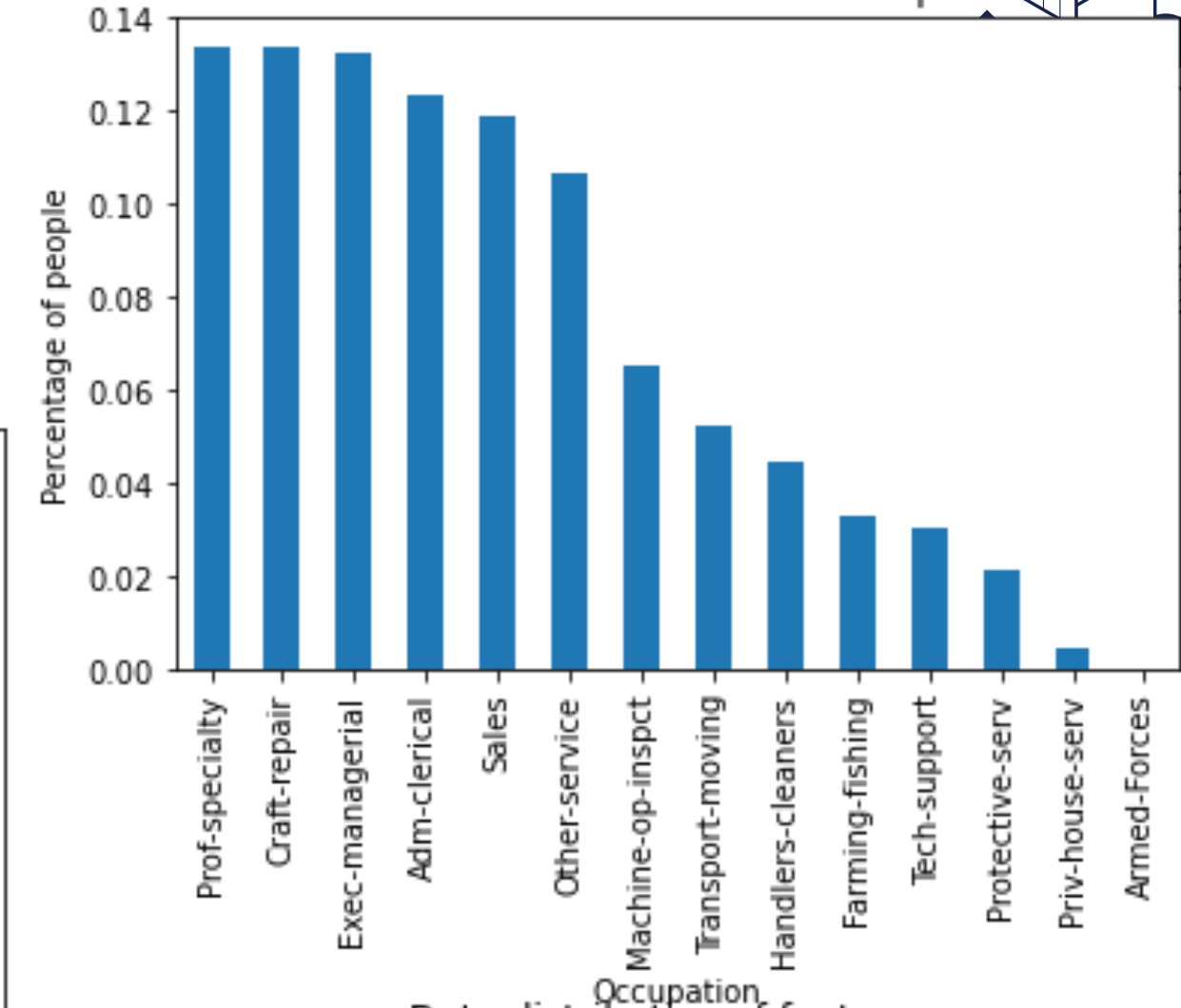
Data distribution of feature education



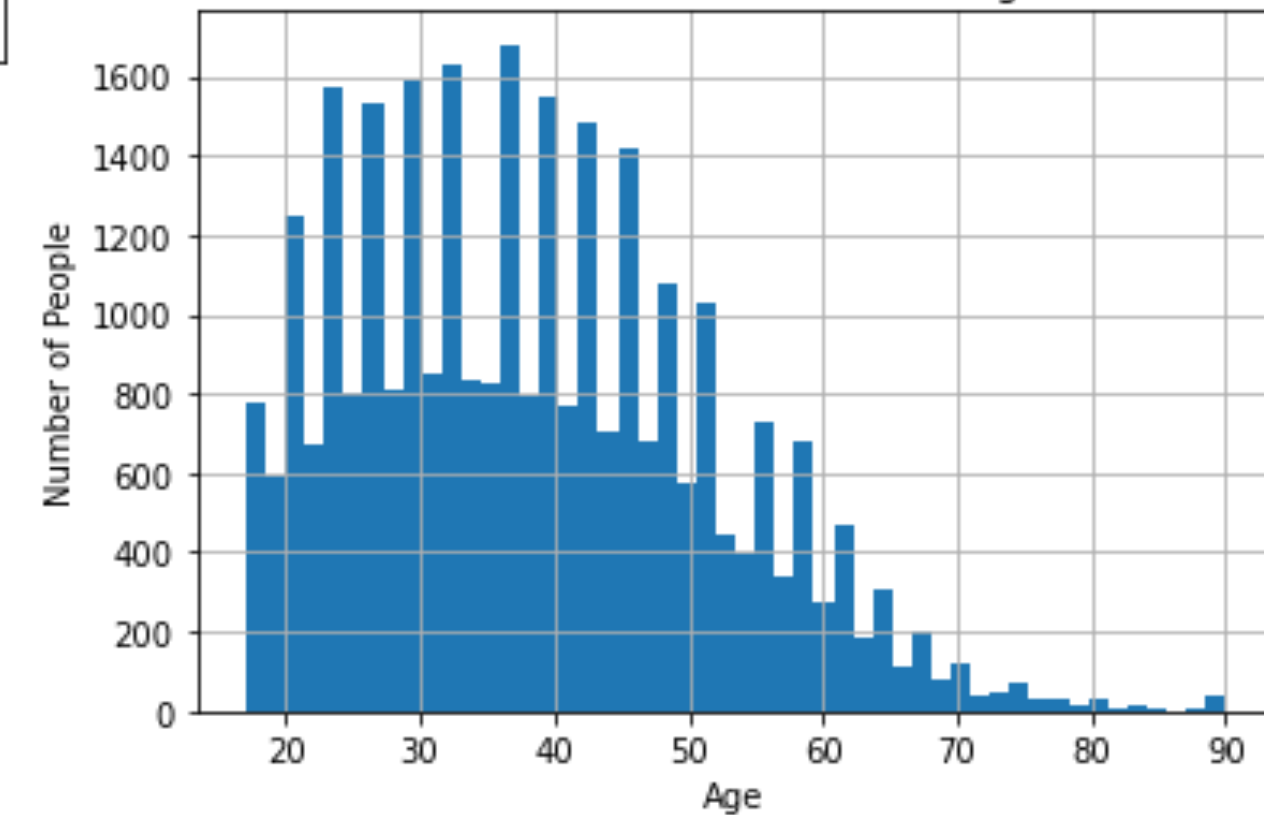
Data distribution of feature workclass



Data distribution of feature occupation



Data distribution of feature age





The diagram illustrates a database schema for a site and variable management system. The tables and their attributes are as follows:

- CATEGORY\_DESCRIPTION**
  - CATEGORY\_DESC
  - CATEGORY\_ID
  - CATEGORY\_NAME
- SITE\_DESCRIPTION**
  - SITE\_ID
  - SITE\_NAME
  - SITE\_TYPE
  - SITE\_DESCRIPTION
  - LAST\_UPDATE
- VARIABLES**
  - VARIABLE
  - VARIABLE\_ID
  - VARIABLE\_NAME
- OCCUPANT\_INFO**
  - OCCUPANT\_ID
  - SITE\_ID
  - SITE\_CODE
  - DESCRIPTION
  - TYPE
  - SITE
  - VARIABLE\_ID
  - FLAG
  - LAST\_UPDATE
- REGION\_INFO**
  - REGION\_ID
  - REGION\_NAME
  - SITE\_ID
  - REGION\_CODE
  - NAME
  - YEAR
  - HEIGHT
  - OBJ
  - ASSET
  - NOTE1
  - NOTE2
  - LAST\_UPDATE
- VARIABLES\_SAT**
  - VARIABLE\_ID
  - SITE\_ID
  - SITE\_CODE
  - DESCRIPTION
  - REGION
  - PRIORITY
  - TYPE
  - STATUS
- TRAINING\_LIST**
  - TRAINING\_LIST\_ID
  - TRAINING\_LIST\_NAME
- SITE\_TYPE**
  - TYPE\_ID
  - SITE\_ID
  - INT TYPE\_DESC
- REGION\_ROW**
  - REGION\_ID
  - SITE\_ID
  - SITE\_CODE
  - (LOW DATA)
  - OBJ
  - VARIABLE
  - TYPE
  - FLAG
- TYPE\_DESCRIPTION**
  - TYPE\_ID
  - TYPE\_CODE
  - TYPE\_NAME
  - TYPE\_SIZE
  - TYPE DESCRIPTION
  - TYPE MODE

Relationships are shown by lines connecting the tables, indicating foreign key relationships between attributes in different tables.

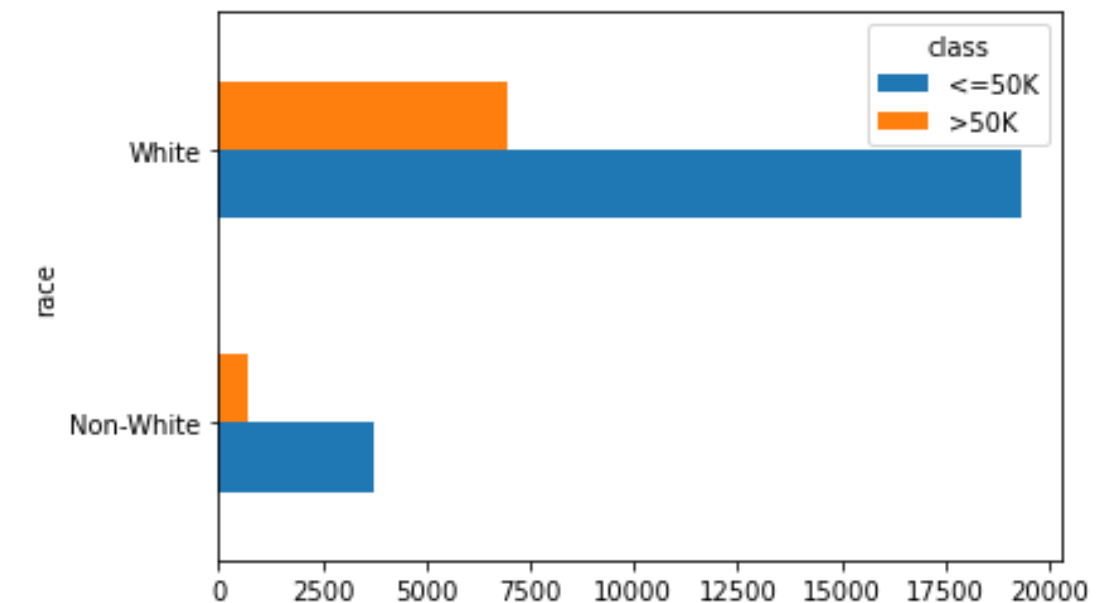
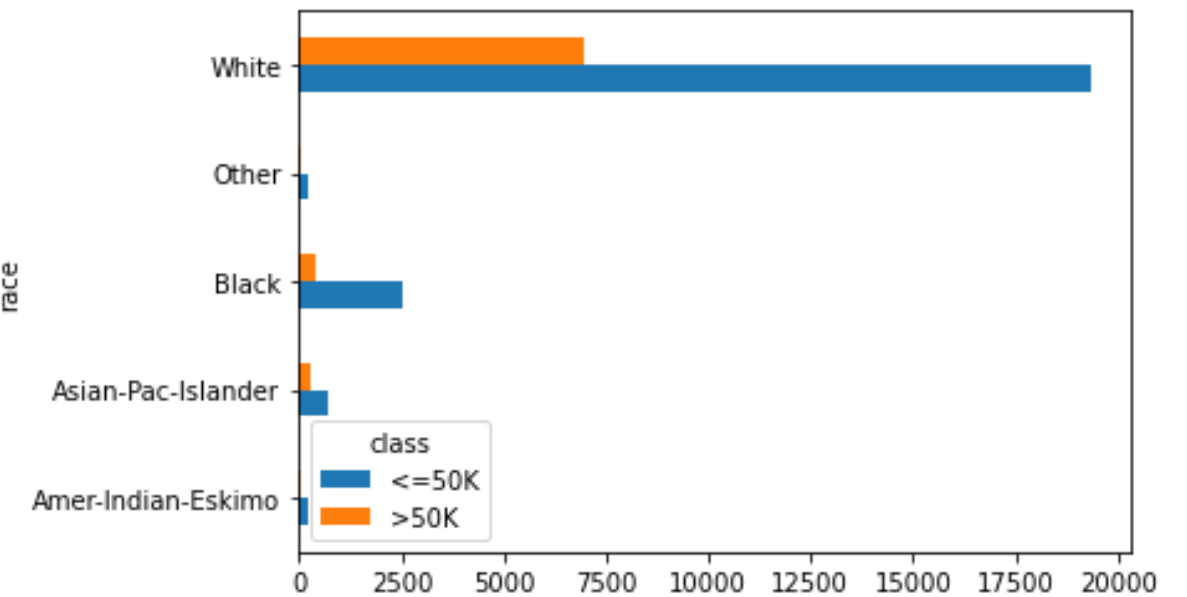
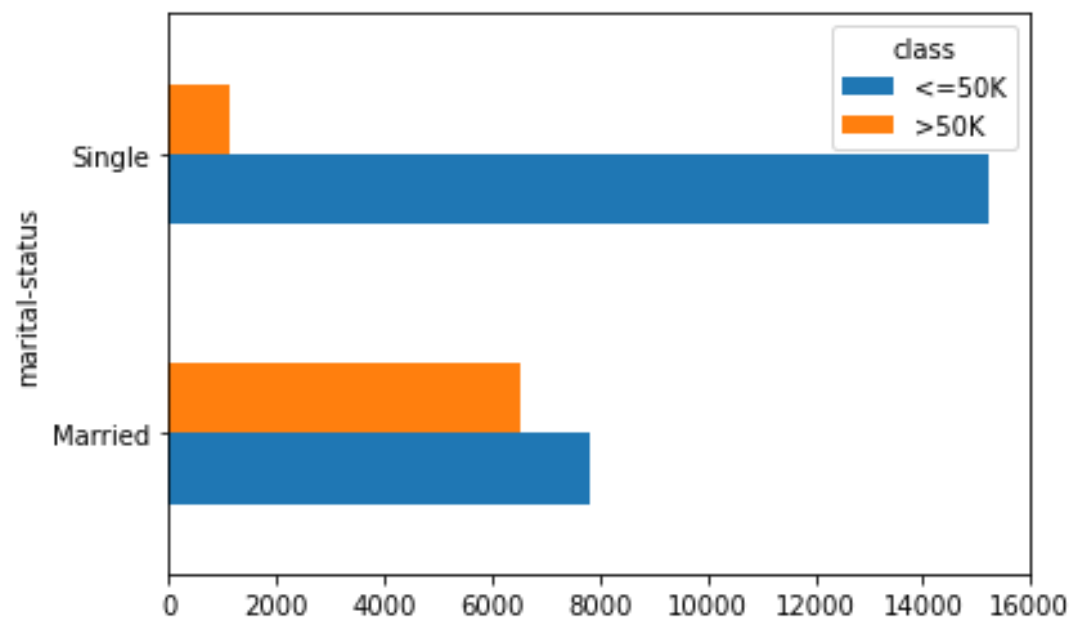
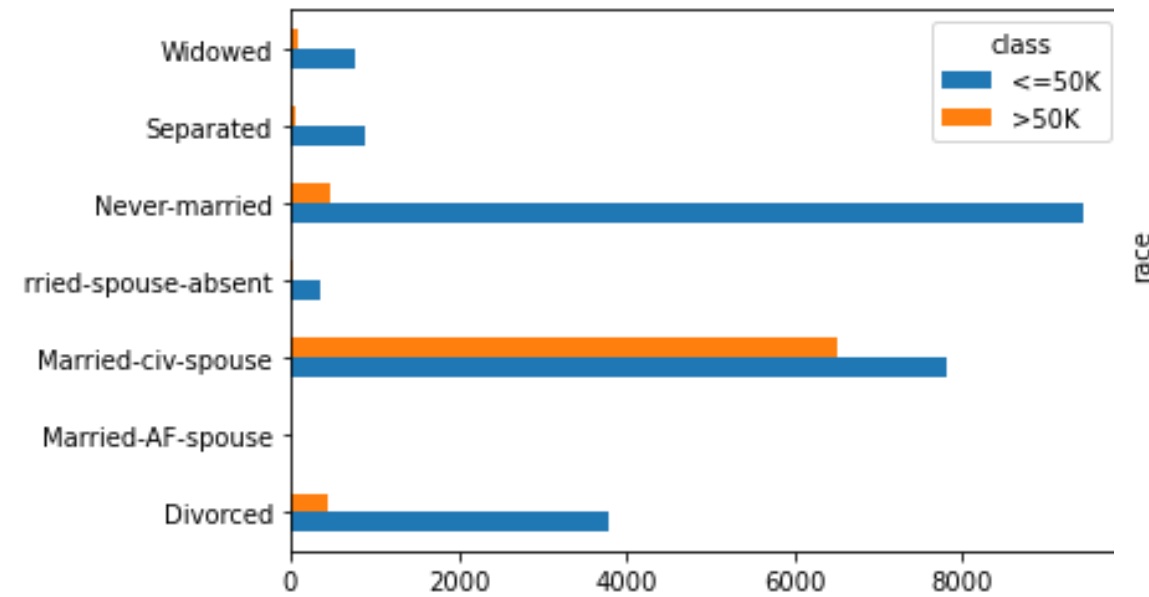
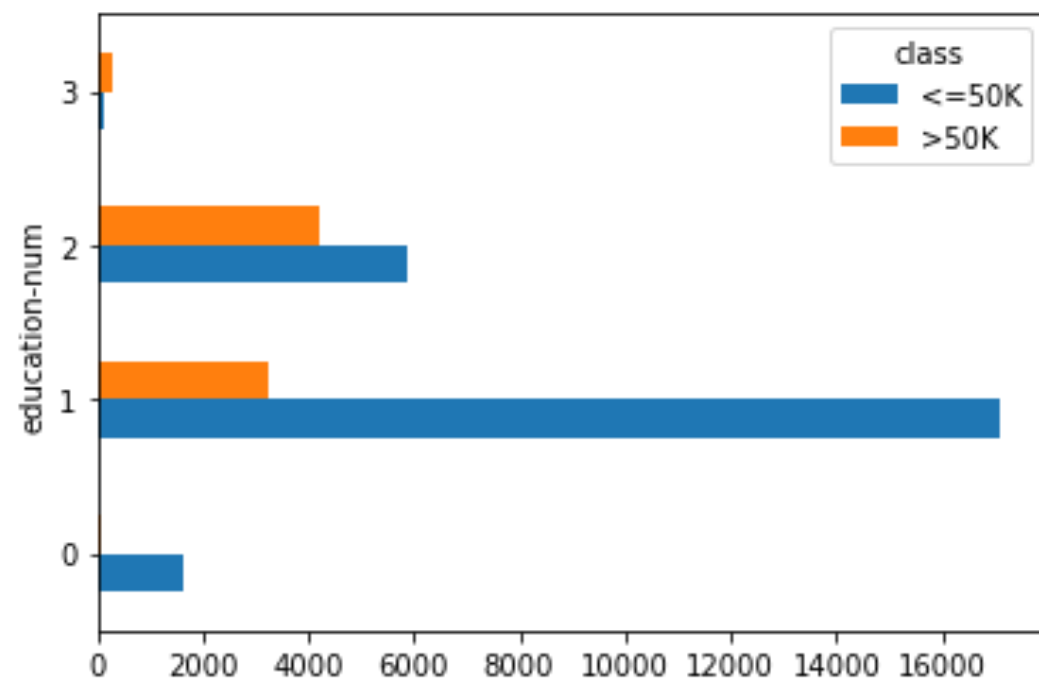
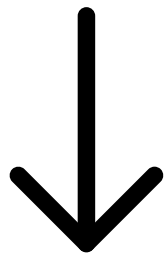
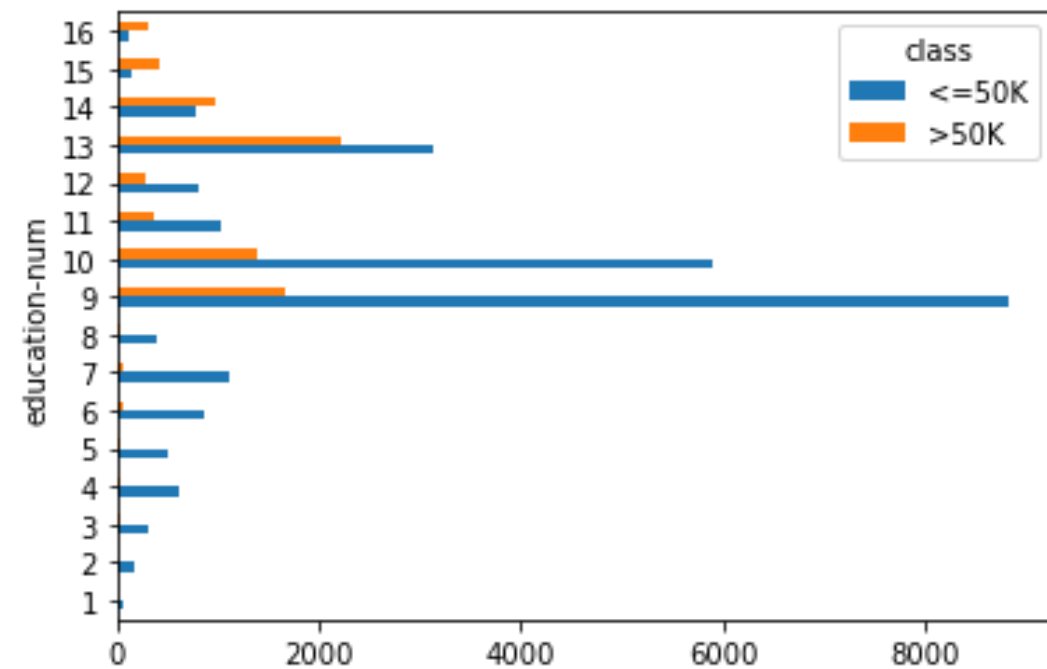
## Categorical Variables

## Feature Selection

The Adult census data contains several categorical variables, such as race and occupation, that need to be converted into numerical variables before they can be used in predictive models. This can be done using techniques such as one-hot encoding or label encoding.

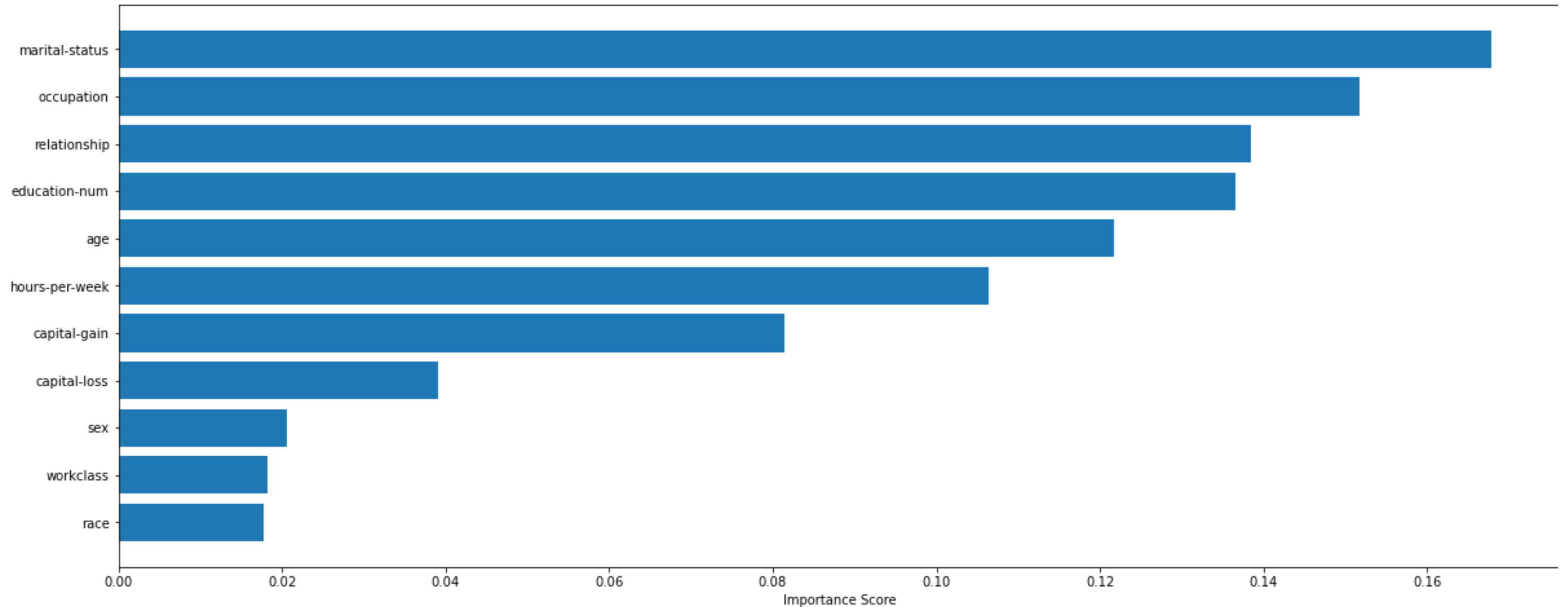
Some Algorithms such as Random Forest, Chi-square Feature Selection were done to select the most important features for the model.

# FEATURE ENGINEERING



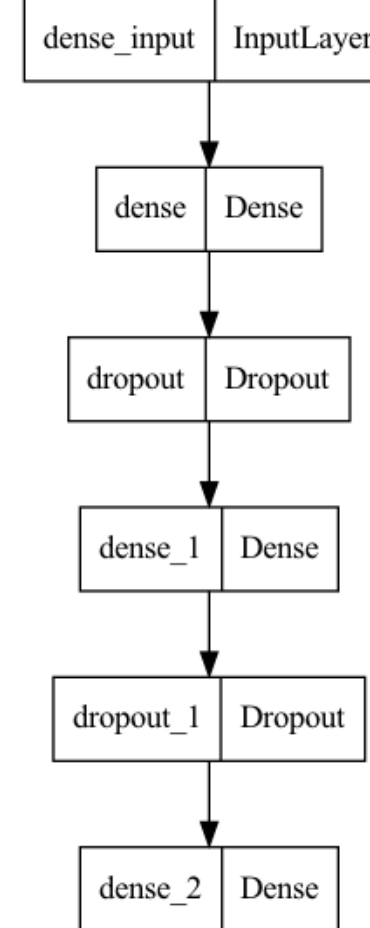
# FEATURE SELECTION

Based on the Random Forest feature importance tests, 3 common features with the least importance are sex, race, and work class.

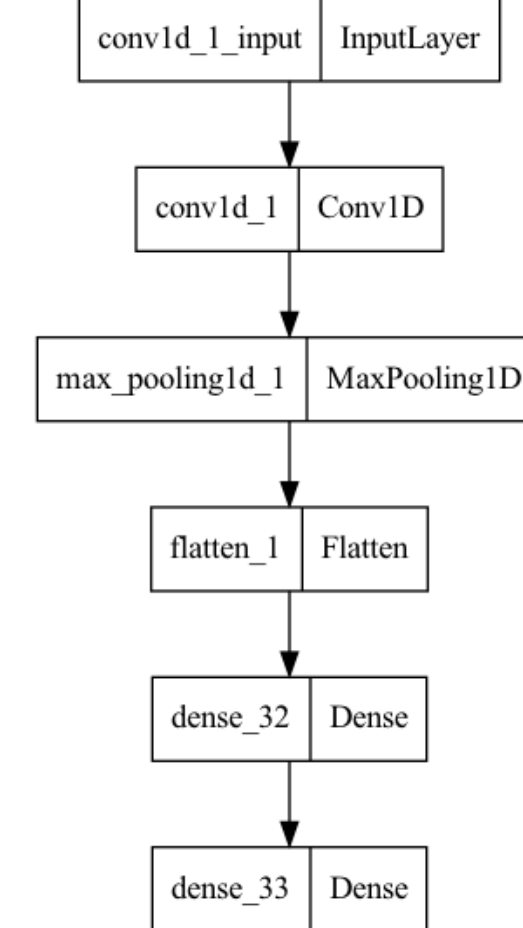


# MODELLING

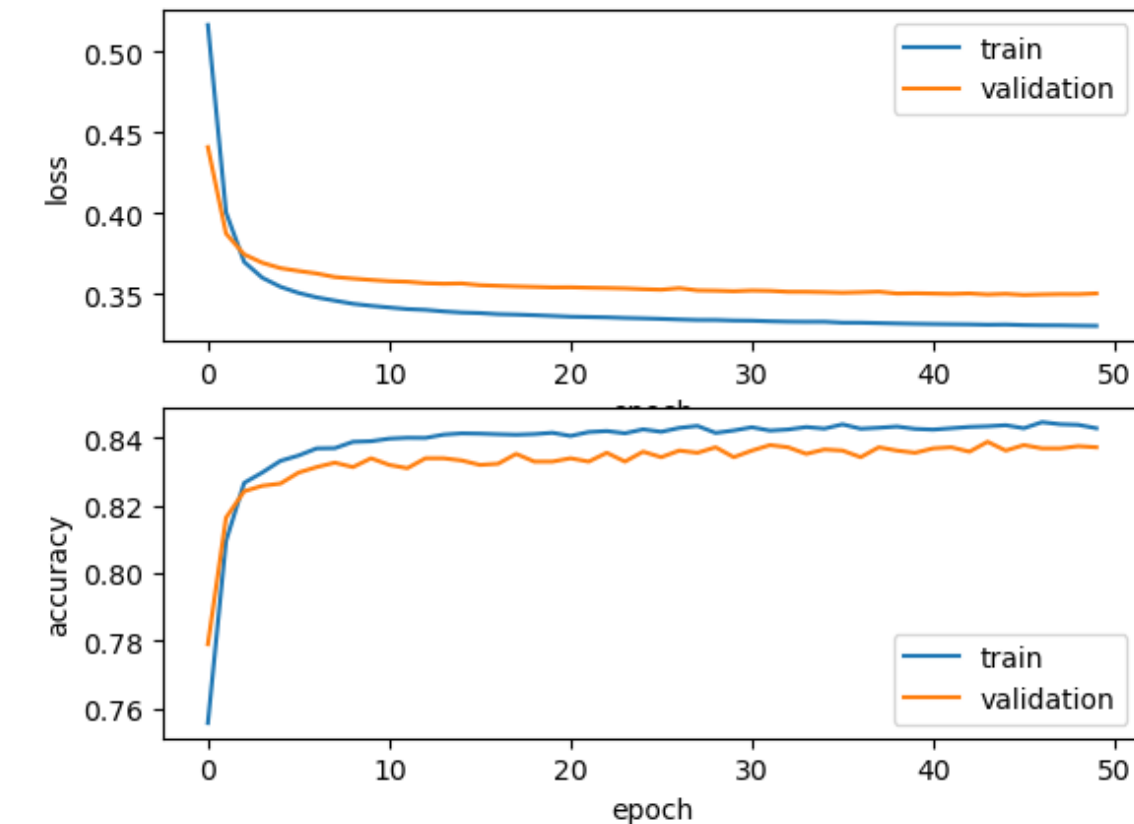
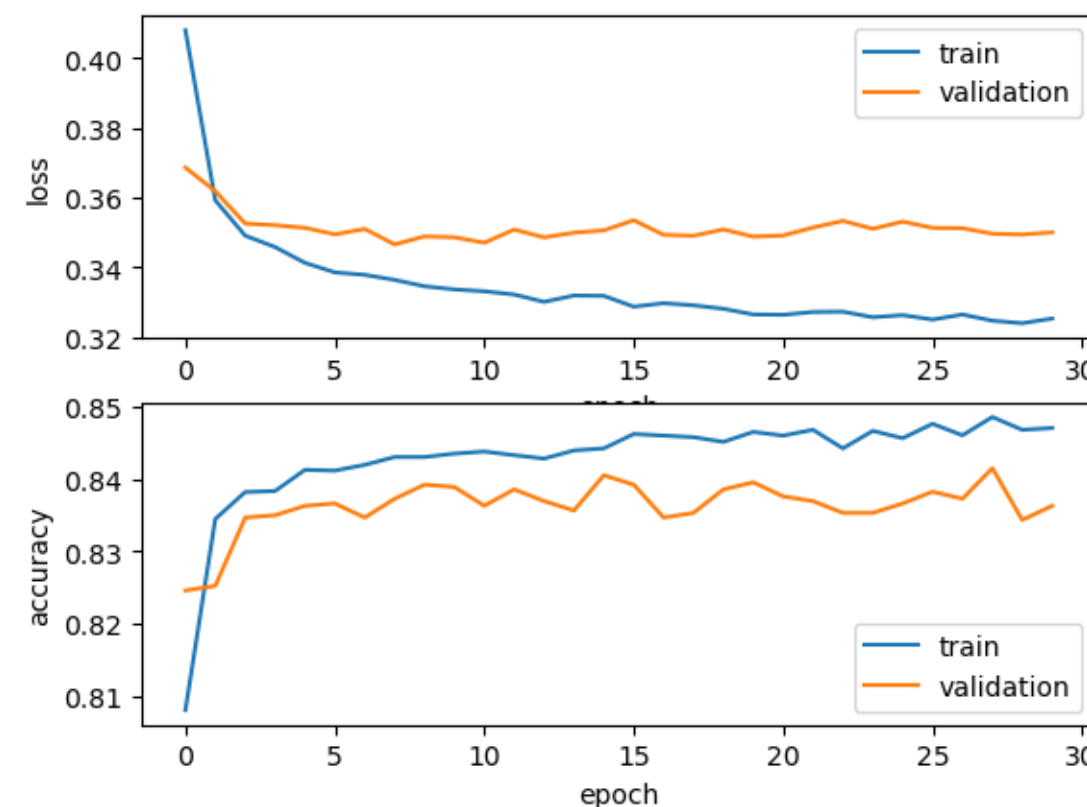
- Various neural network models were explored. DNN and CNN were the most common architectures.
- Model performance was evaluated using accuracy, roc-auc score, precision, recall, and F1 score.
- Models were optimized through hyperparameter tuning and feature selection.
- Models were trained using GPUs on Google Colab and Kaggle to utilize parallelisation.



**Dense Neural Network  
Architecture**



**Convolution Neural Network  
Architecture**





# RESULTS FOR IMBALANCED TRAINING SET

**Default Threshold (0.5)**

**Accuracy Score: 0.85**

**Precision: 0.71**

**Recall: 0.59**

**F1: 0.64**

**AUC: 0.75**

**Best Threshold = 0.212586**

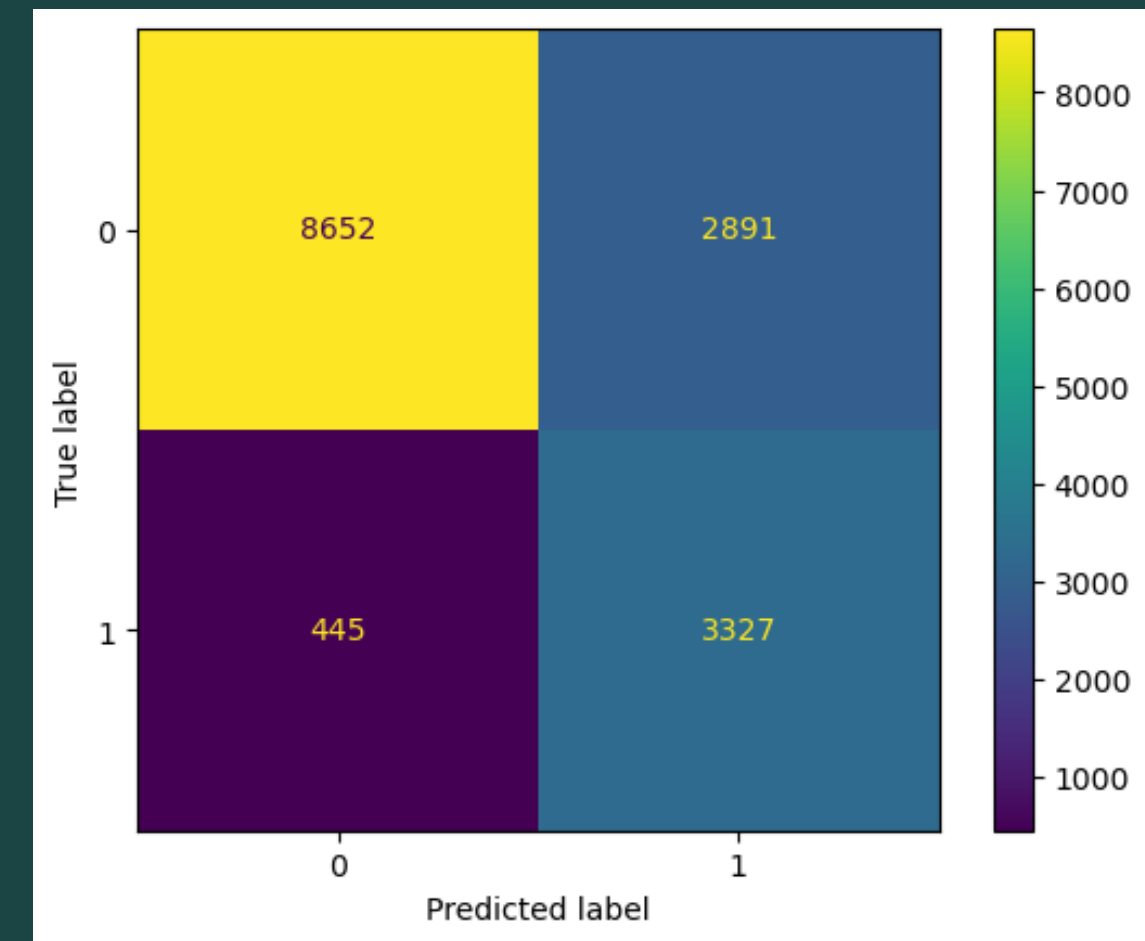
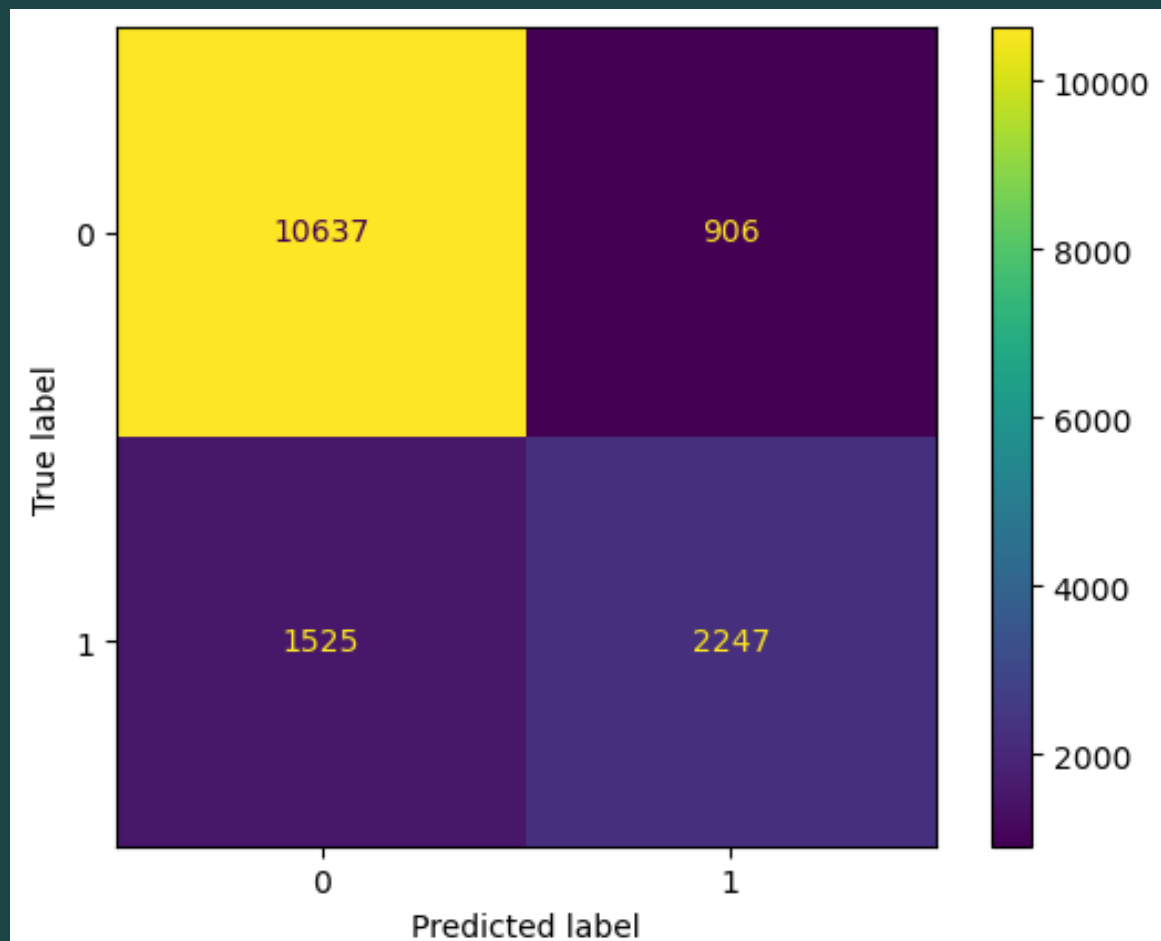
**Accuracy Score: 0.78**

**Precision: 0.53**

**Recall: 0.88**

**F1: 0.66**

**AUC: 0.81**



# RESULTS FOR OVERSAMPLED TRAINING SET

**Default Threshold (0.5)**

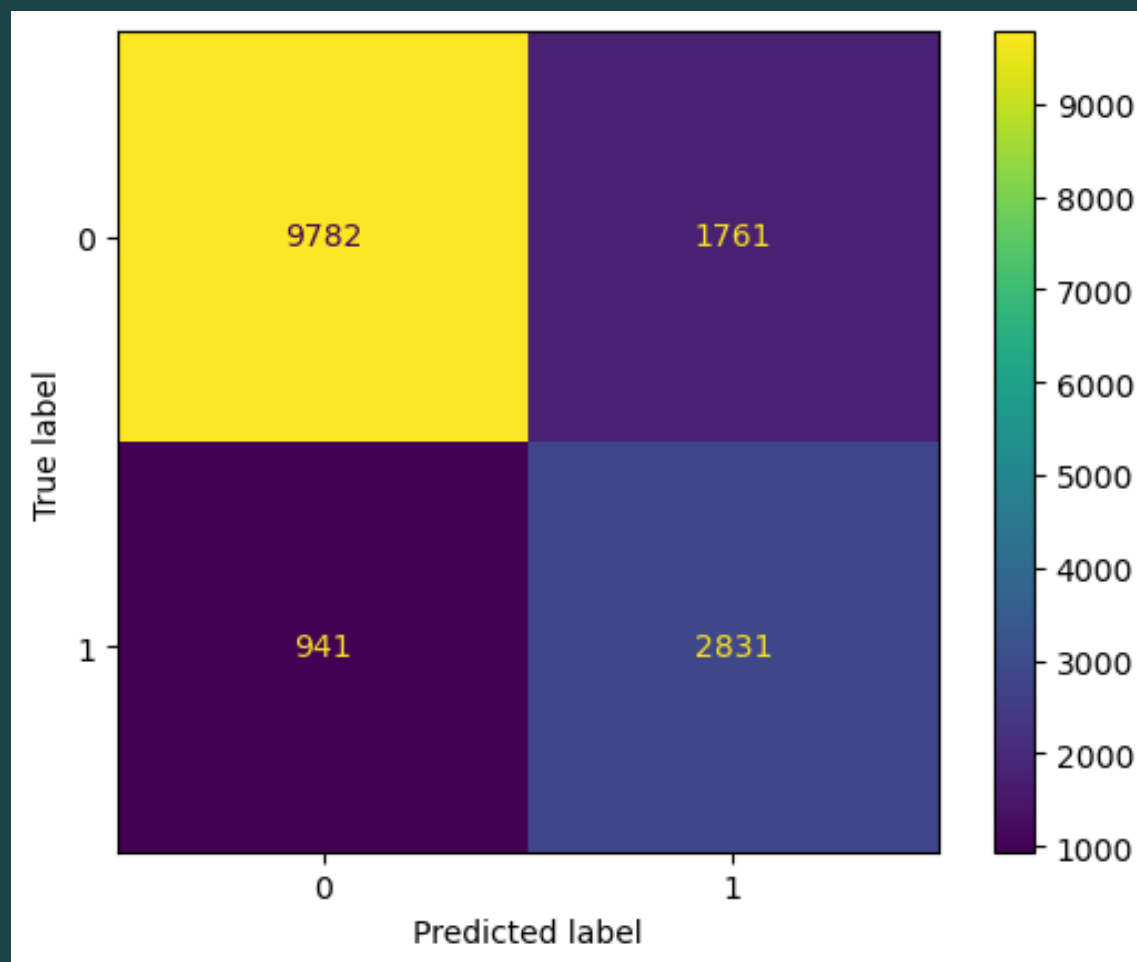
**Accuracy Score: 0.82**

**Precision: 0.61**

**Recall: 0.75**

**F1: 0.68**

**AUC: 0.80**



**Best Threshold = 0.405082**

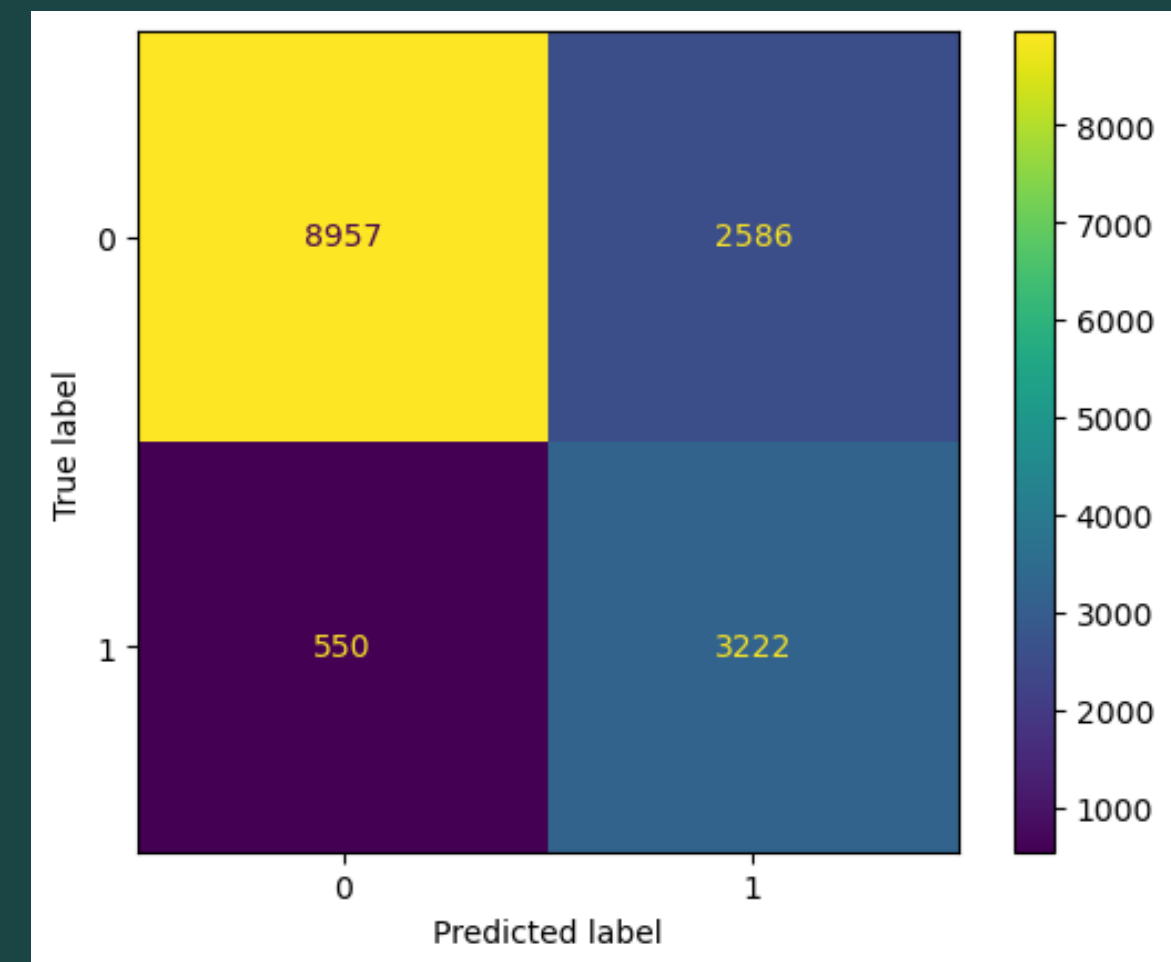
**Accuracy Score: 0.79**

**Precision: 0.55**

**Recall: 0.85**

**F1: 0.67**

**AUC: 0.82**



# CONCLUSION

## Main Problem

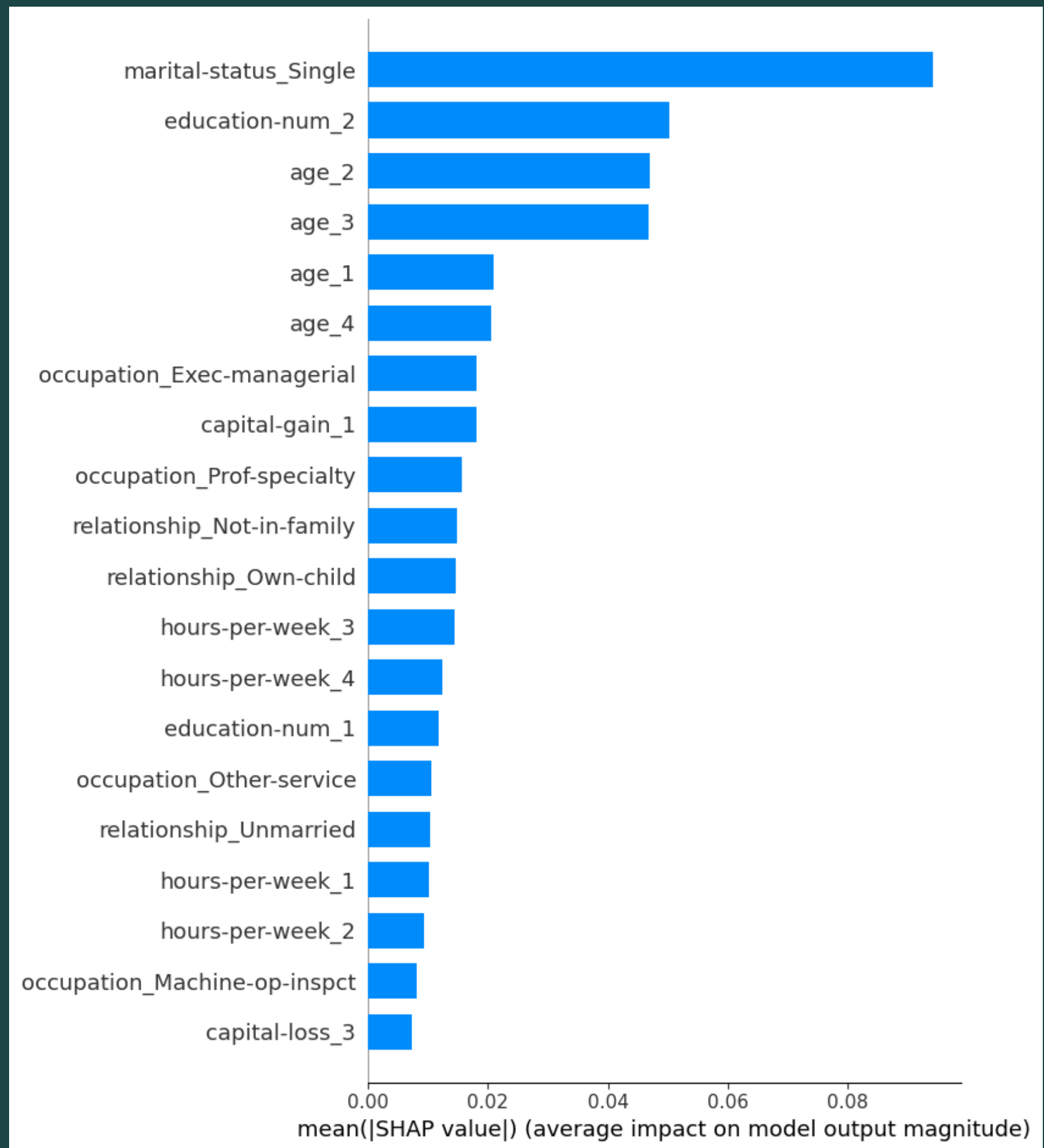
Can we predict an individual's income based on their characteristics?

## Why is it important?

The factors that contribute to it can help us design policies to address income inequality.

## Conclusion

We can observe that factors such as marital status, education, age and occupation are important in determining income.



## MODEL INTERPRETATION USING SHAP VALUES