# Adult Income Predictor Report

## Table of contents

- DSCI 522 - Workflows
- MDS 2024-2025
- Group 24
- Members: Michael Suriawan, Francisco Ramirez, Tingting Chen, Quanhua Huang

## Summary

This report presents the application of a K-Nearest Neighbors (KNN) Classifier to predict an individual's annual income based on selected categorical socioeconomic features from the Adult dataset. The dataset, sourced from the 1994 U.S. Census Bureau by Becker and Kohavi

(1996), contains 48,842 instances and features such as age, education, occupation, and marital status. The model achieved an accuracy of approximately 80%, with a tendency to predict more individuals with incomes below $50K compared to those above. This result emphasizes the importance of socioeconomic factors in determining income levels. However, the findings are limited by the exclusion of numerical features such as age and hours-per-week, which could provide additional predictive power. Further investigation into individual feature contributions and the inclusion of numerical variables like age and hours-per-week could enhance prediction performance.

### Introduction

The Adult dataset, originally curated from the 1994 U.S. Census Bureau database, is a well-known benchmark dataset in machine learning. Its primary objective is to predict whether an individual earns more or less than $50,000 annually based on various demographic and socio-economic attributes. With 48,842 instances and 14 features, the dataset encompasses a mix of categorical and continuous variables, making it a rich resource for classification tasks and exploratory data analysis.

Understanding the factors influencing income levels is important for addressing socioeconomic disparities, informing policy decisions, and developing targeted programs to support underprivileged groups. This problem is particularly interesting as it highlights how demographic and socioeconomic variables interact and impact an individual's earning potential, offering insights into broader societal trends.

The model described in this notebook looks to use a trained "Nearest Neighbors" Classifier to use different socioeconomic features to predict the range of the individual's income. The features in the data set include characteristics such as age, education level, marital status, occupation, among others.

The model looks to predict whether an individual's income exceeds $50K/yr based on the selected categorical socioeconomic features. For simplicity, only selected categorical features from the original data set. These features are specifically encoded based on their content prior to training the kNN classifier used for predictions.

## Data Validation

To ensure that the analysis is not influenced by erroneous data, the inputs to this analysis went through exhaustive validation to avoid any influence of low-quality data in the results of the analysis. Validation process ensured correct column names, no empty observations, missing data below expected thresholds, correct data types, no duplicate observations, no outlier or anomalous values & ensuring correct category levels.

# EDA Analysis

## Data Summary

In our analysis, we focus on the categorical socioeconomic features within the dataset. There are eight categorical features in total; however, we have excluded education and native country from our analysis to minimize potential bias and refine our focus. This decision was made to ensure that the remaining features are more directly relevant to our study objectives.

The exploratory data analysis (EDA) begins with histograms of the six selected categorical features: marital-status (Figure 1), relationship (Figure 2), occupation (Figure 3), work class (Figure 4), race (Figure 5), and sex (Figure 6). These histograms provide a clear view of the distribution of each feature and offer valuable insights into the balance and variability within the dataset, informing the subsequent modeling and analysis phases.

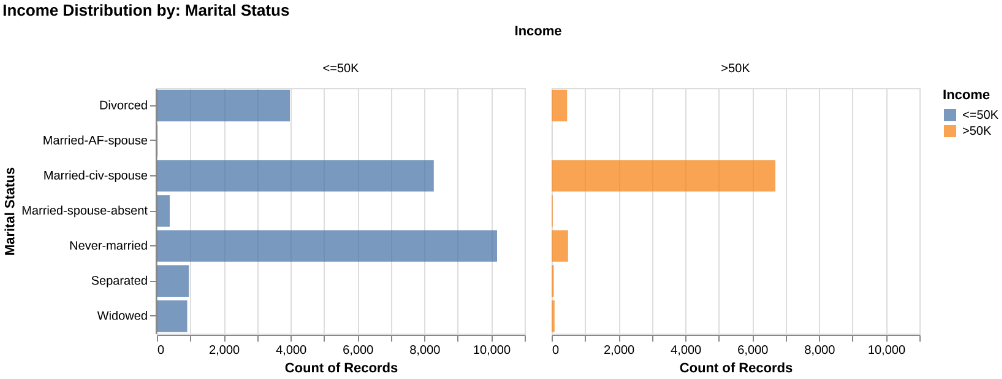

Figure 1: Income Level by Marital Status - Highest count of >50K earners are married under civil law.

Note: The process to create the plots above was inspired by Ostblom (2023).

# Model Development and Evaluation

## Train/Test Split

In preparation for fitting a KNN model, the data was divided into an 80/20 training/test data split. A random seed was used in the splitting process to ensure reproducibility of the analysis.

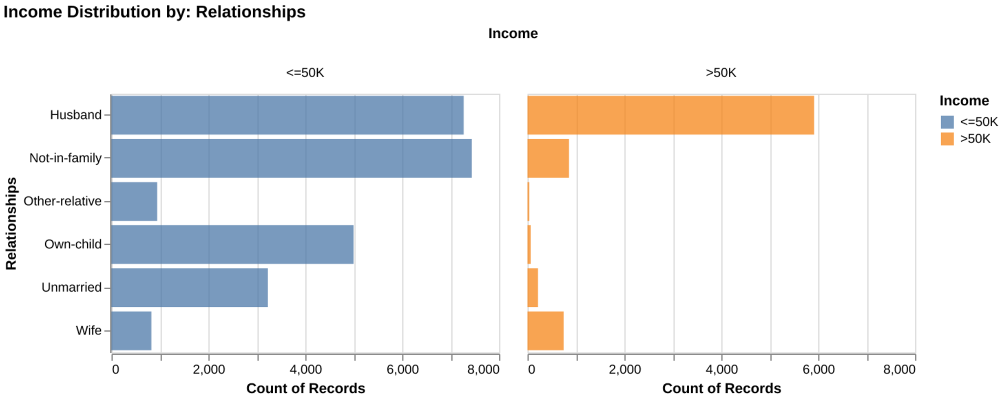

Figure 2: Income Level by Relationship Status - Highest count of >50K earners have a role as husbands in their relationships.

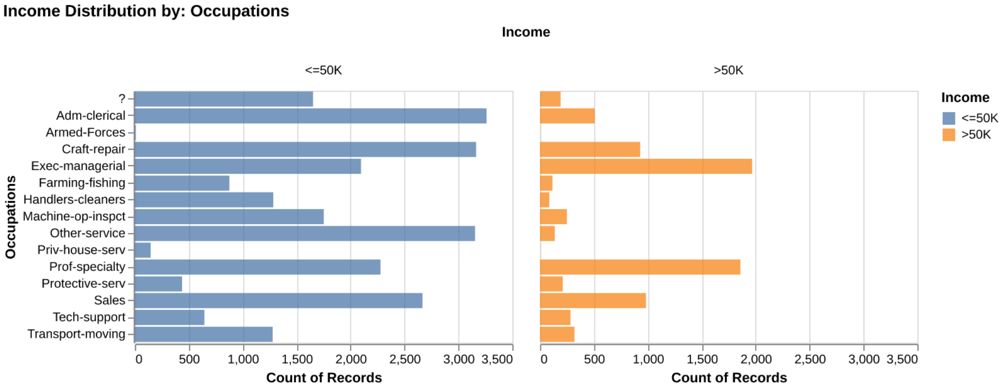

Figure 3: Income Level by Occupation - Highest count of >50K earners have executive or managerial positions in their occupations or have a specialty in their professional field.

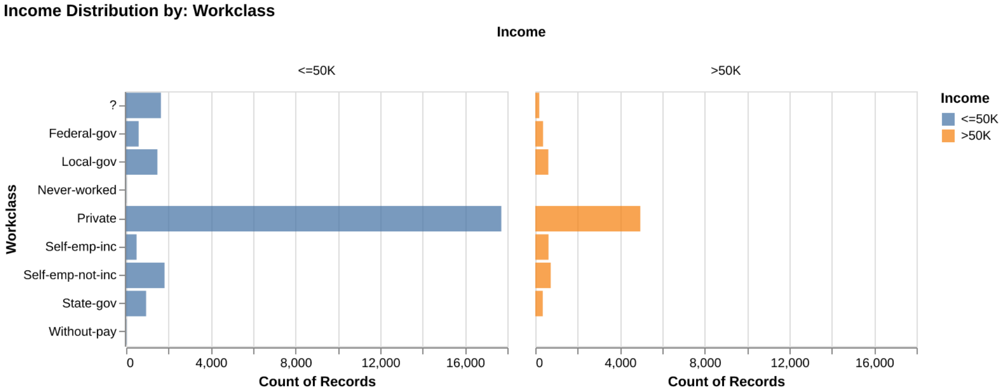

Figure 4: Income Level by Workclass - Highest count of >50K earners are in the private sector.

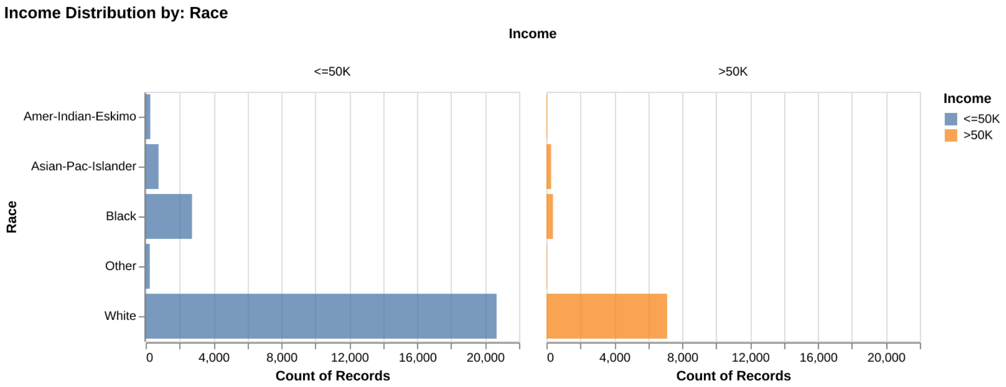

Figure 5: Income Level by Race - Highest count of >50K earners are primarily white.

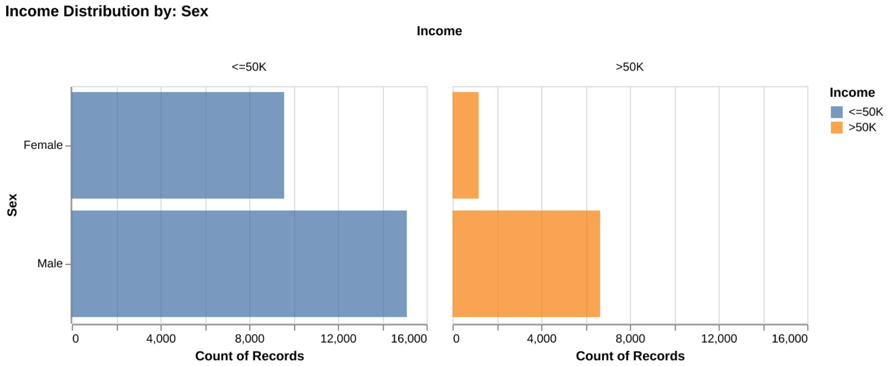

Figure 6: Income Level by Sex - Highest count of >50K earners are primarily male.

## Deep Checks Validation

The following deepchecks validations were performed to ensure quality of the training data:

- No anomalous correlations between target/response variable and features/explanatory variables
- No anomalous correlations between features/explanatory variables

## Feature Selection

For simplicity, the model is focused on using categorical variables available in the data set.

The used features are the following:

As categorical features: "marital-status", "relationship", "occupation", "workclass", "race". As binary features: "sex".

## Preprocessing

The data was preprocessed using One Hot Encoder to encode categorical features, as well as using a Simple Imputer to deal with any missing data in the data set.

## Model Fit

A pipeline was created to describe the preprocessing and KNN modeling steps that were used to train the model.

We chose K-Nearest Neighbors (KNN) for analyzing the income prediction task because of its simplicity, flexibility, and effectiveness in handling non-linear relationships in the data. KNN is an intuitive, non-parametric model that makes predictions based on the closest neighbors of a data point, which allows it to adapt well to complex and non-linear patterns without assuming any specific distribution of the data. Unlike linear models such as logistic regression, which assume a linear relationship between features and the target variable, KNN does not impose such constraints, making it more suitable for datasets where the relationship between variables is more complex.

One of the main advantages of KNN is its minimal preprocessing requirements. Unlike many other models that require significant data transformation, such as feature scaling or normalization, KNN performs well with raw data as long as the features are comparable in scale, where we already applied StandardScaler in preprocessing. Additionally, KNN can be very effective with smaller to medium-sized datasets, which is often the case when dealing with real-world classification tasks. It also performs well in multi-class classification problems, such as predicting income levels, where the goal is to classify a data point into one of several categories (e.g., <=50K or >50K).

Another significant reason for choosing KNN is its interpretability. KNN provides a simple explanation for its predictions—by showing which data points are most similar to the one being classified. This makes it easy to communicate the reasoning behind each prediction to stakeholders. Unlike more complex models like decision trees or neural networks, KNN doesn't rely on intricate structures, and its decision-making process is intuitive. It is also less prone to overfitting compared to highly complex models, as it doesn't require learning a complicated set of parameters but instead depends on local patterns within the data.

While other methods, such as logistic regression, support vector machines, and decision trees, could have been used, they come with certain limitations. For instance, logistic regression assumes a linear relationship between features and outcomes, which is often not the case in complex datasets like income prediction in our project. Decision trees can easily overfit, especially if the tree is too deep or not pruned correctly. Support vector machines require careful parameter tuning and may struggle with larger datasets. In contrast, KNN provides a straightforward, flexible, and interpretable approach that works well for the income classification task at hand, particularly given the non-linear and diverse nature of the relationships in the dataset.

In summary, KNN was selected for this task because it provides a simple, effective, and interpretable model that adapts to the non-linear structure of the income prediction problem without the need for complex preprocessing or parameter tuning. It is particularly useful when the data is small to medium-sized, and the goal is to quickly and accurately classify data based on local patterns.

## Model Test Score and Prediction

Finally, the model is scored on the unseen examples.

Additionally, Figure 7 displays the hard predictions the model does on the test data.

The test score was 0.781.

The Confusion Matrix above displays the types of errors that the model is making. While accuracy is being used as primary scorer, the Confusion Matrix offers information on precision and recall, should they be selected as primary scorers in future developments.

An additional characteristic of the dataset that can be visualized through the confusion matrix is that there is significant class imbalance, as the vast majority of examples in the data set belong to the "<=50K" class. This could be taken into account in future iterations of this analysis for potential score improvement.

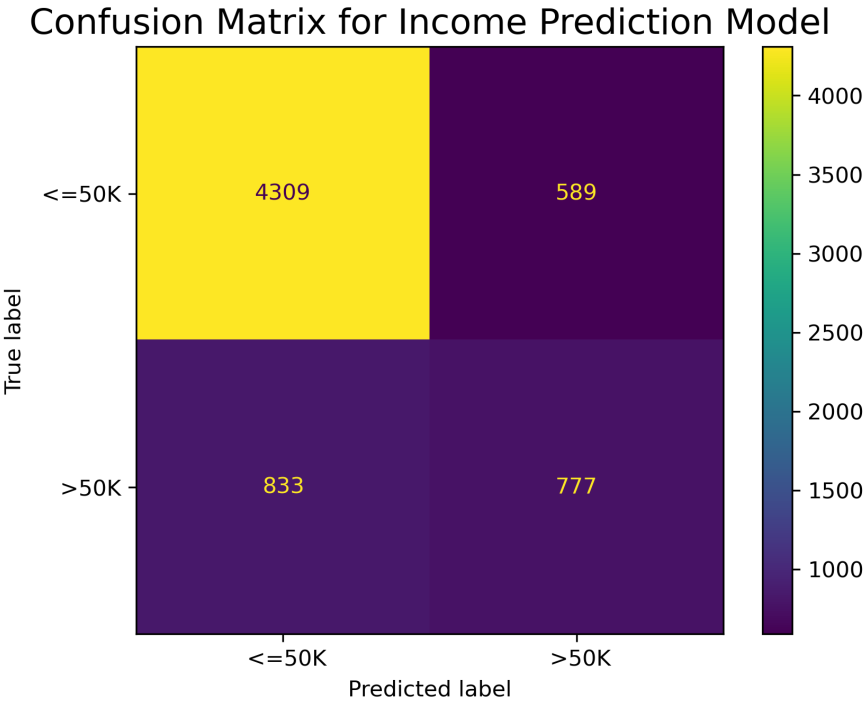

Figure 7: Confusion Matrix

The process followed for creating the model was inspired by Kolhatkar (2023) and Pedregosa et al. (2011)

## Discussion

The KNN model described in this report is able to predict the income of an individual based on the described categorical features with an accuracy of ~80% as seen in the training and test scores.

It was expected that selected categorical features would influence the income range for individuals, particularly those related to occupation and education level.

With above histograms, we notice that our KNN model predicts more individuals to have income that is less than 50K and predict less individuals to have more than 50K income, comparint to the actual results.

These findings support the notion that specific socioeconomic characteristics of individuals have a direct influence on the individual's income level.

However, this analysis opens the question on how each individual feature affects the model. Therefore, further deep-dive could better inform if all features have a significant influence on the model's ability to predict accurately. Additional numerical features, such as age and hours-per-week are likely to improve the model training process and could be evaluated as well.

## References

Becker, Barry, and Ronny Kohavi. 1996. "Adult." UCI Machine Learning Repository.

Kolhatkar, Varada. 2023. *DSCI 571 Supervised Learning 1*. https://pages.github.ubc.ca/MDS-2024-25/DSCI_571_sup-learn-1_students/lectures/notes/00_motivation-course-information.html.

Ostblom, Joel. 2023. *DSCI 531 Data Visualization 1*. https://pages.github.ubc.ca/mds-2024-25/DSCI_531_viz-1_students/intro.html#.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, et al. 2011. "Scikit-Learn: Machine Learning in Python." *Journal of Machine Learning Research* 12 (85): 2825–30. http://jmlr.org/papers/v12/pedregosa11a.html.