Income – Classification Analysis

Introduction:

Income classification is a critical task that holds great significance in today's society. An individual's income level not only affects their quality of life but also plays a pivotal role in various economic and social decisions. Understanding the factors that contribute to high or low income can empower individuals, businesses, and governments to make more informed choices.

In this project, we will delve into the domain of income classification, focusing on a binary classification problem with the goal of predicting whether an individual's income exceeds $50,000 per year or not. To accomplish this, we will leverage a diverse set of features, ranging from personal attributes such as age, education, and occupation to financial aspects like capital gains and losses. These features will serve as the basis for our predictive model.

Task: The task of our income classification project is defined as follows:

1. **Binary Classification**: The primary objective of this project is to develop a model that can categorize individuals into one of two income classes: "> 50 K" or "<= 50 K". This binary classification task simplifies the prediction of income levels.
2. **Feature Set**: The features used for classification encompass a variety of personal, educational, and financial attributes. These features include '**age**', '**workclass**', '**education**', **'marital-status'**, '**occupation**', and more. The selection of these features is driven by the belief that they hold valuable information for predicting income levels.
3. **Dataset Source**: The dataset used in this project is sourced from Kaggle. This dataset has been curated to serve as the foundation of our analysis and modeling efforts. It offers a real-world representation of income-related factors, making our project both practical and applicable.
4. **Modeling Techniques**: Throughout this project, we will explore and implement various deep learning techniques and algorithms to construct a robust predictive model. The selection of methods will be influenced by our pursuit of identifying the most effective approach for the income classification task. Deep learning, with its ability to capture complex patterns and relationships in data, holds the potential to offer valuable insights into income classification, and we aim to harness this power in our analysis.
5. **Data Preprocessing and Analysis**: Data preprocessing will involve tasks such as handling missing values, encoding categorical variables, and scaling features. We will also perform data analysis to gain insights into the dataset's characteristics and distributions.
6. **Performance Evaluation**: We will evaluate the model's performance using standard binary classification metrics, including Accuracy, F1 score, Precision and Recall, and Confusion matrix. This will enable us to assess the model's effectiveness in predicting income levels.

Data Source:

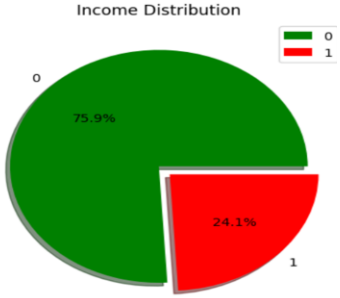
The dataset was collected from Kaggle, a popular platform for data science and machine learning datasets.
Dataset name: Income-classification
Dataset link: https://www.kaggle.com/datasets/lodetomasi1995/income-classification

Data Description:

- Totally there are 14 input features and target class label in the dataset:

Income: 50K, <=50K.
age: continuous.
workclass: Private, Self-emp-not-inc, Self-emp-Inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
fnlwgt: continuous, it indicates the number of people the census believes the entry represents.
education: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.
education-num: continuous. It is code number for different values in 'education' column. It represents the same information as 'education' column.
marital-status: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.
occupation: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.
relationship: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.
race: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
sex: Female, Male.
capital-gain: continuous.
capital-loss: continuous.
hours-per-week: continuous.
native-country: United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands.

Exploratory Data Analysis (EDA):

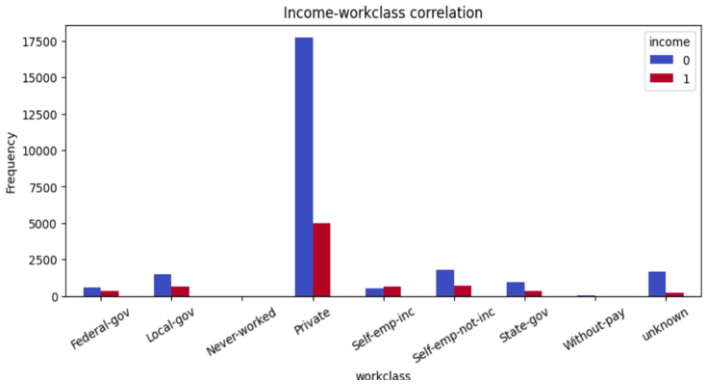


Income feature:

- The Target feature (Income) is having two classes
- <=50k (0) and >50k (1).

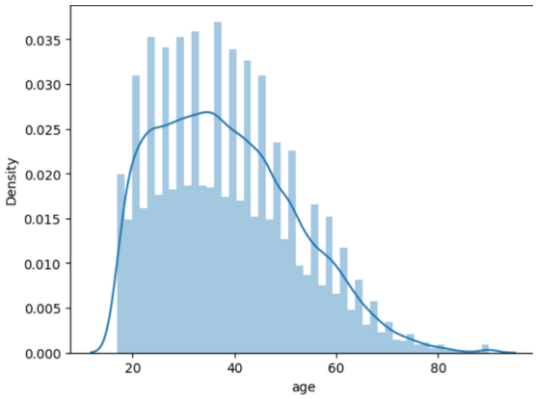
- Most of the people have income <= 50k.

Income-workclass correlation:



- One of the workclasses(with 1836 samples) is not named! It's represented by '?'.
- So '?' considered as 'unknown'.
- Private jobs have topped in both income categories.

Age:

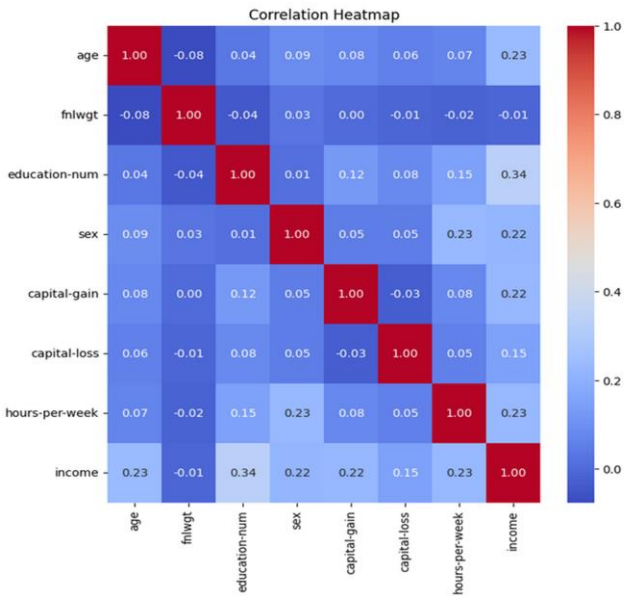


- Minimum age is 17 and maximum is 90.
- Average working age in this dataset is 38.585
- Age feature is left skewed.

- In 'age' column's boxplot, age>80 (approx.) entries are outliers.
- It's natural that people prefer to leave work life at such an age.

Insights and Observations:

- Education-num is highly related to income classification.
- The top occupation (in terms of count) is 'Prof-specialty' and the least one is 'Armed-Forces'.
- In <=50K income category, highest number of people are in 'Adm-clerical' occupation followed by 'Craft-repair' and 'other service'.
- In >50K income category, 'Exec-managerial' topped in terms of count, followed by 'Prof-specialty' occupation.



- The mostly employed people are those who completed High School graduation.
- White people are getting more opportunities as compared to others (in both income categories).
- **In the race feature,** clearly, white people are getting more opportunities as compared to others (in both income categories).
- Fnlwgt, Capital loss features are not effecting income feature.