**Introduction**

Understanding and predicting an individual's income level is essential for policymakers, researchers, and various industries. It is valuable in multiple domains that are related to business, such as advertising and hiring (Matz et el., 2019). In marketing, details about an individual's financial circumstances can be utilized to aim at consumers with appropriate products or services. In recruiting, knowledge of an individual's present income can serve as a reference point to assist companies in negotiating. Apart from that, accurate income level predictions can inform targeted interventions, enabling more effective resource allocation and support for those who need it most, to help resolve income inequality issue. The emergence of machine learning as a powerful predictive tool has opened up new possibilities for addressing data-driven problems like income prediction. This research aims to investigate the potential of machine learning in predicting an individual's income level based on demographic and employment information.

**Literature review**

Traditionally, classical statistical methods, such as linear regression and logistic regression, can be used in income prediction. For example, Yarnold(2019) used linear regression and novometric models respectively to model the relationship between education income, with both models displaying strong effects. Recent studies have demonstrated various more advanced machine learning methods and information that can be used in income prediction. Even modelling based on the digital footprints people leave on Facebook is a viable means of income prediction(Matz et el., 2019). In this study, they found that income is predictable using Facebook likes and Status Updates with machine learning techniques. A similar research buildt a model for income prediction by using users' temporal orientation of Tweets (Hasanuzzaman et al., 2017). These studies all show the possibility of using machine learning models to predict an individual's income with their information.

**Research question**

The research question of this paper is "Is it possible to use machine learning to predict an individual's income level based on demographic and employment information?".

**Presentation of data**

Firstly, we import packages needed for data visualization and analysis.

In this research, an income dataset from Kohavi and Ron (1996) is used.This dataset includes 15 variables, including age, work class, fnlwgt, education, edu\_num, marital status, occupation, relationship, race, sex, capital gain, capital loss, working hours per week, native country, and income. Income is selected as the dependent variable to analyze the impact of other variables on it. The following is the specific explanation of each variable:

Income: anual income, classified as >50K and <=50K.

age: age of each individual, continuous variable.

workclass: classifications of each individual's work, including: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.

fnlwgt: it is a weight assigned by the Census Bureau, representing the number of people the census believes share the same attributes, continuous variable.

education: classifications of education of each individual, including: 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: the number of years of education, continuous variable.

marital-status: classified as: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.

occupation: classified as: 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: classified as: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.

race: classified as: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.

sex: Female, Male.

capital-gain: continuous variable.

capital-loss: continuous variable.

hours-per-week: working hours per week, continuous variable.

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.

Read the dataset remotely from the [github repository](https://github.com/zhuoranliu22/CASA0006" \t "_blank), this repository stores both the dataset and the code:

The dataset does not contain headers, so we manually add these:

Using .info() function to check data types and non-null counts:

Using .isnull().sum() function to check for empty columns, the result shows there are none:

Checking if there are any unusual values in the dataset:

There are values as '?' in the dataset, so we replace them with na values and use .dropna() to drop them:

The data treats Taiwan as an independent country by mistake, so we replace them:

Using .duplicated().value\_counts() to check if there are any duplicated values:

Deleting the duplicated values:

Using label encoder to transfer object variables into numeric values for later analysis:

The followings are the visualizations of the object variables, so that we know roughly the proportion of each class within the variable data:

The plot above is especially caotic, which gives very little information, so we directly print the values to get a clearer understanding:

After dealing with the object variables, we standardize the continuous variables to transfer them into the same scale to make it more comparable:

Using the .describe() function of look at the summary statistics of each variable:

# Methodology

Since there are multiple independent variables and one dependent variable, the regression method is firstly considered to observe their relationship. The dependent variable Income is binary, classified as <50k and >=50k, so logistic regression is more suitble within this context. Logistic regression is a type of generalized linear model that uses the logistic function to model the relationship between a binary dependent variable and one or more independent variables. Therefore, it is the ideal statistical method for this research. However, compared to other more advanced machine learning methods, it has some limitations:

1. Logistic regression presumes a linear relationship between the log odds of the outcome variable and the predictor variables. However, this assumption might not be valid for this research, in which case non-linear relationships among variables is of greater chance.
2. Logistic regression may have difficulty in capturing the complexity of interactions among variables, at least not as well as more advanced models like decision trees, random forests, and deep learning models.
3. Logistic regression is sensitive to outliers, potentially resulting in overfitting and inacurate predictions.

Therefore, decision tree and random forest will be used as well to more accurately build the model and to observe which of these factors have greater impact on income level. In building decision tree classifier, the hyperparameters including maximum depth and minimum sample split will be tuned by using GridSearchCV to find the optimal combination. After building the models, the accuracy of the models will be listed for comparison.

## Logistic Regression

Setting a variable X to be the independent variable set by excluding income from the data:

Building a logistic regression and then fit it:

Print the coefficients of each independent variable and accuracy of the model, then draw the confusion matrix:

Creating a dictionary array to store the dataset:

Exlude the variable names and create X\_matrix, then convert it into an array:

Divide the dataset into training data and testing data based on the default ratio, which is 75:25 division:

Firstly, create a default decision tree classfier:

The score on the training dataset is:

The score on the testing dataset is:

The tree depth with default settings is:

Tune the hyperparameters:

Build models with the 8 values of the max\_depth and store the test scores in variable test:

Visualize the test scores of these models:

Train the final model with the tuned parameters:

For demonstration, the structure of this tree is output by setting the maximum depth to be 3:

The plot shows the feature importance:

Building the random forest model by using the default hyperparameters, as tuning them is too time-consuming in this case:

The score on the testing data is lower than that on the training data, suggesting overfitting problem.

Here is the feature importance of this model:

# Results and discussion

In logistic regression, the confusion matrix shows that 21295 smaples are correctly predicted as type 0, which is <50k, and 3397 samples are correctly predicted as type 1, which is >=50k. From this result, it suggests the model acuracy is higher when the actual outcome is 0, but much lower when the actual outcome is 1, in which case the number of mistakenly predicted samples is even greater than the correct ones. Therefore, it is essential to build more advanced machine learning models to get a more precise classification.

The accuracy of the logistic regression model is 0.8192, and the score on the training data and testing data of decesion tree model and random forest model is listed in the following table. By comparison, the score of random forest model is higher than the other models, but the difference in R2 between training data and testing data of decision tree model is much lower most of the times. It suggests that decision tree model has a better chance in mitigating overfitting problem.

Regarding the feature importance, the top three important variables in logisitic regression model are capital\_gain, edu\_num and sex. In decision tree model, the top three are relationship, fnlweight and age. In random forest model, these are fnlweight, age and capital\_gain. These results show they have a greater impact on income, especially age.

The overall result is that the decision tree model performs better than the other two models, with relationship, fnlweight and age to be the most influential factors on income.

There are several limitations of this research:

1. Outliers were not handled. Logistic regression is sensitive to outliers, so removing them may improve the performance;
2. Even though the decision tree model has the smallest difference in R2, it still has overfitting problem;
3. The hyperparameters of the random forest model were not tuned as in the decision tree model. Tuning the hyperparameters may improve the performance score of this model.

# Conclusion

After conducting the whole research process, the research question can be answered: it is possible to use machine learning to predict one's income level using their demographic and employment information. Three machine learning models are used: logistic regression, decision tree classifier, and random forest classifier, with score up to 0.85. Among the attributes, age, capital gain, fnlweight, relationship, number of years of education, and sex have greater impact on one's income. Therefore, government or companies who need to have a general understanding of people's income level can pay more attention to these attributes, and use machine learning models to make predictions after gaining ethical approval.