**Predicting Income levels using Classification Algorithms**

**Introduction**

This project is an analysis performed to create a prediction model for the prediction of income levels of citizens based on the 1994 census data. The idea is to classify a person in to two types of income categories - >50k and <=50k based various demographics such as age, work class, education, occupation. The dataset chosen is the “Adult Census Income” from Kaggle. Various data pre processing methods are applied and relationships between variables are visualized using plots. Machine learning models of decision trees and K – nearest neighbours are created and are compared on their performance results. Between the two models, the KNN model performed slightly better with an overall accuracy of 85%. Our model assists in giving informed decisions and better services in the financial, banking domains which is mainly required for financial institutions to know an individual’s credit score, or whether an individual requires finance assistance or not.

**Related Work**

Rehman et al. [1] used machine learning techniques and data mining to predict the annual income of individuals. The Decision Tree, Naïve Bayes and logistic regression were used and resulted in 82%, 80% and 76% accuracies respectively.

Bekena [2] used Random Forest Classifier to predict the income levels of people based on the data related to their education, gender, occupation and others. The developed model showed an accuracy of 85% and was very weak in predicting high income individuals. But, the model was succeed in identifying the key factors that explain the difference between high and low income.

Navoneel Chakrabarty and Sanket Biswas [3] used a gradient boosting classifier model and worked on a similar model of dataset to predict yearly income level categories of people based on set of attributes. They obtained a high accuracy of 88.16% using their approach.

**Methods**

**Data collection**

We have seen multiple datasets available on Kaggle, a public platform for various data sources. Based on the features and number of records available we have chosen the Adult Census Income dataset. This dataset is provided by UCI Machine learning repository. The data was extracted from the 1994 census bureau database. The dataset comprises of multiple attributes such as gender, education level, occupation, marital status, native country related to a person and an income variable which has two classes, to describe the person’s income as over $50k or under $50k in a year.

**Data Mining Pipeline**

We have followed a well-planned data mining pipeline to implement our project through different stages.

Data Pre-processing-

Data pre-processing is performed to explore the data, identify missing values, clean the data and modify/add any existing variables in the dataset. The dataset is loaded into the colab notebook and is explored using pandas library. There are a total of 15 variables, for a total of 32561 records. There are 9 categorical variables in the data. The data doesn’t consist of any missing values. The categorical variables have multiple classes, and all the categorical variables are encoded using the one-hot encoding method. Encoding of these variables allows them to be used in the model training. The classes are assigned binary numbers of 1 and 0 based on class value. Few data values are replaced within the occupation column for better understanding. Also, the dataset doesn’t contain large outliers in the numerical values.

Data Visualization

Data visualization is performed on multiple combinations of variables to understand the data much precisely. The categorical and quantitative variables are plotted using different styles available within the seaborn and matplotlib libraries. A histogram is created for the age variable giving its distribution. Correlation matrix is generated to understand the dependencies between different variables. Correlation between the income variable is important from this plot and variables with less or negative correlation values can be discarded. A count plot is created for the distribution of occupation types. The income distribution between the two levels is by gender is observed through a side-by-side box plot. The relation between income category and working hours per week is represented through a violin plot.

Data splitting

The data post pre – processing is ready to be used in our machine learning analysis. For understanding the performance of the model, we require a testing set of data. Therefore, the current dataset is split into two parts of training set and test set. The training set is 80% of the data and is used to train machine learning models. Test set is 20% of the data and will be used for model evaluation purpose. The data is divided initially into X and y variables as part of attribute and classes. The income class is the class variable, and all other variables are predicting variables. The splitting is done using sklearn library to obtain four variables related to training and testing.

Model Training

This stage involves the training of machine learning models through python library of sklearn. The machine learning task in this project is a classification task and we aim to classify the group of annual income into >$50k or <=$50k.

Firstly, we have trained the decision tree model. Decision tree model is a machine learning model which implements the classification task in terms of a tree structure by using different attributes at each level. The model is fit on the training data. It is then evaluated for accuracy by making prediction on the test data. The classification report is created on comparison for the generated prediction values and the test values of income class.

Similarly, we have then implemented a K nearest neighbours model. KNN model is a model which makes use of the proximity or distance between two values to classify new values. The model is trained on the data and predictions are made on testing data. These predictions are also used to create classification report of this model.

Software used

We have used python as the primary programming language in our project for all the data mining stages. The machine learning libraries of python like sklearn, pandas are utilised. Other sub modules of sklearn like neighbors, model\_selection, train\_test\_split, decisiontreeclassifier, classification\_report is used. For visualization purpose the libraries of seaborn and matplotlib have been used. The entire project has been carried through colab workspace.

**Results and Discussion**

**Visualization Results**

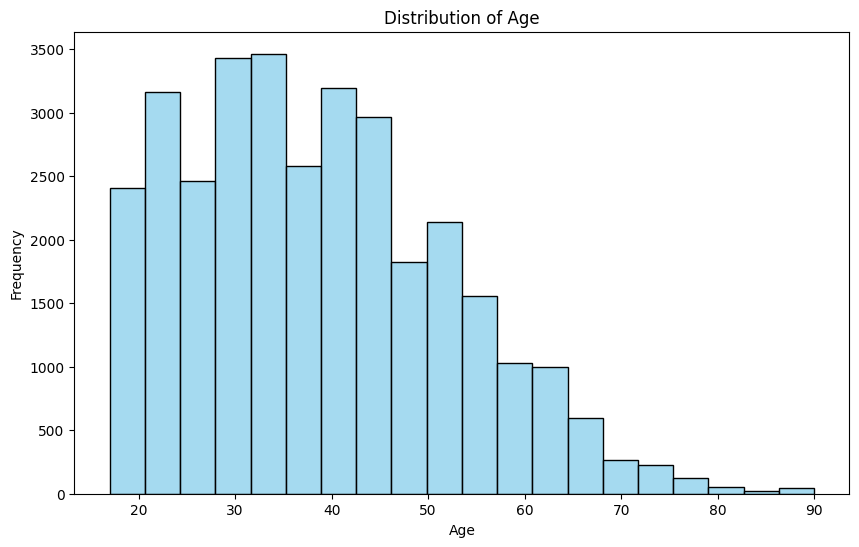
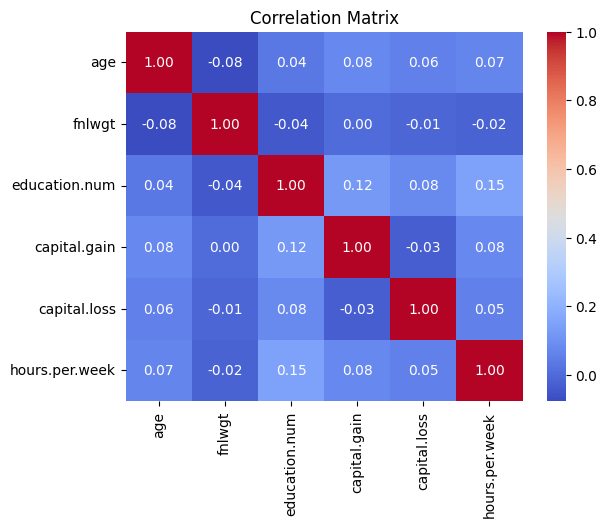
****

Fig.1 Histogram of age

The majority of the age range of people lie between 30-50.

****

**Fig.2** Correlation matrix of variables

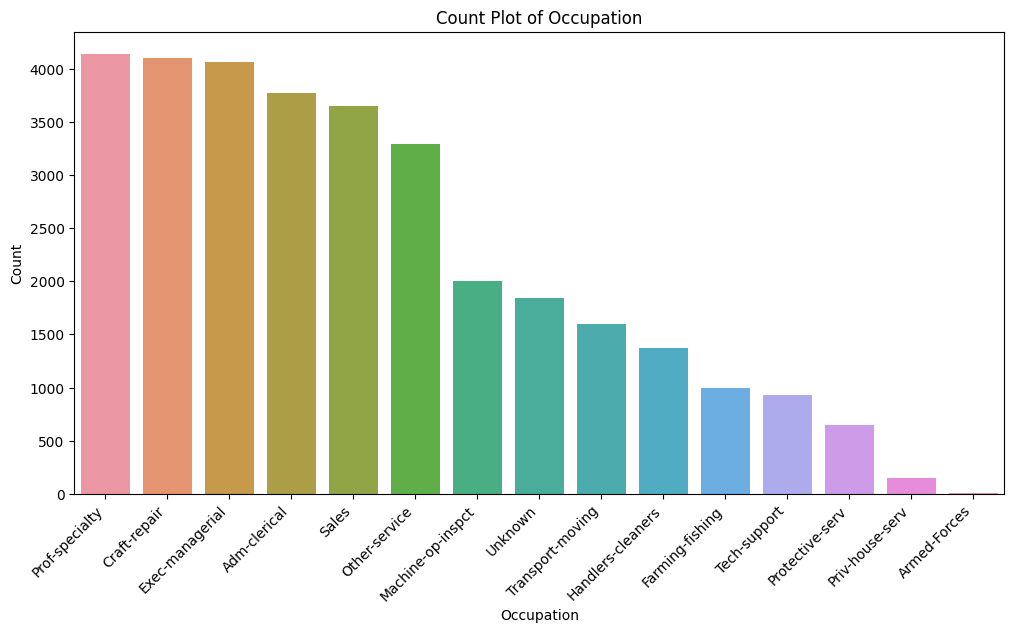


Fig.3 Distribution of occupation types

Prof-speciality and craft-repair have the highest frequency.

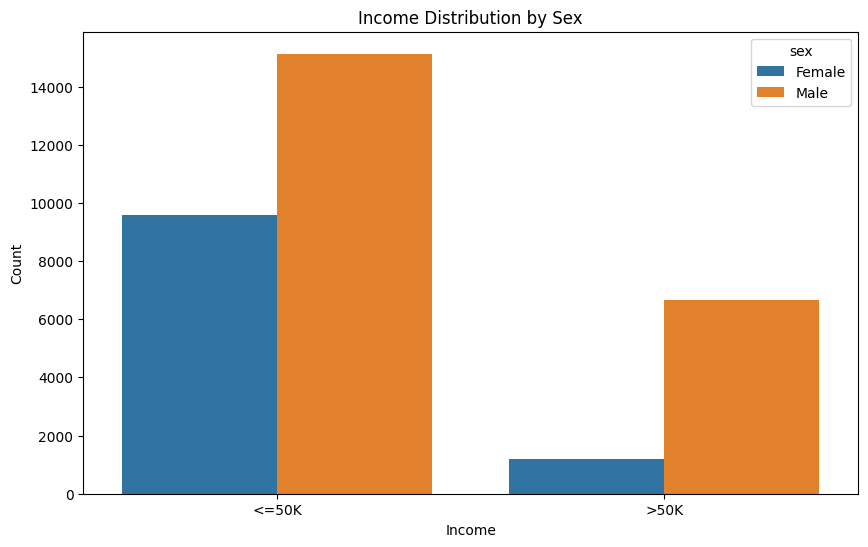


Fig.4 Income distribution

Males have higher number of occurrences in both the categories.

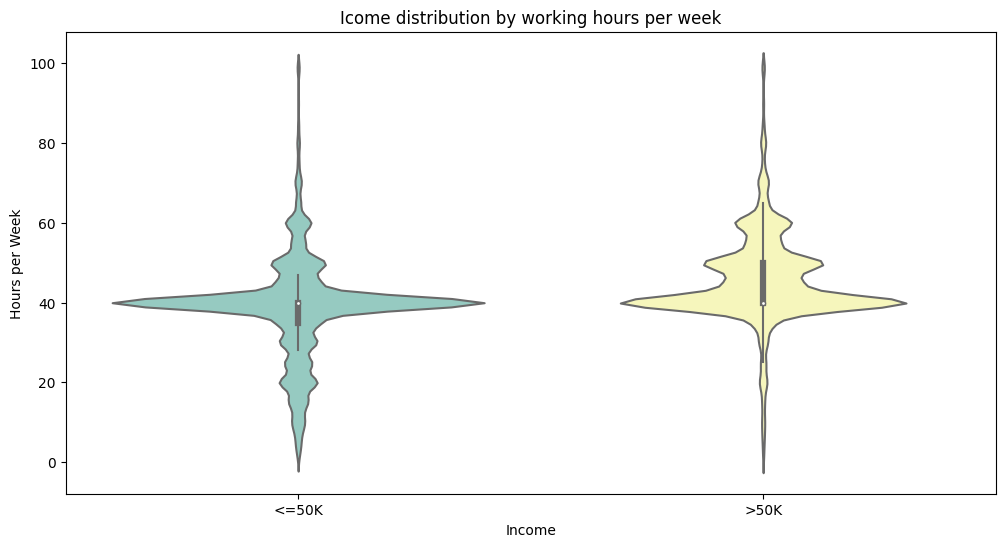


Fig.5 Violin plot of income and working hours per week

Income category of >50k have higher working hours.

**Model evaluation**

The decision tree model has given an accuracy of 82.6%.

The classification report is as follows-

precision recall f1-score support

<=50K 0.88 0.89 0.89 4976

>50K 0.64 0.61 0.62 1537

accuracy 0.83 6513

macro avg 0.76 0.75 0.75 6513

weighted avg 0.82 0.83 0.82 6513

The confusion matrix for the model prediction-

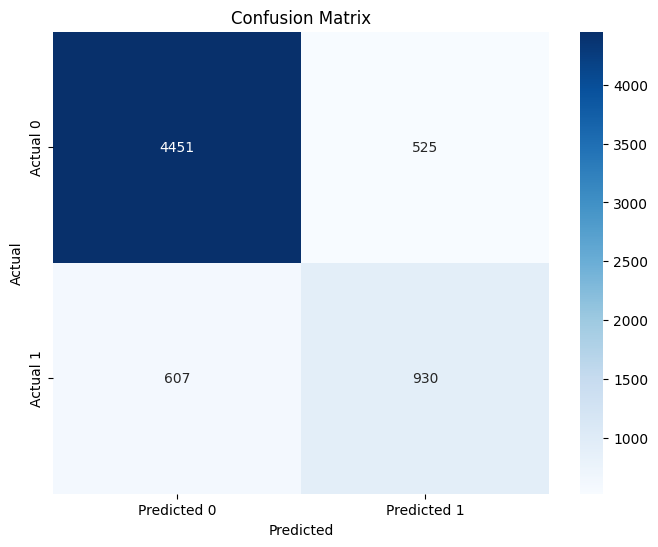


Fig. 6 Decision Tree prediction confusion matrix

Overall, it can be said that decision tree model is an efficient and accurate classifier in this task. It classifies many of the classes with income <50k correctly but fails to classify high number of income classes >50k.

The evaluation of the KNN model:

The accuracy obtained from this model is 85.6%.

Classification report-

precision recall f1-score support

<=50K 0.88 0.94 0.91 4976

>50K 0.76 0.58 0.66 1537

accuracy 0.86 6513

macro avg 0.82 0.76 0.78 6513

weighted avg 0.85 0.86 0.85 6513

Confusion matrix-

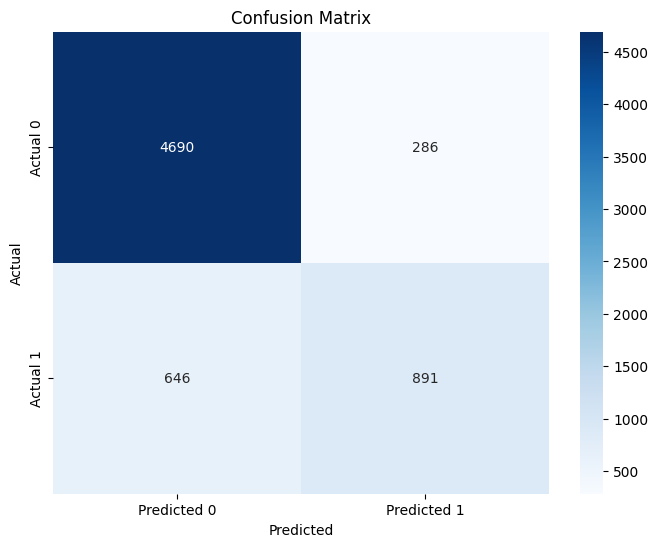


Fig. 7 Confusion matrix of KNN predictions

The KNN model has similar performance statistics to the decision tree model. It slightly performs better with an accuracy of 85%. The KNN model shows a similar trend of classifying the income category of <50k correctly but inaccurate predictions for the other category.

**Conclusion**

With this project, we have developed a machine learning model for classifying an individual’s annual income in to two categories of income less than $50k or greater than $50k. This classification task is useful for study and research by government, financial institutions, banks, and other organizations. We have performed various data pre-processing steps on the census data taken from Kaggle to obtain useful information. Data visualization has been performed to analyse our data well. Decision tree and KNN models are trained on the data to perform the classification task. After proper evaluation steps, we have found that both models perform in a similar fashion and the KNN model has a higher accuracy of 85%.

The limitations of this project include the available attributes. There might be other factors of an individual also which are related to an individual’s income such as location, role of work, experience etc. Also, our model could not show very high accuracy in classification for a particular category of income. This might be due to the limited data available for the group.

In the future, we aim to extend our work in to applying other machine learning classifiers such as gradient boost, random forests. We also aim to explore further datasets with more records and attributes of data through multiple sources.

**Data and Software Availability**

The dataset we have used can be found at - <https://www.kaggle.com/datasets/uciml/adult-census-income>.

GitHub link of the project - <https://github.com/GVSU-CIS635/gvsu-cis635-term-project-income-prediction>.

**References**

[1] Rehman, A. U., Saleem, R. M., Shafi, Z., Imran, M., Pradhan, M., & Alzoubi, H. M. (2022, February). Analysis of Income on the Basis of Occupation using Data Mining. In 2022 International Conference on Business Analytics for Technology and Security (ICBATS) (pp. 1-4). IEEE.

[2] Bekena, S. M. (2017). Using decision tree classifier to predict income levels.

[3] N. Chakrabarty and S. Biswas, "A Statistical Approach to Adult Census Income Level Prediction," 2018 International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), Greater Noida, India, 2018, pp. 207-212, doi: 10.1109/ICACCCN.2018.8748528.