**Progress Update Report**

**Team members:**

**Susmitha Inturi**

**Venkata jayanth pasam**

**Progress so far**

Data Gathering

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 pre-processing

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. 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.

The data is also visualized using multiple types of plots using the libraries matplotlib, and seaborn to understand the distribution. Plots for income distribution by sex, and working hours per week are created. The data distribution of occupation and age are created through bar plot and histogram. A correlation matrix is generated to understand level of relationships between the variables. The finalwgt attribute is excluded from the data as it has less correlation values.

Data Splitting and Model Building

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 data is divided into training and testing sets using a test ratio size of 20%. The splitting is done using sklearn library to obtain four variables related to training and testing.

The decision tree model is selected for the classification task. The model is loaded from the sklearn library. It is fitted using the training data of X\_train and y\_train.

**Data:**

The dataset link that we are using in the project is-

<https://www.kaggle.com/datasets/uciml/adult-census-income>

**Challenges**

Encountered difficulties and how we solved them:

We had to understand the importance of variables with no description such as fnlwgt, education.num and other numerical variables. The correlation matrix helped us to analyse the relevance of such variables. The values made us to decide which variables to include and exclude in further processing.

On training the model initially, we encountered a value error indicating that the model could not work on string type of data present within the occupation class as it could not be converted to float type. This made us to go back to our pre-processing step and ensure that all the variables of type object were encoded into numbers. We could then proceed with the model training.

**Collaboration**

As planned, we ensured that we met twice in every week to maintain the progress of the project. We have met in person and discussed the progress made so far and planned our next steps. Both of us are contributing very efficiently to the project by sharing the work and discussing every step. We are exploring different sources and working on our project to prepare an effective model.

**Next Steps**

According to our proposed pipeline, and our progress so far, the stages left in our project are model testing and implementation. We aim to test the trained decision tree model on the testing data and observe different performance metrics to analyse the performance of the model.

The accuracy of the model might be a potential challenge for us in these steps. We might have to improve our training process or pre-processing steps in case of lower accuracy. We may also need to check with other type of model for performance improvement.