**4/4/2023**

***Professor: Claudia Gomez***

*Submitted By: Group 08*

*Tripathi Vidhi (041057189) (*[*trip0046@algonquinlive.com*](mailto:trip0046@algonquinlive.com)*)*

*Walia Kiran (041057641) (*[*wali0034@algonquinlive.com*](mailto:wali0034@algonquinlive.com)*)*

*Shah Saiyam (041080364)* *(*[*shah0371@algonquinlive.com*](mailto:shah0371@algonquinlive.com)*)*

*Panchal Paras (041057680) (*[*panc0062@algonquinlive.com*](mailto:panc0062@algonquinlive.com))

**Predicting Classification of Income Level**

Data Science Foundations (23W\_CST2105\_300)

# BUSINESS UNDERSTANDING

### Background:

The **adult income** dataset is taken from UCI machine learning from which we can explore the possibility of predicting income level based on the individual’s personal information and classify them into various categories. **The output of this exercise can be useful to keep government records to develop public policies related to taxation, welfare, and social security.**

### Problem statement:

Develop a classification model on labeled dataset and predicting income level on unseen data with higher accuracy using demographic and socioeconomic variables.

### Objective:

* To identify which variables are most important for predicting income level, and to evaluate the effect of other variables in the accuracy of the model.
* To assess the overall performance of the model using various evaluation metrics such as accuracy, precision, recall and F1-score.

# DATA UNDERSTANDING

### Data Capture and explanation

Our team gathered this data from Kaggle, containing 14 attributes including age, workclass, fnlwgt, education, educational-num, marital status, occupation, relationship, race, gender, and target variable is income, which is divided in to two class: ‘<=50K’ and ‘>50K’.

### Data Exploration

* Dataset contains 9-character variables and 6 numerical variables. Three columns: workclass, occupation and native\_country contain mtng values which we imputed with mean values.

# DATA PREPARATION & FEATURE ENGINEERING

In projects that involve analysis of data and machine learning, and data preparation feature engineering are critical steps. Following steps were taken to make machine readable clean dataset:

* Libraries that will be used throughout the model: Pandas, NumPy, Matplotlib, Scikit-learn.

### Data Preparation

1. Data cleaning: Identifying and dealing with missing values, duplicate values using SimpleImputer class from sklearn library, and dealing with the outliers which were identified by plotting various boxplots of individual variable and with respect to label i.e., income variable.
2. Data Transformation: Converted all the classes of categorical variable to a machine understandable format done using the Sklearn preprocessing module LabelEncoder(). MinMaxScaler from Sklearn was used to shrink the data to a range, usually 0 to 1.
3. Feature Engineering & selection: Pearson correlation matrix was used to find bivariate relationship between variables and their impact on the target label. Correlation with the target variable was plotted for all the variables. Used SelectKBest from sklearn library to find the k highest score of input variables and select top 7 variables.
4. Data Splitting: Sklearn library test\_train\_split function was used to split the dataset into training set and testing set into the ratio of 80-20.

# MODELING AND PIPELINES CREATION

### Classification models:

To evaluate the model, we imported precision score, recall score, accuracy score, f1 score from the sklearn library for all the models.

1. Support Vector Machines:

k-fold cross\_val\_score was from sklearn library used to check the overfitting of the data with CV= 5 and found cross-validation scores.

1. Decision Tree Model:

DecisionTreeClassifier from sklearn library, highly intuitive and easy to classify as the tree grows. For hyperparameter tuning, used gridsearchCV from sklearn library.

1. Random Forest Classifier (Ensemble model):

RandomForestClassifier and KFold, cross\_val\_score imported from sklearn library and tested with input parameters: n\_estimators: 100, max\_depth: 7, min\_samples\_split: 2, min\_samples\_leaf: 2, max\_features: auto, random\_state: 42.

1. KNeighborsClassifier (KNN)

KNeighborsClassifier from sklearn library, a non- parametric algorithm i.e., it stores the data from the training set while training and when new data is fed, it classifies based on the similarity.

1. Logistic Regression

LogisticRegression from sklearn.linear\_model is used and fits the regressor over the training data. The outcomes will be in binary form (0 or 1) as per predefined labels. Evaluation is done by importing the confusion matrix and accuracy score from sklearn.

**Pipelines:**

Diagram

Description automatically generated with medium confidence

# MODEL EVALUATION AND SELECTION

### Evaluation metrics:

The effectiveness of classification models can be evaluated using a variety of evaluation metrics:

1. Accuracy
2. Precision
3. Recall
4. F1 Score
5. ROC Curve: A graphical representation of true positive rate (TPR) against the false positive rate (FPR) at different thresholds of classification.
6. AUC Score: A metrics used to evaluate binary classification problem with a value ranging from 0 to 1. The higher the values indicating the better performance of the model.

### Results interpretation & Model selection

* The evaluation measure that we used for model selection for income classification is accuracy. It is proportion of accurate predictions made by a model to all the predictions made and higher the value better is the model.
* ROC curve having TPR and FPR was plotted to compare the models and to select the best performing model.

Chart

Description automatically generated

# CONCLUSION

### Results analysis

Random Forest is having accuracy of 84.84% which is the highest among all the models. As well as from ROC curve, AUC = 0.73 which is second highest.

### References:

<https://www.kaggle.com/datasets/lodetomasi1995/income-classification>

[http://cseweb.ucsd.edu/classes/sp15/cse190-c/reports/sp15/048.pdf](file:///C:\Users\Abhi\Desktop\ML\Project\ http:\cseweb.ucsd.edu\classes\sp15\cse190-c\reports\sp15\048.pdf)