A logo for university of texas

Description automatically generated

# FINAL REPORT FOR

# DATA MINING AND INFORMATION VISUALIZATION

Table of Contents

1. Executive Summary ------------------------------------------------------------------ 3
2. Problem Statement ------------------------------------------------------------------- 4
3. Data Description ---------------------------------------------------------------------- 5
4. Explanatory Data Analysis ---------------------------------------------------------- 7
5. Data Pre-Processing ----------------------------------------------------------------- 18
6. Model Selection ----------------------------------------------------------------------- 19
7. Evaluation Metrics -------------------------------------------------------------------- 21
8. Conclusion ------------------------------------------------------------------------------ 27
9. References ----------------------------------------------------------------------------- 28

# Executive Summary

The U.S. Adult Income dataset is a resource used to examine the correlation between various demographic factors and income levels in adult populations. This analysis aims to identify significant predictors of income level to better understand economic dynamics across different demographic groups.

Our project here, addresses this problem, with our huge data available for the Kaggle. Based on this problem, we built -

1. A reliable machine learning model, that can predict the income of an individual is above 50k or not with 83.7% accuracy.

2. A Visualization, that shows the parameters under which our machine learning model has been built.

When we got this dataset, initially we tried to find the null values as they would affect our model building. Next, we used imputation method based on the data distribution. After that we have performed data cleaning where we found a column to be a replica of one of the existing columns. After removing that column from the dataset, we tried to understand the key features that are influencing the model by performing explanatory data analysis. It helped us in understanding the main features that must be taken into consideration before building our machine learning model. Then we proceeded with model building.

In model building, we considered two methods for developing decision tree model.

1. Gini Index

2. Gain Ratio

After building these models we tried to find the accuracy and efficiency of the model by performing a confusion matrix for both the models. From which we were able to decide which method is best to develop decision tree model.

This model can be very useful because there are several features that are responsible for an individual’s income prediction. By using this model, we can understand what are the parameters that are influencing the person’s income. By using these parameters, the US government can perform necessary action so that they can find a way to increase the ratio of individuals who are earning less than 50k. This will help both economically and demographically for the US.

# PROBLEM STATEMENT

To build a reliable classification model that predicts the income class of individuals, a comprehensive approach must be taken, involving several key stages: data collection, preprocessing, exploratory data analysis, feature engineering, model building, validation, and deployment. This systematic approach will allow for the development of a model that can accurately classify individuals into income groups such as "<=50K" or ">50K." Such a model can significantly aid government agencies in making data-driven decisions on resource allocation, policy planning, risk assessment, and optimizing demographic dividends.

A group of people standing on top of coins

Description automatically generated

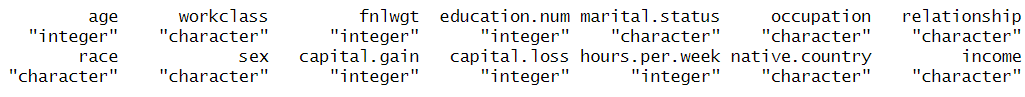
# DATA DESCRIPTION

Data description is the foundational step where data on various attributes of individuals is described. This may involve using existing survey data, such as the U.S. Adult Income dataset, or collecting new data to ensure diverse demographic coverage and up-to-date information. The dataset includes:

A screenshot of a document

Description automatically generated

The columns of our dataset are:



The d-types and values for each column in of our dataset are:

A computer screen shot of a computer code

Description automatically generated

The summary of our dataset is:

A screenshot of a computer

Description automatically generated

A group of people standing around a computer

Description automatically generated

# EXPLANATORY DATA ANALYSIS

EDA is crucial for understanding the underlying distributions of the data and identifying patterns or anomalies. The steps involving are:

1. Visualization for categorical columns:

A screenshot of a computer

Description automatically generated

A close-up of a computer screen

Description automatically generated

1. Visualization for numerical columns:

A graph with a red line

Description automatically generated

A graph with a line and a line

Description automatically generated

A graph with a line

Description automatically generated

A graph with a line graph

Description automatically generated

A graph with a line graph

Description automatically generated

A graph with a red line

Description automatically generated

1. Visualization for income vs other numerical features:

A graph of a graph with red rectangles and black lines

Description automatically generated

In box plot analysis for income vs age, we can see that for both income greater than 50k and below 50k, people whose age is more are earning more compared to people whose age is less.

A graph of a graph with black and red lines

Description automatically generated with medium confidence

In box plot analysis for income vs fnlwgt, we can see that for both income greater than 50k and below 50k, both are equally distributed i.e fnlwgt doesn’t affect the earning of a person.

A graph with red and black lines

Description automatically generated

In box plot analysis for income vs education, we can see that for both income greater than 50k and below 50k, as the education background is high the person is earning higher. Like if a person has not studied much then his earning is less compared to higher educational background people.

A graph of a graph

Description automatically generated with medium confidence

In box plot analysis for income vs fnlwgt, age we can see that for both income greater than 50k and below 50k, both are equally distributed i.e fnlwgt doesn’t affect the earning of a person.

A graph of a loss

Description automatically generated with medium confidence

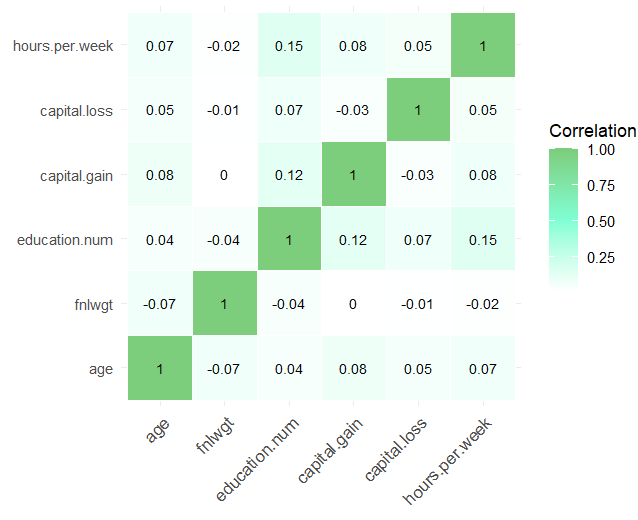
In box plot analysis for income vs capital, we can see that for both income greater than 50k and below 50k, both are equally distributed i.e capital doesn’t affect the earning of a person.

A screenshot of a graph

Description automatically generated

In box plot analysis for income vs hours per week, we can see that for both income greater than 50k and below 50k, both are equally distributed i.e hours per week doesn’t affect the earning of a person.

1. Correlation analysis:



From the correlation plot we can see that there are not many parameters that are affecting the variables. Like there are no two variables that are much correlated. We can see that hours per week and education are the only two variables that are somewhat correlated with a positive correlation of 0.15. Whereas for the rest there is no such correlation.

1. Violon Plot Analysis:

A graph showing a couple of colored shapes

Description automatically generated with medium confidence

From violon plot we can say that as the age increases the persons who are earning more is increasing compared to those who are young. As we can see there is more density for people around age 30 to 60 for earning greater than 50k. In case of people earning less than 50k the density is more between 0 to 30. So finally, as the age increases the income of the person increases.

1. Stacked Plot Analysis:

A graph with red and blue squares

Description automatically generated

From this stacked plot analysis, we can see that for married civ spouse. The number of people who are earning is high compared to other martial status. The number of people who are earning more than 50k is more in case of never married. As we know that people who tend to earn more try to earn before marriage, so we are seeing more number of people in that category. In cases like martial AF spouse, married spouse absent, separated and widowed there are less number of people who are earning.

# DATA PRE-PROCESSING

This stage involves cleaning the data to handle issues such as missing values, errors, or inconsistencies in the dataset. Data preprocessing steps include:

1. Handling missing values: We used mode imputation for missing values.
2. Feature Engineering: There is a column called education which is a replica of existing column from our dataset education.num. So, we are removing that column.

The original dataset head values:

A white background with black text

Description automatically generated

A screenshot of a computer

Description automatically generated

The dataset head values after pre-processing:

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

# MODEL SELECTION

For the given dataset, as we can see from the correlation plot from EDA there is not much correlation among the variables. Atmost the highest correlation is 0.15. Which is not even a moderate correlation. As for the rest they are almost 0. The second thing is that there is no linearity. As for the last factor, as we can see from the dataset there are so many categorical columns. On top of that the problem statement of this dataset is of type classification problem.

By considering all the above factors we have decided to proceed with decision tree classifier. Even in decision tree classifier building we have considered two factors. They are:

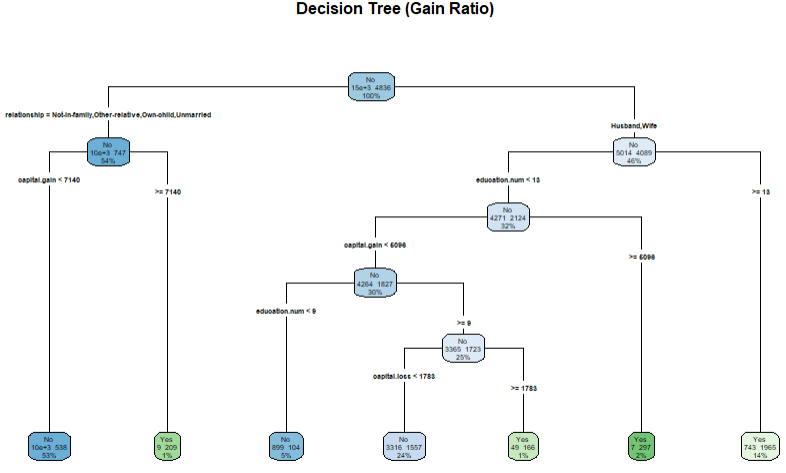
1. Gini Index
2. Gain Ratio

We have built the models using the above two factors. The decision tree using Gini index as factor is:

A diagram of a decision tree

Description automatically generated

The decision tree using Gain Ratio as factor is:



Visualization of decision tree for both Gini index and Gain ratio is:

A diagram of a graph

Description automatically generated

# EVALUATION METRICS

1. Decision Tree for Gini Index:

Important Features for this model:

A graph with different colored squares

Description automatically generated

The above graph shows the top 10 important features that are determining the decision tree model. The top influential feature is capital.gain with overall score of 100. Next its education.num followed by martial.statusMarried-civ-spouse with overall score of more than 75. Age and martial.statusNever-married are affecting moderately with an overall score near 50. At last occupationExec-managerial and capital.loss are showing minimal affect in the overall score. Other than these there are no such features which are affecting the model.

Confusion matrix for decision tree based on Gini index:

A graph with purple squares and numbers

Description automatically generated

The summary of the model is:

A screenshot of a computer

Description automatically generated

The accuracy of this model is 0.8376. Which indicates that it is performing well. The AUC score for this model is 0.8245405.

The ROC curve for this model is:

A graph of a positive rate

Description automatically generated

From the ROC curve we can notice that the curve is affecting more positively which indicates that our model is performing well. Here it indicates that true positive values are highly dominant compared to true negative values.

1. Decision Tree for Gain Ratio:

A graph with green and red squares

Description automatically generated

The above graph shows the top 10 important features that are determining the decision tree model. The top influential feature is capital.gain with overall score of 100. Here for this model martial.statusMarried-civ-spouse is also affecting with an overall score of almost 100. Next its education.num overall score of more than 75. Age and martial.statusNever-married are affecting moderately with an overall score more than 50. At last occupationExec-managerial and capital.loss are showing minimal affect in the overall score. Other than these there are no such features which are affecting the model.

Confusion matrix for decision tree based on Gini index:

A graph with purple squares and numbers

Description automatically generated

The summary of the model is:

A screenshot of a computer

Description automatically generated

The accuracy of this model is 0.838. Which indicates that it is performing well. The AUC score for this model is 0.8266527.

The ROC curve for this model is:

A graph of a positive rate

Description automatically generated

From the ROC curve we can notice that the curve is affecting more positively which indicates that our model is performing well. Here it indicates that true positive values are highly dominant compared to true negative values.

# CONCLUSION

* As we can compare

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Accuracy | Specificity | AUC Score | Kappa | Sensitivity |
| Gini Index | 0.8376 | 0.9488 | 0.8245405 | 0.497 | 0.4888 |
| Gain Ratio | 0.838 | 0.9483 | 0.8266527 | 0.4992 | 0.4921 |

* By comparing both decision trees, the model which was built using gain ratio is more efficient compared to model built using gini index.
* From the model we can conclude that capital gain, martial status: married–civ-spouse and education are the top features that are affecting the income of a person.
* So, the associate committee can look into it and try to implement some schemes so that the number of people who are earning more than 50k can increase.
* As for future scope, if we get more data like if more features are included like persons mentality. Then it will increase the accuracy of the model. Even like how many people are dependent on them.

# REFERENCES

https://www.kaggle.com/datasets/amirhosseinmirzaie/americancitizenincome/data