# Income Level Prediction:

# Decision Tree, Random Forest, Neural Network, and Logistic Regression

**By**

**Ayman Amer Abdulhafed Alkubati**

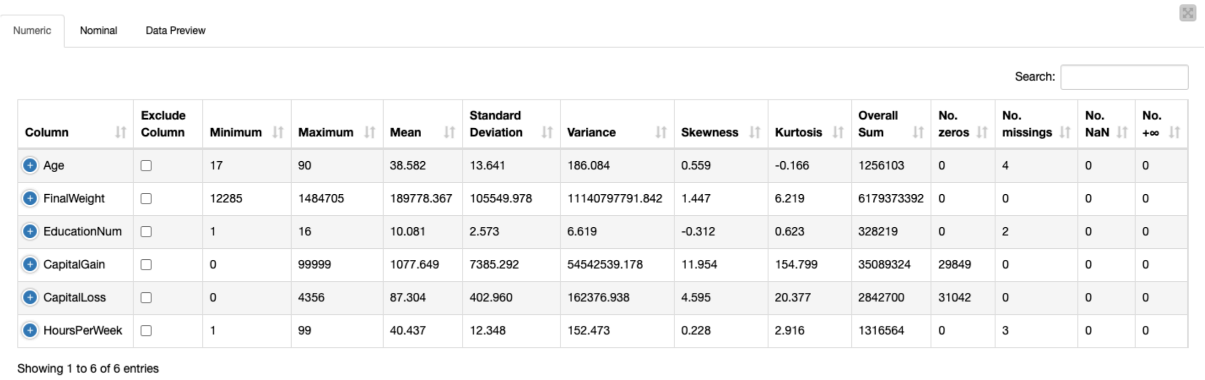
In this report, using the CRISP-DM methodology, a dataset taken from the US census of 1994 database will be analyzed and processed, developing four different models. These are the Decision Tree model (DT), Random Forest (RF) model, Neural Network (MLP type), and Logistic Regression (LR) model, using Knime analytics platform. The CRSISP-DM methodology consists of six steps, namely: business understanding, data understanding, data preparation, model building, testing and evaluation, and deployment.

**Business Understanding:**

Having a dataset extracted from the 1994 US Census database, there are plenty of information to be gathered and many questions can be answered. In our case, focusing on the income level of a particular set of the population, using various variables like work class, education, capital gain and occupation, the annual income of individuals can be predicted to be higher or lower than $50,000. The impact of these different factors that leads to different income levels can be analyzed, whether race, age, marital status play any role in how money one is earning.

**Data Understanding:**

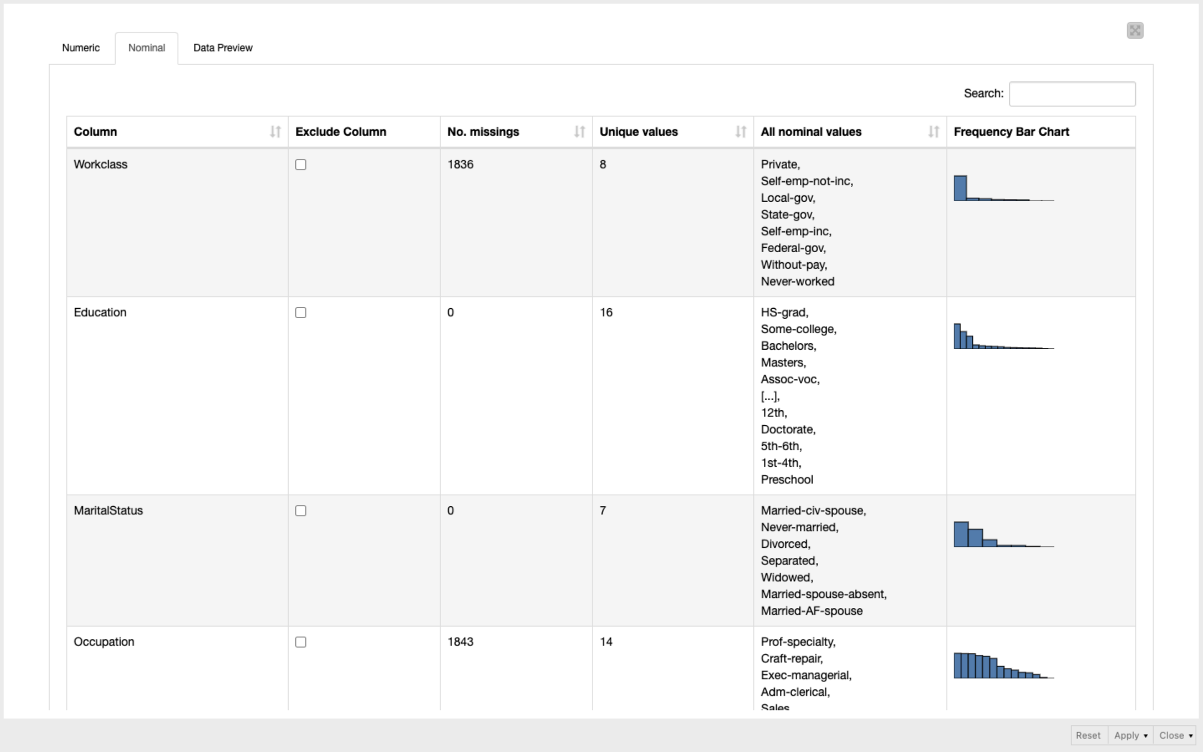
Our dataset contains 32,561 rows and 15 columns varying between numerical variables like age, capital gain and loss, and hours per week worked like it is shown in Figure 1, and categorical nominal data such as the level of education achieved, type of occupation, and race in Figure 2. We also need to understand other features of the dataset such as how many missing values are there in each column, and to pay attention to the skewness of data if there is. All that will need to be dealt with in the next step.



| Column | Exclude Column | Minimum | Maximum | Mean | Standard Deviation | Variance | Skewness | Kurtosis | Overall Sum | No. zeros | No. missings | No. NaN | No. +∞ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Age | □ | 17 | 90 | 38.582 | 13.641 | 186.084 | 0.559 | -0.166 | 1256103 | 0 | 4 | 0 | 0 |
| FinalWeight | □ | 12285 | 1484705 | 189778.367 | 105549.978 | 11140797791.842 | 1.447 | 6.219 | 6179373392 | 0 | 0 | 0 | 0 |
| EducationNum | □ | 1 | 16 | 10.081 | 2.573 | 6.619 | -0.312 | 0.623 | 328219 | 0 | 2 | 0 | 0 |
| CapitalGain | □ | 0 | 99999 | 1077.649 | 7385.292 | 54542539.178 | 11.954 | 154.799 | 35089324 | 29849 | 0 | 0 | 0 |
| CapitalLoss | □ | 0 | 4356 | 87.304 | 402.960 | 162376.938 | 4.595 | 20.377 | 2842700 | 31042 | 0 | 0 | 0 |
| HoursPerWeek | □ | 1 | 99 | 40.437 | 12.348 | 152.473 | 0.228 | 2.916 | 1316564 | 0 | 3 | 0 | 0 |

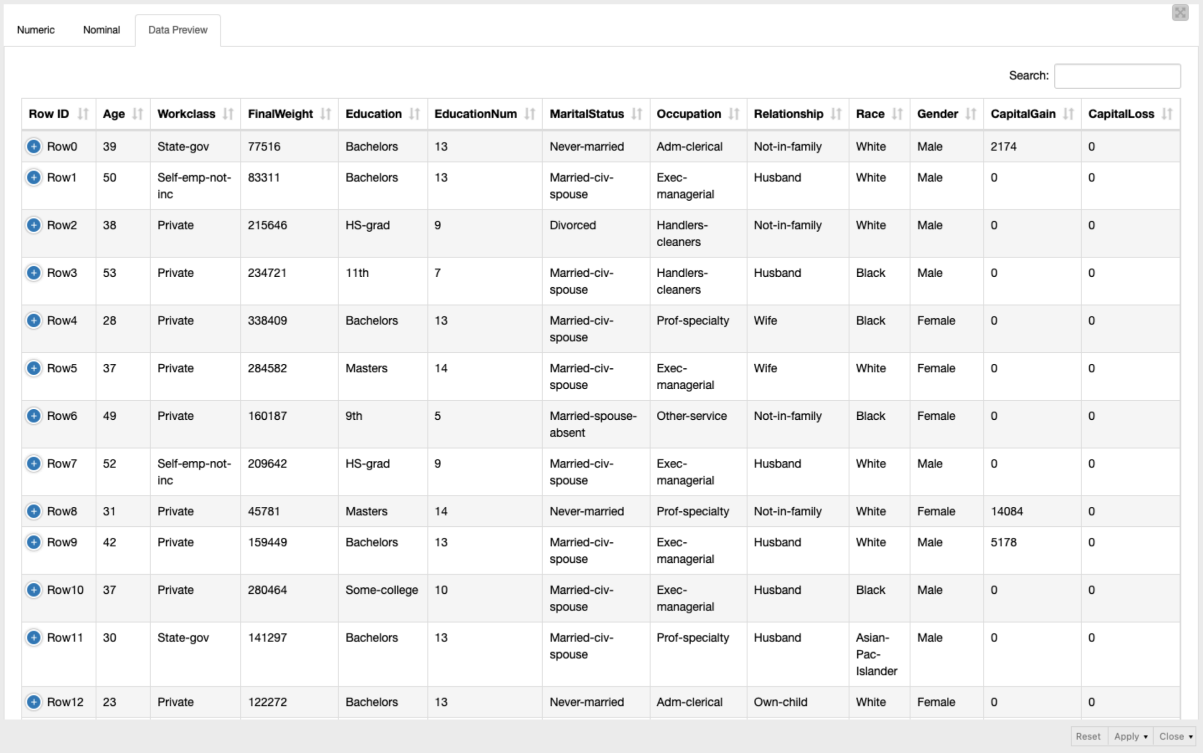
Showing 1 to 6 of 6 entries

Figure 1: Numeric data on Data explorer.



| Column | Exclude Column | No. missings | Unique values | All nominal values | Frequency Bar Chart |
|---|---|---|---|---|---|
| Workclass | ☐ | 1836 | 8 | Private,<br>Self-emp-not-inc,<br>Local-gov,<br>State-gov,<br>Self-emp-inc,<br>Federal-gov,<br>Without-pay,<br>Never-worked | |
| Education | ☐ | 0 | 16 | HS-grad,<br>Some-college,<br>Bachelors,<br>Masters,<br>Assoc-voc,<br>[...],<br>12th,<br>Doctorate,<br>5th-6th,<br>1st-4th,<br>Preschool | |
| MaritalStatus | ☐ | 0 | 7 | Married-civ-spouse,<br>Never-married,<br>Divorced,<br>Separated,<br>Widowed,<br>Married-spouse-absent,<br>Married-AF-spouse | |
| Occupation | ☐ | 1843 | 14 | Prof-specialty,<br>Craft-repair,<br>Exec-managerial,<br>Adm-clerical,<br>Sales | |

Figure 2: Nominal data on the Data explorer node.



| Row ID | Age | Workclass | FinalWeight | Education | EducationNum | MaritalStatus | Occupation | Relationship | Race | Gender | CapitalGain | CapitalLoss |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Row0 | 39 | State-gov | 77516 | Bachelors | 13 | Never-married | Adm-clerical | Not-in-family | White | Male | 2174 | 0 |
| Row1 | 50 | Self-emp-not-inc | 83311 | Bachelors | 13 | Married-civ-spouse | Exec-managerial | Husband | White | Male | 0 | 0 |
| Row2 | 38 | Private | 215646 | HS-grad | 9 | Divorced | Handlers-cleaners | Not-in-family | White | Male | 0 | 0 |
| Row3 | 53 | Private | 234721 | 11th | 7 | Married-civ-spouse | Handlers-cleaners | Husband | Black | Male | 0 | 0 |
| Row4 | 28 | Private | 338409 | Bachelors | 13 | Married-civ-spouse | Prof-specialty | Wife | Black | Female | 0 | 0 |
| Row5 | 37 | Private | 284582 | Masters | 14 | Married-civ-spouse | Exec-managerial | Wife | White | Female | 0 | 0 |
| Row6 | 49 | Private | 160187 | 9th | 5 | Married-spouse-absent | Other-service | Not-in-family | Black | Female | 0 | 0 |
| Row7 | 52 | Self-emp-not-inc | 209642 | HS-grad | 9 | Married-civ-spouse | Exec-managerial | Husband | White | Male | 0 | 0 |
| Row8 | 31 | Private | 45781 | Masters | 14 | Never-married | Prof-specialty | Not-in-family | White | Female | 14084 | 0 |
| Row9 | 42 | Private | 159449 | Bachelors | 13 | Married-civ-spouse | Exec-managerial | Husband | White | Male | 5178 | 0 |
| Row10 | 37 | Private | 280464 | Some-college | 10 | Married-civ-spouse | Exec-managerial | Husband | Black | Male | 0 | 0 |
| Row11 | 30 | State-gov | 141297 | Bachelors | 13 | Married-civ-spouse | Prof-specialty | Husband | Asian-Pac-Islander | Male | 0 | 0 |
| Row12 | 23 | Private | 122272 | Bachelors | 13 | Never-married | Adm-clerical | Own-child | White | Female | 0 | 0 |

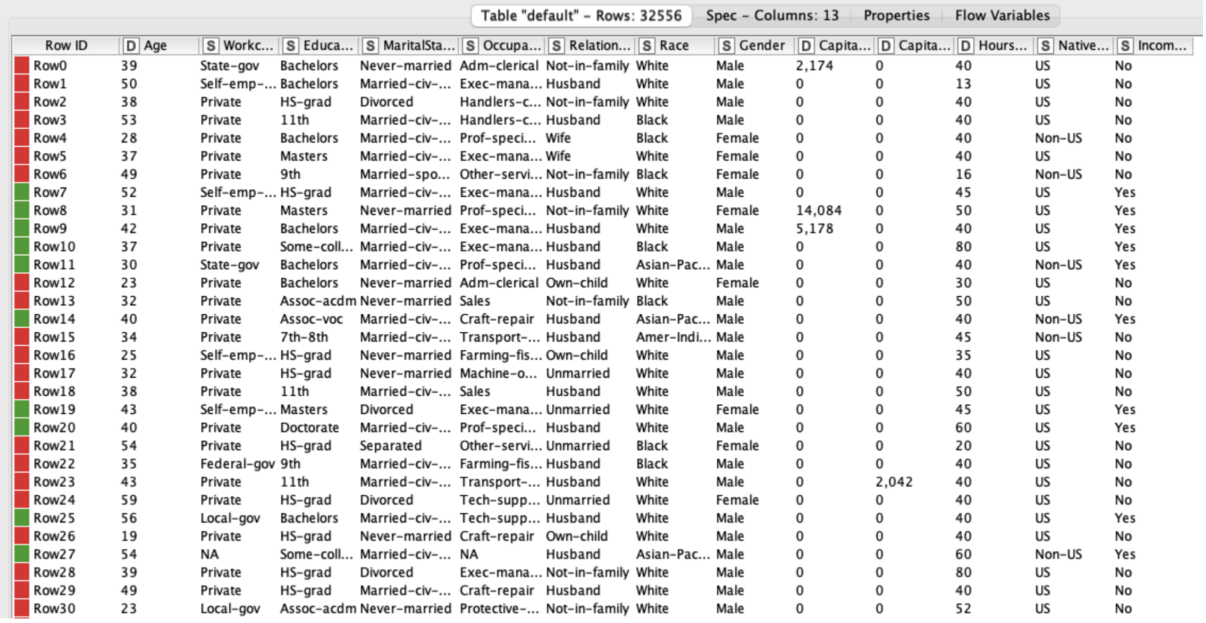
Figure 3: Data Preview on the Data explorer node.

**Data Preparation:**

Filtered out two columns, final weight for it being sampling data which does not carry valuable information for us in this analysis and education due to the fact that there are two columns serving the same purpose, Education and EducationNum, we will exclude the latter. Ending up with a dataset that has 13 columns instead of 15.

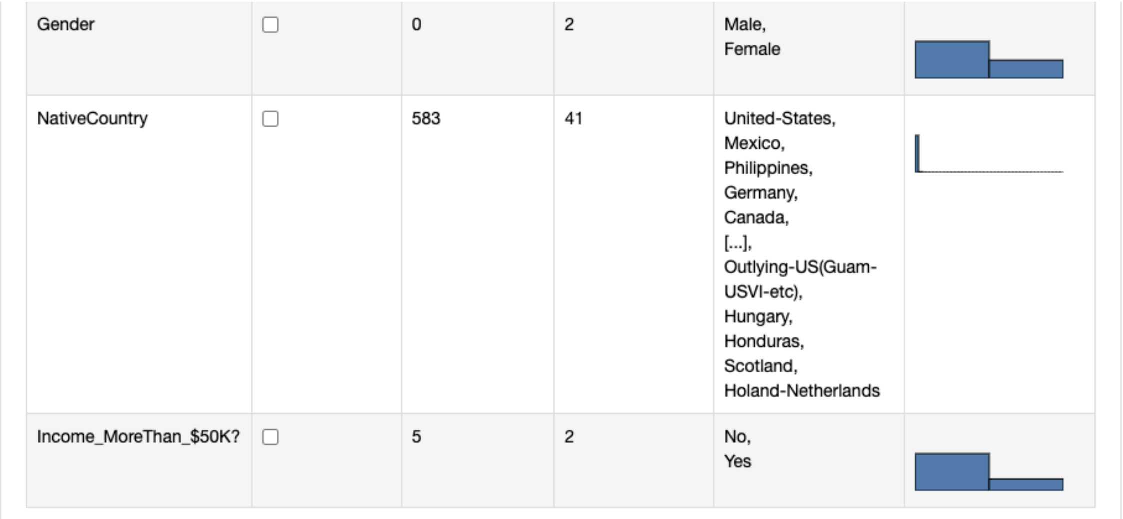
The imputation of missing values is also conducted to make sure these values does not have a major effect on the accuracy of the model we are building. As we do not need to deal with missing numerical values for them not existing in this particular dataset, the missing strings will be replaced with the word "Missing", making sure that we do not exclude these rows and actually make sense of the logic behind these missing values if there was any. Another way to make the distinction clear between our two addressed values of income level, above and below $50,000, we can add a colored layer on each row, let it be green and red colors to distinguish between each as depicted in Figure 4.

When checking how each country is represented in the dataset, as shown in Figure 5, we find that the NativeCountry column shows that the United States dominates the number of values and other countries appear less frequently. We can do that by creating two categories, one for United States called US, and another which consists of all the other countries grouped under one value called Non-US.



| Row ID | D Age | S Workc... | S Educa... | S MaritalSta... | S Occupa... | S Relation... | S Race | S Gender | D Capita... | D Capita... | D Hours... | S Native... | S Incom... |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 🟥 Row0 | 39 | State-gov | Bachelors | Never-married | Adm-clerical | Not-in-family | White | Male | 2,174 | 0 | 40 | US | No |
| 🟥 Row1 | 50 | Self-emp-... | Bachelors | Married-civ-... | Exec-mana... | Husband | White | Male | 0 | 0 | 13 | US | No |
| 🟥 Row2 | 38 | Private | HS-grad | Divorced | Handlers-c... | Not-in-family | White | Male | 0 | 0 | 40 | US | No |
| 🟥 Row3 | 53 | Private | 11th | Married-civ-... | Handlers-c... | Husband | Black | Male | 0 | 0 | 40 | US | No |
| 🟥 Row4 | 28 | Private | Bachelors | Married-civ-... | Prof-speci... | Wife | Black | Female | 0 | 0 | 40 | Non-US | No |
| 🟥 Row5 | 37 | Private | Masters | Married-civ-... | Exec-mana... | Wife | White | Female | 0 | 0 | 40 | US | No |
| 🟥 Row6 | 49 | Private | 9th | Married-spo... | Other-servi... | Not-in-family | Black | Female | 0 | 0 | 16 | Non-US | No |
| 🟩 Row7 | 52 | Self-emp-... | HS-grad | Married-civ-... | Exec-mana... | Husband | White | Male | 0 | 0 | 45 | US | Yes |
| 🟩 Row8 | 31 | Private | Masters | Never-married | Prof-speci... | Not-in-family | White | Female | 14,084 | 0 | 50 | US | Yes |
| 🟩 Row9 | 42 | Private | Bachelors | Married-civ-... | Exec-mana... | Husband | White | Male | 5,178 | 0 | 40 | US | Yes |
| 🟩 Row10 | 37 | Private | Some-coll... | Married-civ-... | Exec-mana... | Husband | Black | Male | 0 | 0 | 80 | US | Yes |
| 🟩 Row11 | 30 | State-gov | Bachelors | Married-civ-... | Prof-speci... | Husband | Asian-Pac... | Male | 0 | 0 | 40 | Non-US | Yes |
| 🟥 Row12 | 23 | Private | Bachelors | Never-married | Adm-clerical | Own-child | White | Female | 0 | 0 | 30 | US | No |
| 🟥 Row13 | 32 | Private | Assoc-acdm | Never-married | Sales | Not-in-family | Black | Male | 0 | 0 | 50 | US | No |
| 🟩 Row14 | 40 | Private | Assoc-voc | Married-civ-... | Craft-repair | Husband | Asian-Pac... | Male | 0 | 0 | 40 | Non-US | Yes |
| 🟥 Row15 | 34 | Private | 7th-8th | Married-civ-... | Transport-... | Husband | Amer-Indi... | Male | 0 | 0 | 45 | Non-US | No |
| 🟥 Row16 | 25 | Self-emp-... | HS-grad | Never-married | Farming-fis... | Own-child | White | Male | 0 | 0 | 35 | US | No |
| 🟥 Row17 | 32 | Private | HS-grad | Never-married | Machine-o... | Unmarried | White | Male | 0 | 0 | 40 | US | No |
| 🟥 Row18 | 38 | Private | 11th | Married-civ-... | Sales | Husband | White | Male | 0 | 0 | 50 | US | No |
| 🟩 Row19 | 43 | Self-emp-... | Masters | Divorced | Exec-mana... | Unmarried | White | Female | 0 | 0 | 45 | US | Yes |
| 🟩 Row20 | 40 | Private | Doctorate | Married-civ-... | Prof-speci... | Husband | White | Male | 0 | 0 | 60 | US | Yes |
| 🟥 Row21 | 54 | Private | HS-grad | Separated | Other-servi... | Unmarried | Black | Female | 0 | 0 | 20 | US | No |
| 🟥 Row22 | 35 | Federal-gov | 9th | Married-civ-... | Farming-fis... | Husband | Black | Male | 0 | 0 | 40 | US | No |
| 🟥 Row23 | 43 | Private | 11th | Married-civ-... | Transport-... | Husband | White | Male | 0 | 2,042 | 40 | US | No |
| 🟥 Row24 | 59 | Private | HS-grad | Divorced | Tech-supp... | Unmarried | White | Female | 0 | 0 | 40 | US | No |
| 🟩 Row25 | 56 | Local-gov | Bachelors | Married-civ-... | Tech-supp... | Husband | White | Male | 0 | 0 | 40 | US | Yes |
| 🟥 Row26 | 19 | Private | HS-grad | Never-married | Craft-repair | Own-child | White | Male | 0 | 0 | 40 | US | No |
| 🟩 Row27 | 54 | NA | Some-coll... | Married-civ-... | NA | Husband | Asian-Pac... | Male | 0 | 0 | 60 | Non-US | Yes |
| 🟥 Row28 | 39 | Private | HS-grad | Divorced | Exec-mana... | Not-in-family | White | Male | 0 | 0 | 80 | US | No |
| 🟥 Row29 | 49 | Private | HS-grad | Married-civ-... | Craft-repair | Husband | White | Male | 0 | 0 | 40 | US | No |
| 🟥 Row30 | 23 | Local-gov | Assoc-acdm | Never-married | Protective-... | Not-in-family | White | Male | 0 | 0 | 52 | US | No |

Figure 4: Colored Rows based on IncomeLevel values.



| | | | | | |
|---|---|---|---|---|---|
| Gender | □ | 0 | 2 | Male,<br>Female | |
| NativeCountry | □ | 583 | 41 | United-States,<br>Mexico,<br>Philippines,<br>Germany,<br>Canada,<br>[...],<br>Outlying-US(Guam-USVI-etc),<br>Hungary,<br>Honduras,<br>Scotland,<br>Holand-Netherlands | |
| Income_MoreThan_$50K? | □ | 5 | 2 | No,<br>Yes | |

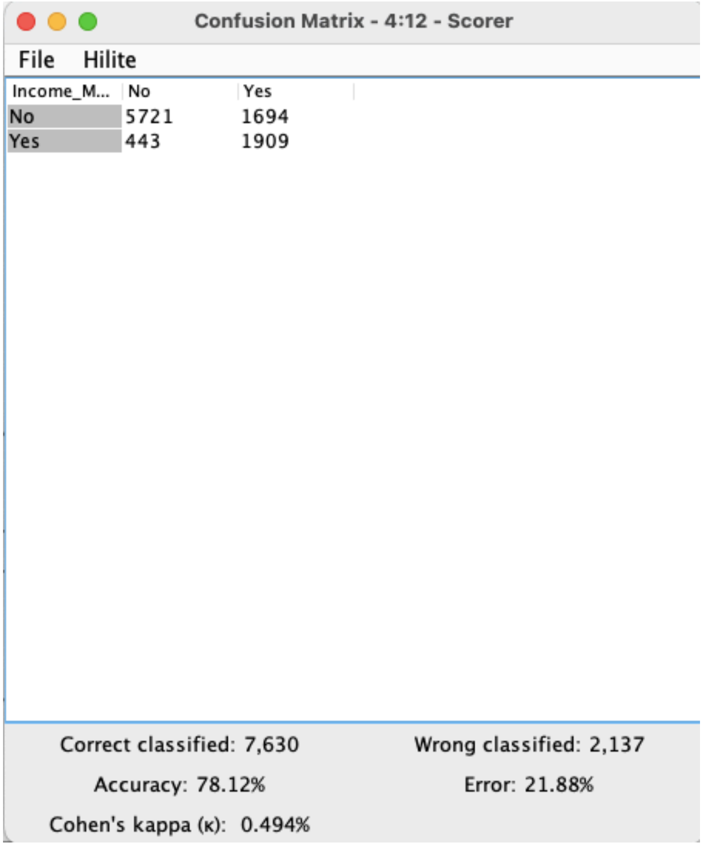
Figure 5: NativeCountry column.

**Model Building and Testing:**

A- Decision Tree Model:

After understanding the business problem, exploring and transforming the data, according to the CRISP-DM methodology we can continue with the fourth step which is working on building our models. We start with partitioning the data into two groups, one for training which is 70% of the whole dataset, and a testing set of 30%, stratified on IncomeLevel column, while using a random seed value of 12345.

In the next step, we built a Decision Tree model, having the target variable being the column named IncomeLevel. To evaluate the model and check its accuracy, Knime offers various tools, in this analysis we can use a tool called Scorer and also plotting it on the ROC curve. For the low specificity where the minority class of values get predicted with a lower success rate of around 64%, we decided to choose equal size sampling to help balancing the training data, and we can see the new accuracy rate in Figure 6. While Figure 7 shows the plotted ROC curve derived from this decision tree model. Lastly, the Decision Tree graphical model of the first three levels is shown in Figure 8.



| Income_M... | No | Yes |
|---|---|---|
| No | 5721 | 1694 |
| Yes | 443 | 1909 |

Correct classified: 7,630 Wrong classified: 2,137

Accuracy: 78.12% Error: 21.88%

Cohen's kappa (κ): 0.494%

Figure 6: Decision Tree Scorer.

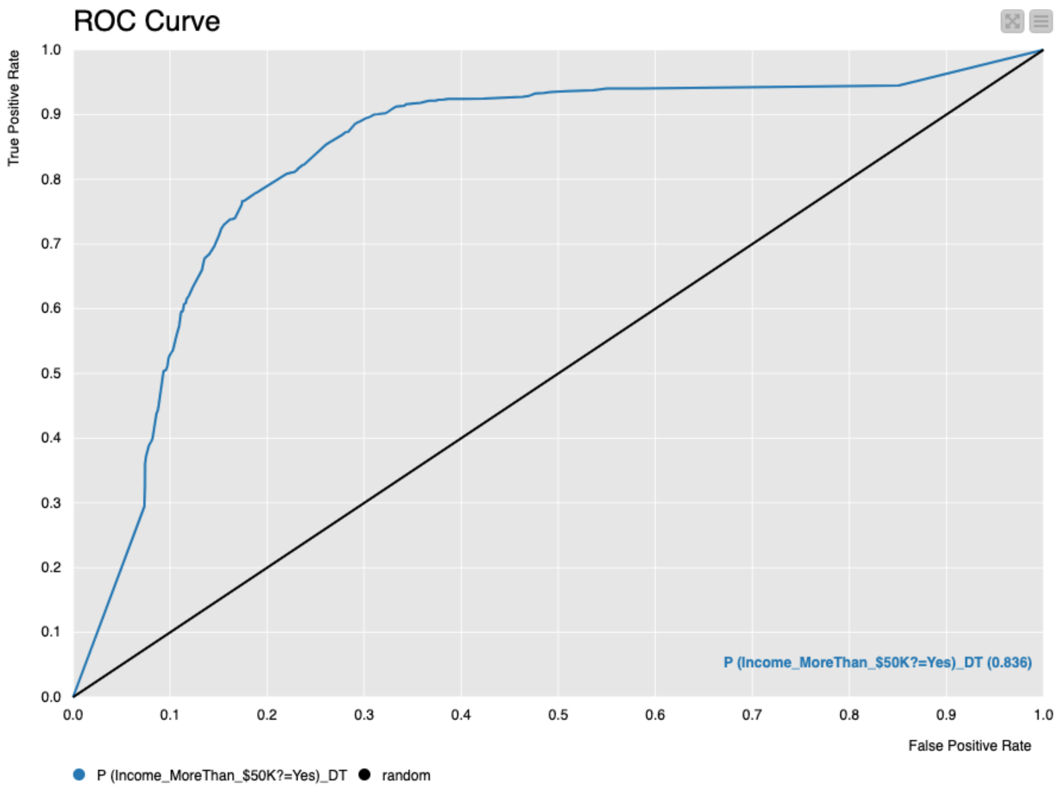
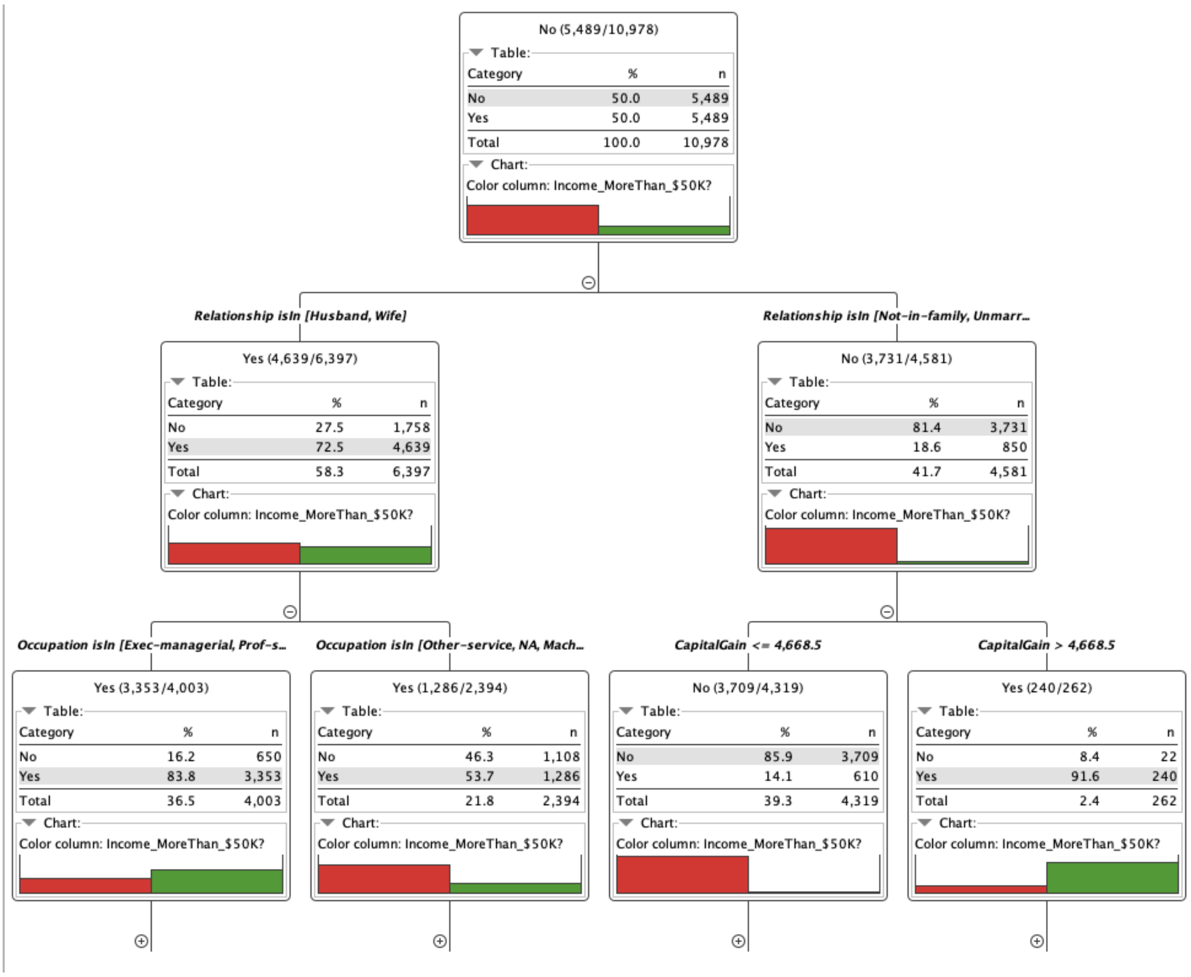


Figure 7: Decision Tree ROC Curve.



No (5,489/10,978)

Table:

| Category | % | n |
|---|---|---|
| No | 50.0 | 5,489 |
| Yes | 50.0 | 5,489 |
| Total | 100.0 | 10,978 |

Chart:

Color column: Income_MoreThan_$50K?

Relationship isIn [Husband, Wife]

Yes (4,639/6,397)

Table:

| Category | % | n |
|---|---|---|
| No | 27.5 | 1,758 |
| Yes | 72.5 | 4,639 |
| Total | 58.3 | 6,397 |

Chart:

Color column: Income_MoreThan_$50K?

Relationship isIn [Not-in-family, Unmarr...

No (3,731/4,581)

Table:

| Category | % | n |
|---|---|---|
| No | 81.4 | 3,731 |
| Yes | 18.6 | 850 |
| Total | 41.7 | 4,581 |

Chart:

Color column: Income_MoreThan_$50K?

Occupation isIn [Exec-managerial, Prof-s...

Yes (3,353/4,003)

Table:

| Category | % | n |
|---|---|---|
| No | 16.2 | 650 |
| Yes | 83.8 | 3,353 |
| Total | 36.5 | 4,003 |

Chart:

Color column: Income_MoreThan_$50K?

Occupation isIn [Other-service, NA, Mach...

Yes (1,286/2,394)

Table:

| Category | % | n |
|---|---|---|
| No | 46.3 | 1,108 |
| Yes | 53.7 | 1,286 |
| Total | 21.8 | 2,394 |

Chart:

Color column: Income_MoreThan_$50K?

CapitalGain <= 4,668.5

No (3,709/4,319)

Table:

| Category | % | n |
|---|---|---|
| No | 85.9 | 3,709 |
| Yes | 14.1 | 610 |
| Total | 39.3 | 4,319 |

Chart:

Color column: Income_MoreThan_$50K?

CapitalGain > 4,668.5

Yes (240/262)

Table:

| Category | % | n |
|---|---|---|
| No | 8.4 | 22 |
| Yes | 91.6 | 240 |
| Total | 2.4 | 262 |

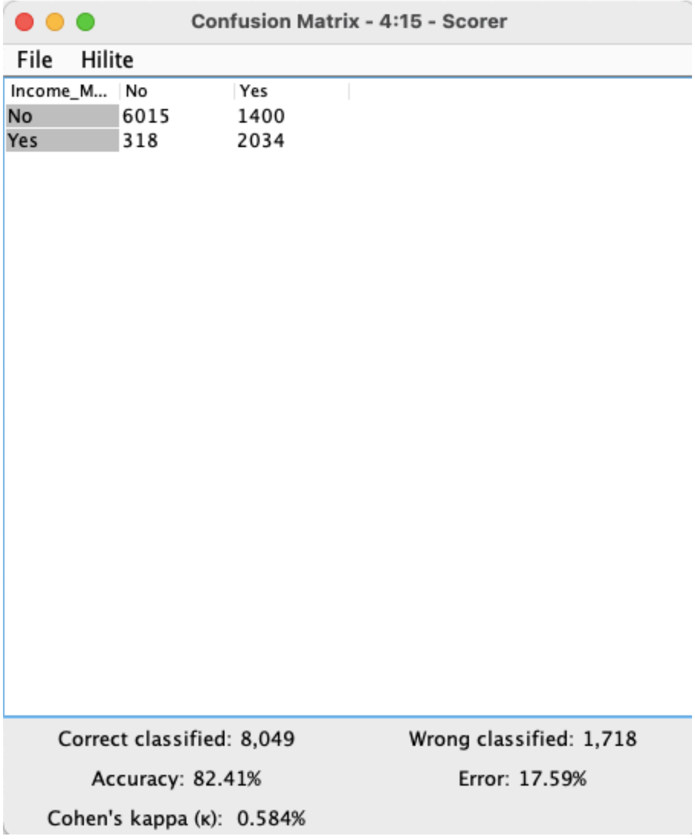
Chart:

Color column: Income_MoreThan_$50K?

Figure 8: Three levels Decision Tree.

B- Random Forest Model:

Similar to how we created the Decision Tree model, we can use the exact same settings we used to create the Random Forest model to be able to compare the accuracy that each model produce. And as the fifth step of the CRISP-DM methodology focuses on testing and evaluating the output of the newly built models, we will use the two tools we used previously, the Scorer and ROC Curve, displayed in Figure 9 and 10 respectively.



| Income_M... | No | Yes |
|---|---|---|
| No | 6015 | 1400 |
| Yes | 318 | 2034 |

Correct classified: 8,049 Wrong classified: 1,718

Accuracy: 82.41% Error: 17.59%

Cohen's kappa (κ): 0.584%

Figure 9: Random Forest Scorer.

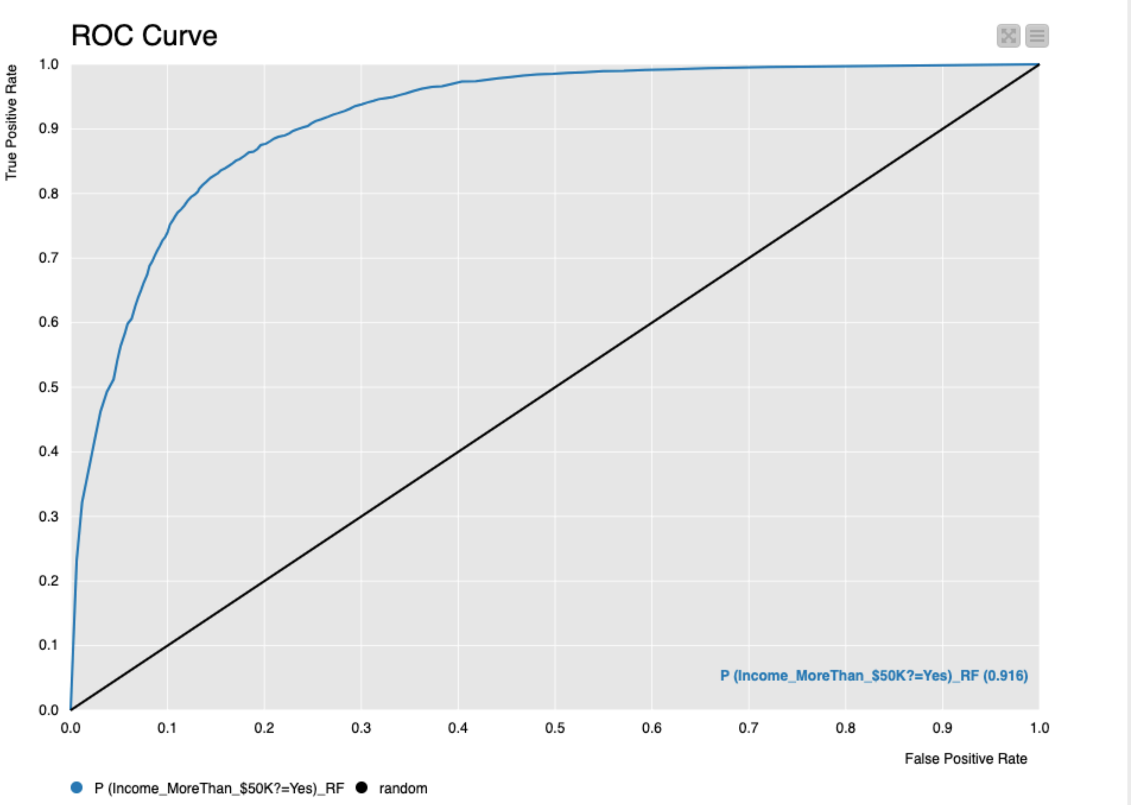
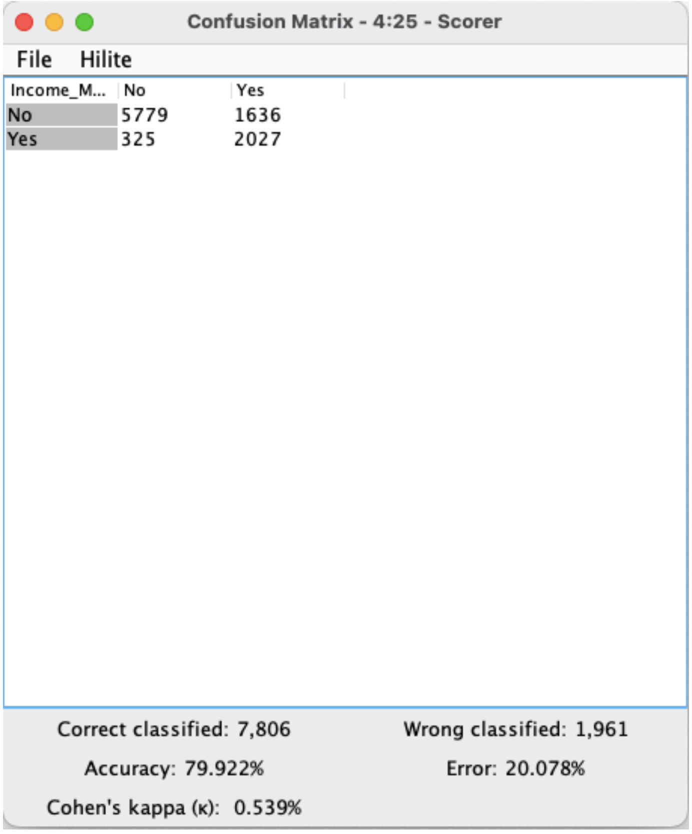


Figure 10: Random Forest ROC Curve.

C- Neural Network (MLP type) Model:

The third model utilizes a neural network (MLP) structured with multiple connected nodes to preprocess the data before feeding it into the machine learning algorithm. Specifically, the variables and data are converted into numerical categories via a one-to-many node. The data then passes through a normalization node to standardize the data range. After preprocessing, the data goes into partition and MLP learner nodes, which have the IncomeLevel column set as the target variable. Accuracy metrics are generated by ROC Curve and Scorer nodes in KNIME. The full workflow with intermediate outputs is visualized in the accompanying figures.



| Income_M... | No | Yes |
|---|---|---|
| No | 5779 | 1636 |
| Yes | 325 | 2027 |

Correct classified: 7,806

Wrong classified: 1,961

Accuracy: 79.922%

Error: 20.078%

Cohen's kappa (κ): 0.539%

Figure 11: Neural Network Scorer.

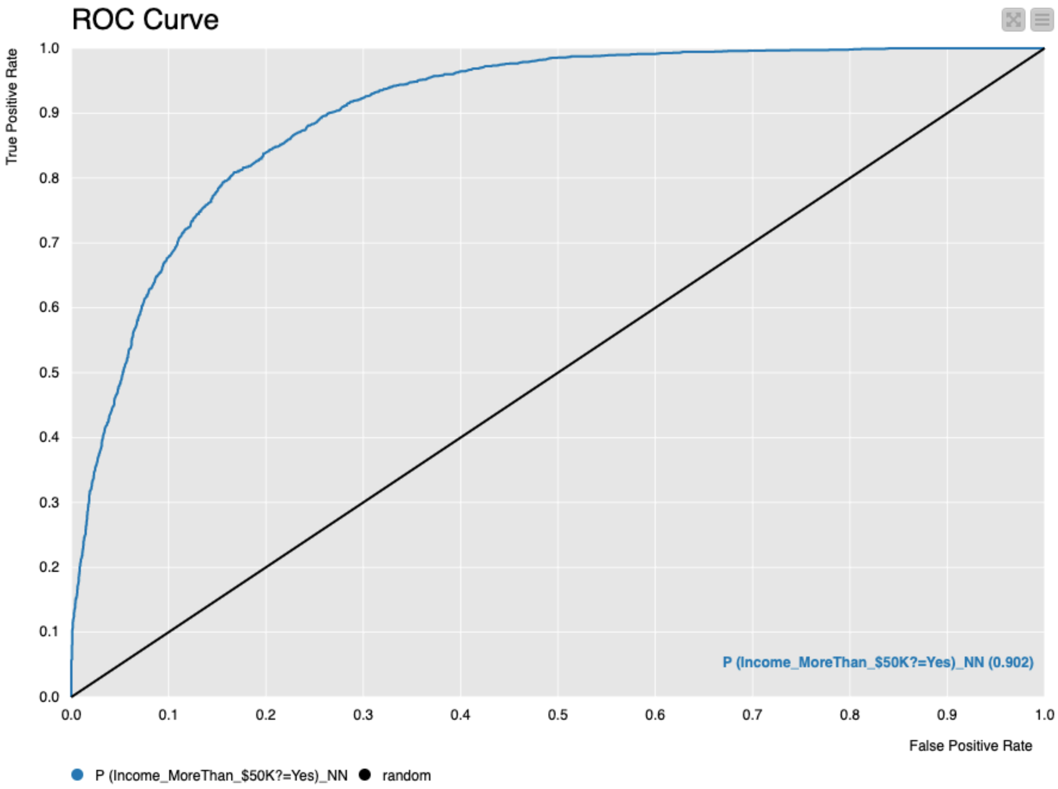
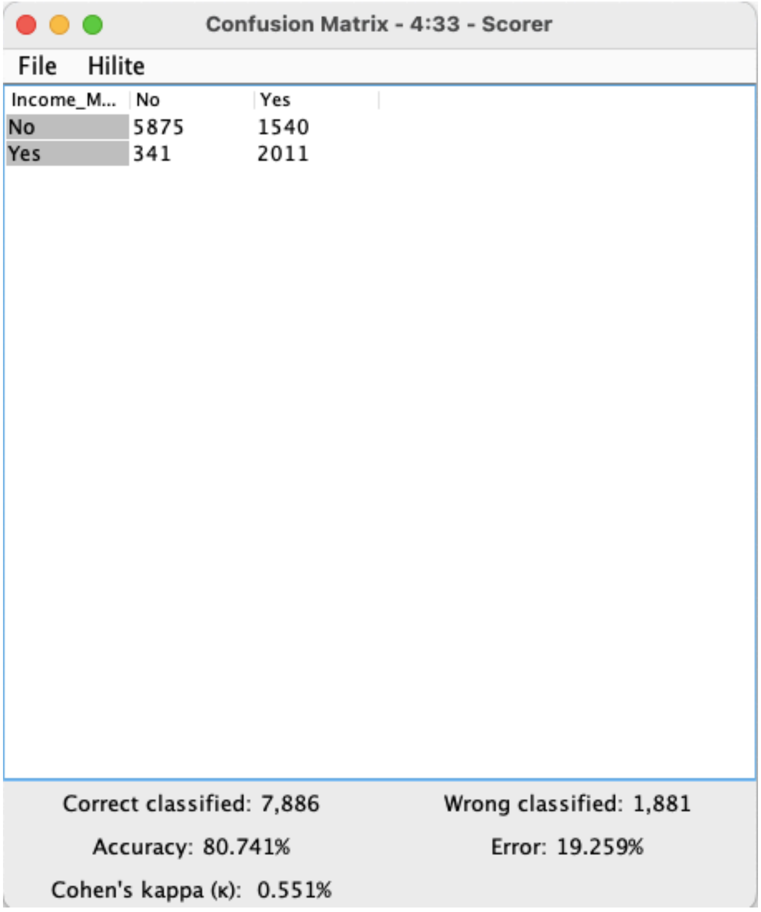


Figure 12: Neural Network ROC Curve.

D- Logistic Regression (LR) Model:

The final model employs logistic regression (LR), structured in KNIME with a series of nodes for data preprocessing before the machine learning algorithm. As in the prior neural network model, the raw data is converted to numerical categories via a one-to-many node, then normalized across variables through a normalization node. The processed data is partitioned and fed into the LR learner, with the IncomeLevel set again as target output variable. Model accuracy is evaluated by routing the LP model predictions into ROC Curve and Scorer nodes native to KNIME. These provide numeric metrics and visual evaluation of how well income level is predicted. The complete workflow - data transformations, LP model building, accuracy checking - is visualized in the accompanying figures within the KNIME interface.



| Income_M... | No | Yes |
|---|---|---|
| No | 5875 | 1540 |
| Yes | 341 | 2011 |

Correct classified: 7,886 Wrong classified: 1,881

Accuracy: 80.741% Error: 19.259%

Cohen's kappa (κ): 0.551%

Figure 13: Logistic Regression Scorer.

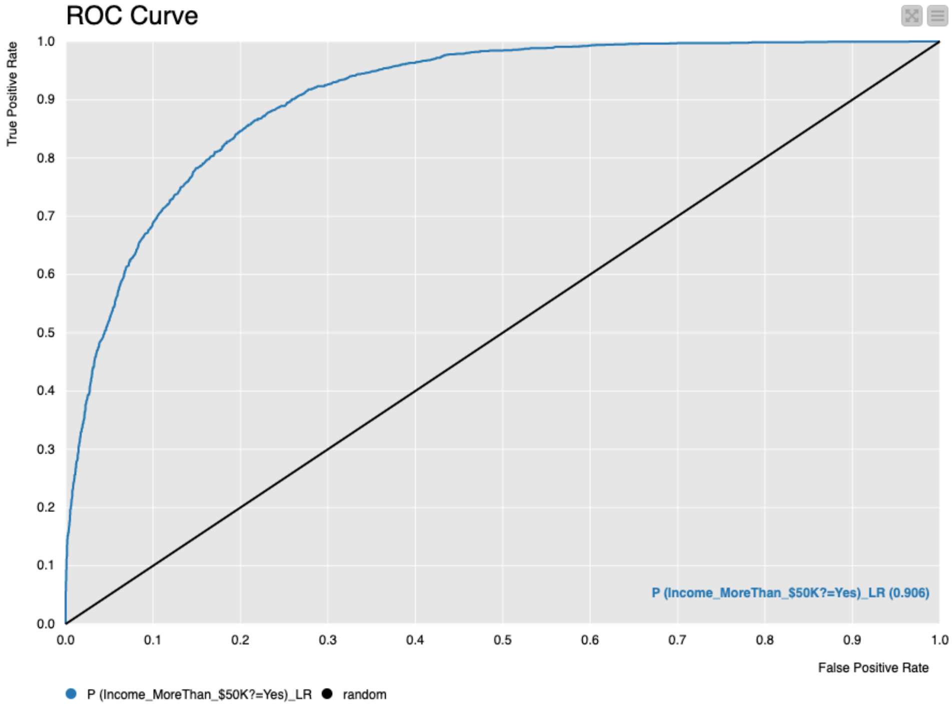


Figure 14: Logistic Regression ROC Curve.

**Testing and Evaluation:**

To compare the performances of all the models we used in this report, Decision Tree, Random Forest, Neural Network, and Logistic Regression, we created an ROC curve that combines the output of each into one chart, Figure 15. The Random Forest model performed better than the other models with 0.916 accuracy. On the other hand, the model that yielded the lowest prediction results is Decision Tree model with 0.836 accuracy. Therefore, the Random Forest model is preferred to the other models we built for having more accuracy and less error rate.

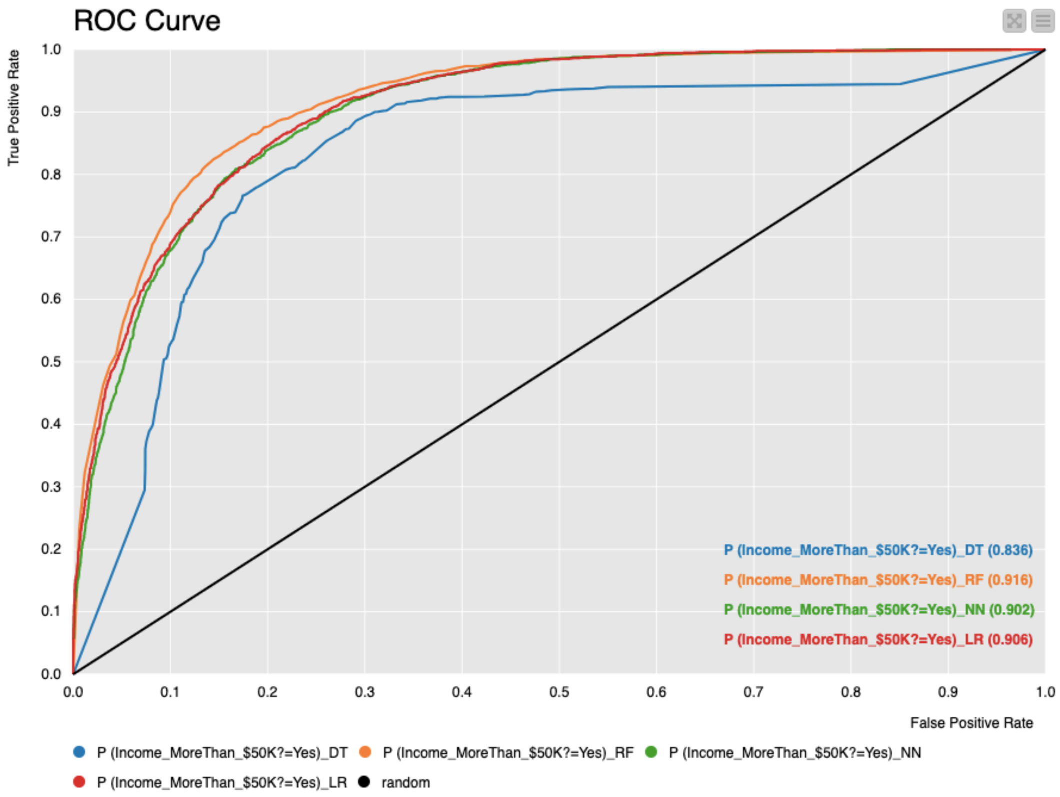


Figure 15: ROC Curve combining all the models.

**Deployment**:

In the sixth and final step of the CRISP-DM methodology, deployment of the superior model can take place to support decision making processes. In the graph below, Figure 16 shows the Variable Importance chart which shows how much the model is utilizing each variable to make accurate predictions. In our Random Forest model used in this analysis, capital gain was the most important, followed by relationship and marital status of individuals participated in the consensus data we used, while the native country, race, and gender had the least influence.

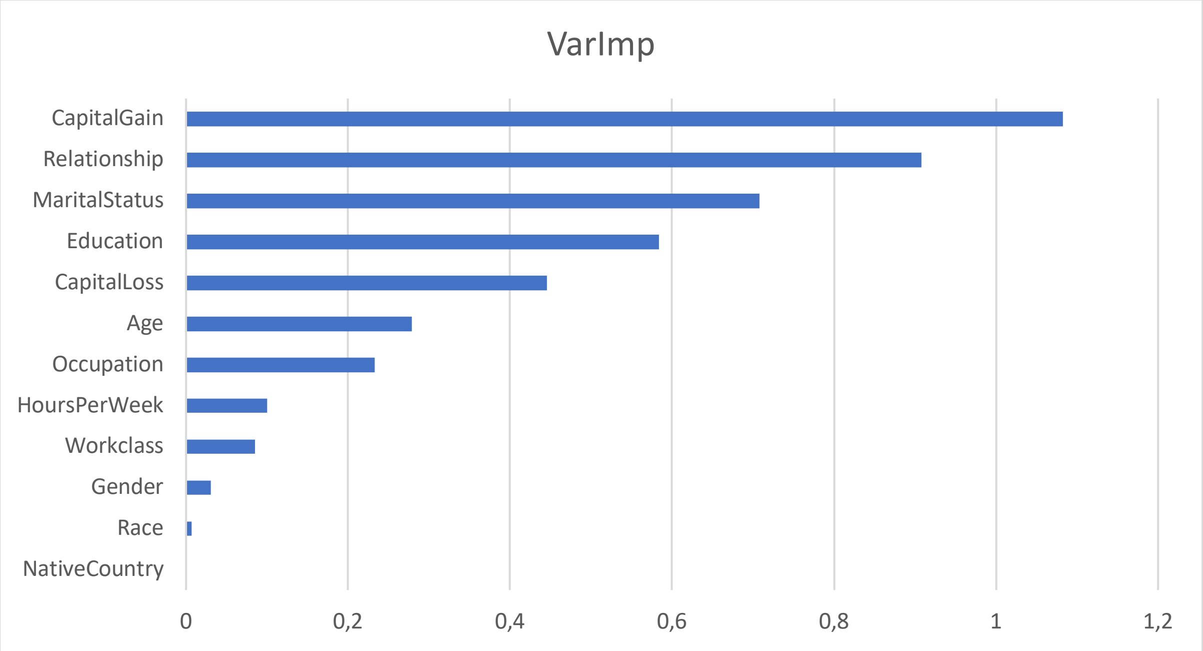


Figure 16: Random Forest Importance Graph

In conclusion, this report utilized the CRISP-DM methodology to develop and evaluate models for predicting income level from US census data. Four models were built and tested: Decision Tree, Random Forest, Neural Network, and Logistic Regression. The process followed the phases of CRISP-DM - establishing the business need, understanding the data, preparing the data, training and tuning models, and evaluating performance.

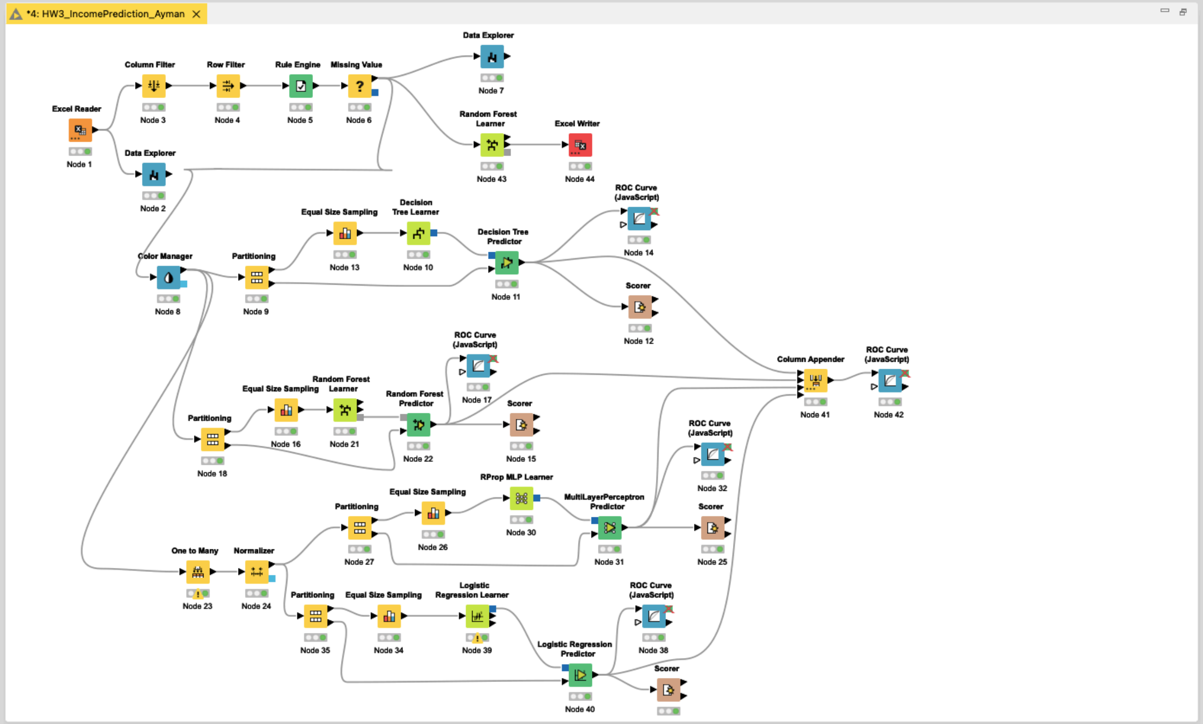


Figure 17: Knime Workflow.