Logo

Description automatically generated

ALY6020 Module 1 Project

**Authored by: Prateek Singh**

Course: ALY6020 – Predictive Analytics

Submission Date: 04/19/2022

**Instructor’s name: Kristen Drobnis**

# Introduction

In this assignment we are modeling a KNN or K Nearest Neighbor algorithm to classify the Income earned by the citizens of US across all the races into two groups, namely Less than or equal to 50K annual income and Greater than 50K annual income. We wanted to build a model which can accurately classify citizens under the mentioned groups or cluster to ensure equal pay opportunities. We were determined to understand all the factors which affects the Income in US, hence number of bar graphs were generated to see the distribution of all the independent variables data against the average money earned by the citizens. Income is our dependent variable which is our Target or Label for the experiment.

The Dataset contains demographic and financial information of the citizens with a total of 15 variables and 48842 observations. As a part of dataset cleansing, we assigned the header names for the columns as the dataset was missing it. We also checked for the missing values and replaced them with the mode value of the columns because the columns were having categorical values. All the variables doesn’t seem to be affecting the Income of the citizens and hence variables like Final\_Weight, Education, Capital\_Gain and Capital\_Loss were dropped from the dataset. Education variable was dropped because we have Education\_Num as another supporting variable and it is creating the redundancy in the dataset. Also, Education was categorical and Education\_Num was numerical variable hence the decision to keep Education\_Num was taken. We used Jupyter Notebook and Python to accomplish the assignment.

# Analysis

In our dataset we have total of 6 Numeric and 9 Object variables. Upon cleansing the dataset, we were left with 3 Numeric and 8 Object Variables. On an average citizen of US in the dataset age 39 years old, work for an average of 40 hours a week and commonly has at least “HS-grad” education. Maximum citizens in the dataset belongs to Private work class and belongs to “Less than and equal to 50K” Income category. (*Refer “Fig 1: Statistical Dataset Analysis of Numeric Series” and “Fig 2: Statistical Dataset Analysis of Object Series” in appendix*)

Variables WorkClass, Occupation and Native\_Country had missing values in form of “?” which we got rid by first replacing it with NAN values and then dropping them. We also verified if there was any missing information left in the dataset. The comparison can be seen in the Fig 3 and Fig 4. (*Refer “Fig 3: Checking for the missing or null values” and “Fig 4: Verifying if all the Missing values were converted and replaced” in appendix*)

Finally, after cleansing the dataset we retrieved the factor or unique values for all the Object variables and assigned them the continuous labels starting from indices 0 before mapping it to our original dataset. (*Refer “Fig 5: Labels assigned for all the factors of Object Variables” in appendix*)

Finally, we wanted to visually comprehend that which particular factors has maximum influence in the affluency of the income and hence we plot several bar plots of those factors against the average Income. We can see that a self-employed individual earns maximum income when compared with someone who works without pay or has never worked. (*Refer “Fig 6: Working Class vs Average Income” in appendix*)

We can see that as the level of education increases the average income of the citizens are also increasing. (*Refer “Fig 7: Education vs Average Income” in appendix*)

Few other variables were compared against the average income but no concrete conclusions can be derived, however we understood that the maximum population of the working citizens in US is dominated by Males. (*Refer “Fig 8: Marital Status vs Average Income”, “Fig 9: Occupation vs Average Income”, “Fig 10: Relationship vs Average Income”, “Fig 11: Race vs Average Income” and “Fig 12: Gender vs Average Income” in appendix*)

We finally plotted hours worked per week by the citizens against the average income earned and realize that the bar is maximum at 40 hours which makes sense as that is the most common working hours in US. (*Refer “Fig 13: Hours Worked Per Week vs Average Income” in appendix*)

Coming to building the KNN Algorithm model, I took the sample size of 5000 (which is approximately 10% of the entire dataset) setting the seed as 42. The Sample set was then divided into independent variable set and the target variable set. These sets are used to create the training dataset to train the model and test dataset to test the trained model. Train and Test dataset was divided into 7:3 ratio respectively. Both the dataset has train set data, train set labels, test set data and test set labels. Then we plot the scatter plot of both the income levels to find how our labels are distributed. We can clearly see that the income level which is less than or equal to 50K is dominantly distributed as discusses earlier and there are lot of overlap points. (*Refer “Fig 14: Distribution of two Labels of Income earned (above 50K or below 50K)” in appendix*) This can prove to be difficult for our model to be able correctly predict and classify the unknown datapoint.

Further we build our model by calculating the nearest neighbors distance and testing several values of K, we identified our best working model. We started with using square root of sample size () of test data set which comes out to be 39. Then altered couple of values up and down to see the accuracy of the model. We cannot afford to keep low values of K such as 3,6 or 10 because of the large dataset observations and it will not be able to justify the predictions made.

# Conclusion

Upon analyzing all the values of K, we found the accuracy of our model as below,

* K = 36; Accuracy = 77.9%
* K = 37; Accuracy = 78.1%
* K = 38; Accuracy = 77.8%
* K = 39; Accuracy = 78%
* K = 40; Accuracy = 78%

We can conclude to keep the value of K as 37 as it yield the maximum accuracy for the model.

# Reference

1. Grosz, J. (n.d.). Working With Nearest Neighbors. Canvas. Retrieved April 16, 2022, from <https://northeastern.instructure.com/courses/105748/pages/lesson-1-5-running-the-algorithm?module_item_id=7143680>
2. Load and return the iris dataset. scikit. (n.d.). Retrieved April 16, 2022, from <https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_iris.html>
3. Pingulkar, H. (2021, October 22). Scikit-Learn: NAMEERROR: Name 'Load\_iris' is not defined. Stack Overflow. Retrieved April 16, 2022, from <https://stackoverflow.com/questions/69675761/scikit-learn-nameerror-name-load-iris-is-not-defined>
4. Tournoij, M. (2015, March 17). TypeError: 'function' object is not subscriptable - Python. Stack Overflow. Retrieved April 16, 2022, from <https://stackoverflow.com/questions/29101836/typeerror-function-object-is-not-subscriptable-python>
5. Kilian Q. Weinberger, John Blitzer and Lawrence K. Saul (2005), Distance Metric Learning for Large Margin Nearest Neighbor Classification retrieved on 18th July,2021 from <https://proceedings.neurips.cc/paper/2005/file/a7f592cef8b130a6967a90617db5681b-Paper.pdf>
6. BitDegree. (2019, November 25). A guide on splitting datasets with Train\_test\_split function. BitDegree. Retrieved April 19, 2022, from <https://www.bitdegree.org/learn/train-test-split>

# Appendix

Table

Description automatically generated

Fig 1: Statistical Dataset Analysis of Numeric Series

Graphical user interface, text, application, email

Description automatically generated

Fig 2: Statistical Dataset Analysis of Object Series

Text

Description automatically generated

Fig 3: Checking for the missing or null values

Text

Description automatically generated

Fig 4: Verifying if all the Missing values were converted and replaced

A screenshot of a computer

Description automatically generated with low confidence

Fig 5: Labels assigned for all the factors of Object Variables

Chart

Description automatically generated with medium confidence

Fig 6: Working Class vs Average Income

Chart, histogram

Description automatically generated

Fig 7: Education vs Average Income

Chart, bar chart

Description automatically generated

Fig 8: Marital Status vs Average Income

Chart, bar chart

Description automatically generated

Fig 9: Occupation vs Average Income

Chart

Description automatically generated

Fig 10: Relationship vs Average Income

Chart, bar chart

Description automatically generated

Fig 11: Race vs Average Income

Icon

Description automatically generated with medium confidence

Fig 12: Gender vs Average Income

Chart, histogram

Description automatically generated

Fig 13: Hours Worked Per Week vs Average Income

Chart, scatter chart

Description automatically generated

Fig 14: Distribution of two Labels of Income earned (above 50K or below 50K)

NOTE: Supporting Jupyter Notebook file is attached along with the report.