***Prediction of Salary Class of an Individual Using Logistic Regression***

Machine Learning Deep Learning

Ruoming Jin

Vineeth Billakanti

811192961

Masters in Computer Science

Kent State University

[vbillaka@kent.edu](mailto:vbillaka@kent.edu)

**Abstract:**

Logistic regression is a statistical analysis approach that uses prior data set observations to predict a binary result, such as yes or no. By analyzing the relationship between one or more existing independent variables, a logistic regression model can predict a dependent data variable.

Problem:

The problem is to predict the salaried class of an individual is greater than $50,000 or less than $50,000 based on an individual’s credentials like education level, age, gender, experience, occupation, etc. For example, the salary of an individual whose experience is above 15 years is most likely to be greater than $50,000. Prediction is not made by considering only one factor but all the factors that affect the income of an individual. I am going to explain how the problem is predicted by using Logistic Regression method.

**Introduction:**

A prediction is a guess about what will happen in the future. A forecast is based on information or experience in certain cases, but not always. Future occurrences are not always guaranteed, therefore specific data about the future is often hard to confirm. However, a forecast may be valuable in developing strategies for likely developments. In this study, the pay of an organization's employee is forecasted using prior compensation growth rates. The pay history of a person has been watched, and it may then be determined automatically based on that wage after a given length of time.

Now, the Logistic regression is a statistical analysis approach that uses prior data set observations to predict a binary result, such as yes or no. By analyzing the relationship between one or more existing independent variables, a logistic regression model can predict a dependent data variable.

**Methodology:**

Machine Learning (ML) is a branch of computer science that allows computers to understand data in a similar manner that humans do. To put it another way, machine learning (ML) is a sort of artificial intelligence that uses an algorithm or technique to extract patterns from raw data. The goal of machine learning is to enable computers to learn from their own experiences without the need for explicit programming or human involvement.

Logistic regression is a classification model rather than a regression model. For binary and linear classification issues, logistic regression is a simple and more efficient technique. It is a classification model that is simple to implement and delivers excellent results with linearly separable classes. It is a widely used categorization method in the business. Like the Adaline and perceptron, the logistic regression model is a statistical approach for binary classification that may be adapted to multiclass classification.

About Dataset:

The dataset is taken from Kaggle. The US Adult Census dataset is a repository of 32,561 entries with 15 variables. The dataset contains information about age, work class, education, occupation, relationship, country, income. Let’s look at dataset details.



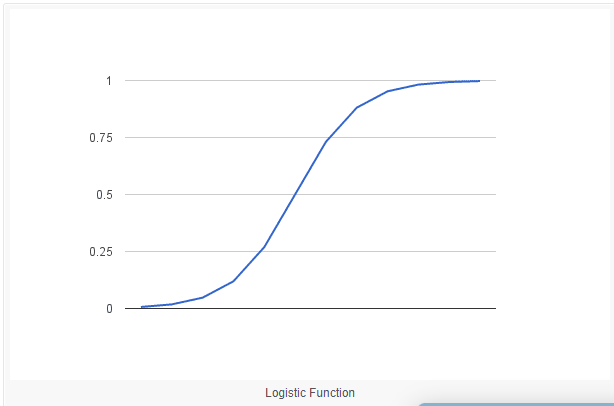
Here is the link of Dataset: [Adult Census Income | Kaggle](https://www.kaggle.com/uciml/adult-census-income)

Approach:

We may utilize the Naive Bayes approach, Linear Regression, and Logistic Regression to solve this prediction issue. Because the prediction variable (salary class) is dependent on a number of factors (both category and numerical), the Naive Bayes approach and Logistic Regression are superior than Linear Regression for this issue. Because our dataset contains more categories than numerical factors, Logistic Regression is the ideal approach for this issue.

The Supervised Learning approach includes one of the most prominent Machine Learning algorithms: logistic regression. It's used to predict a categorical dependent variable from a group of independent factors. As a result, the result must be a discrete or categorical value. It may be Yes or No, 0 or 1, true or false, and so on, but instead of providing precise values like 0 and 1, it delivers probabilistic values that are somewhere between 0 and 1. Furthermore, unlike linear regression, logistic regression does not need a linear connection between the input and output variables. This is due to the odds ratio being transformed via a nonlinear log transformation.



****

**Experimentation:**

This prediction has been implemented through the following approaches:

* numpy:- numpy (‘Numerical Python’) is a Python library. It is a scientific computing core package that includes a strong n-dimensional array object as well as integration tools for C, C++, and other languages. It also has applications in linear algebra, random number generation, and other fields.
* matplotlib.pyplot:- matplotlib.pyplot is a set of command-style functions that allow matplotlib to behave similarly to MATLAB. Each pyplot function modifies a figure in some way, such as creating a figure, a plotting area in figure, Charting certain lines in a plotting area, decorating the plot with labels and so on.
* pandas:- For data processing and analysis, pandas is a software library created in Python. Its data structures and methods for manipulating numerical tables and time series are particularly useful. It’s open-source software licensed under BSD license, which has three clauses.
* read\_csv:- Python is a great language for doing data analysis, primarily because of the fantastic ecosystem of data-centric python packages. pandas is one of those packages and make importing and analyzing data much easier.
* Predict:-predict the values where find at the time period that how much there in that graph.

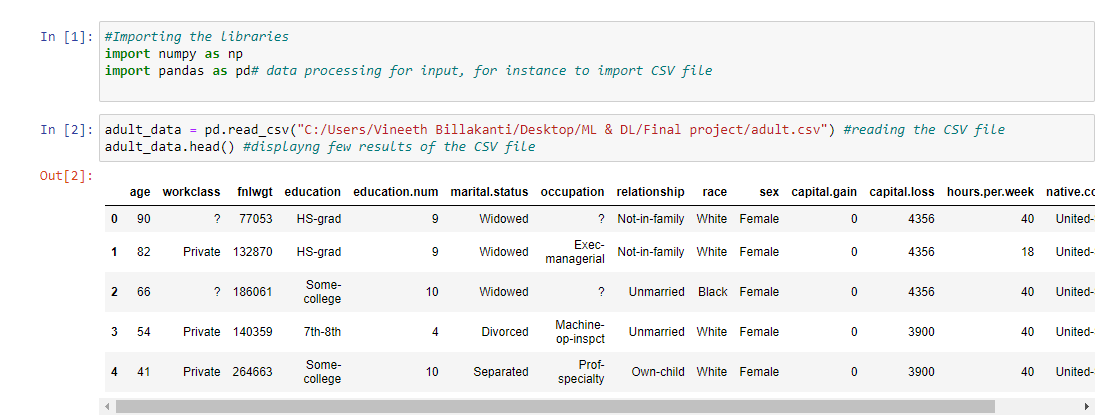
To begin, create an excel file dataset, which will then be opened in Jupyter notebook using ANACONDA Navigator. From there, we can use Anaconda Navigator to read the data set. First, in the Jupyter notebook, we’ll import two variables : numpy and pandas. These functions are used as follows:

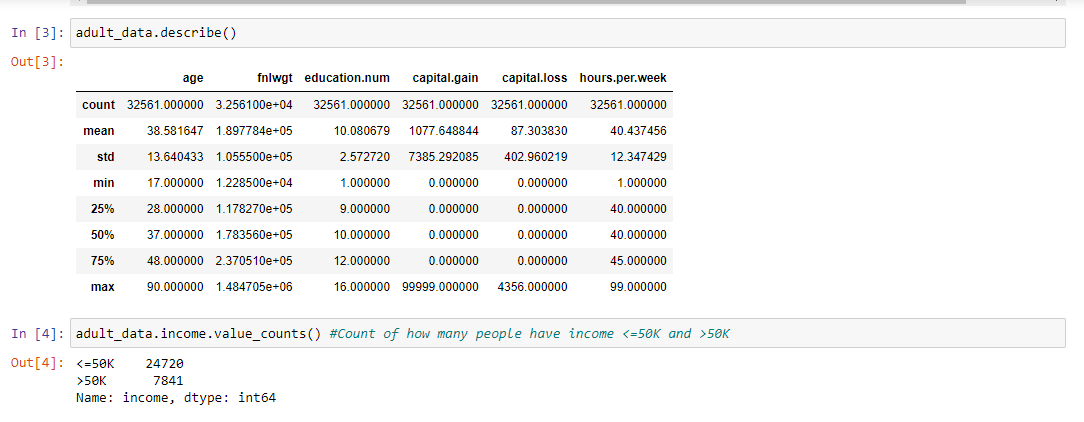
* NumPy is used for antialiasing a dynamic array or large set.
* matplotlib.pyplot is used for making graph
* pandas are mainly used like database where we store the dataset from the excel file. Through the data set we are forming graph and predicting.

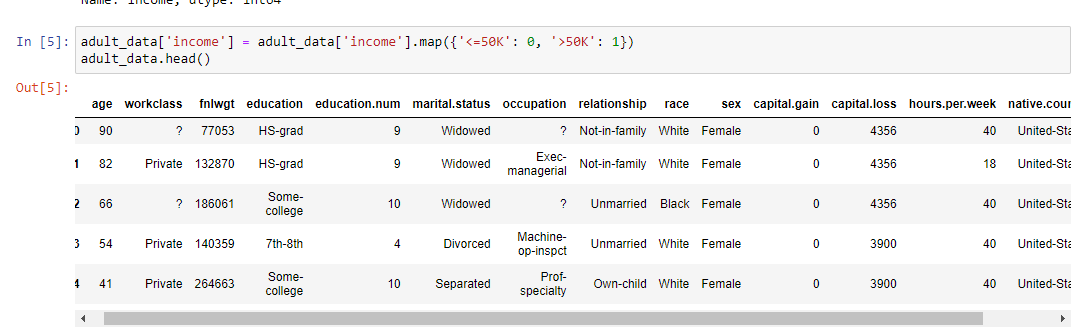
Then import the dataset on the pandas through the help of “<any variable>. read\_csv” and then showing that field that are importing from the excel file. Then the salary is predicted by using the logistic regression method.

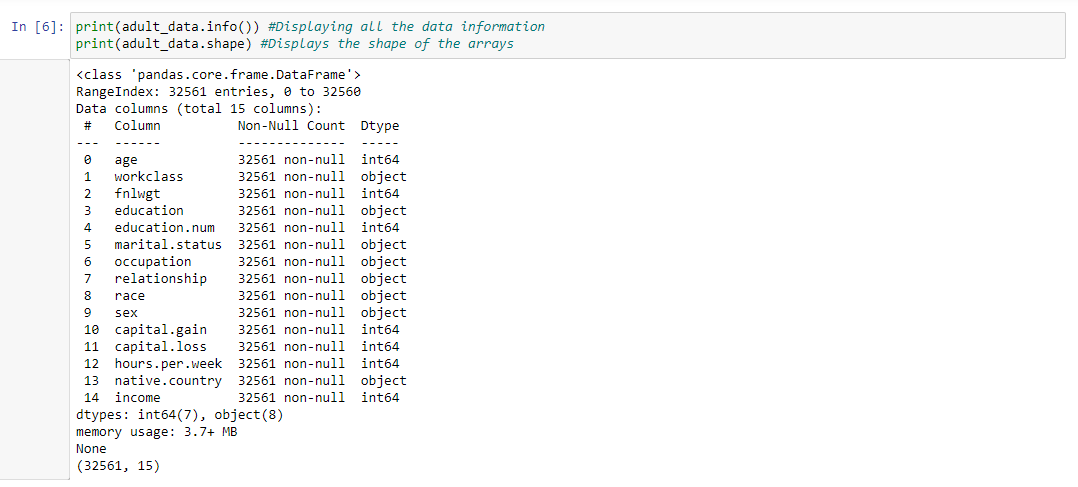
**Python Code and Output:**

Importing the Libraries and reading the .CSV file(Data Set):





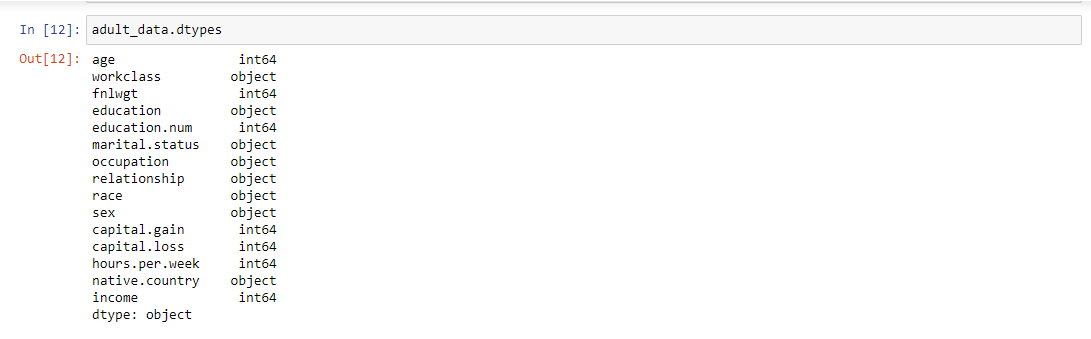




Data Cleaning:

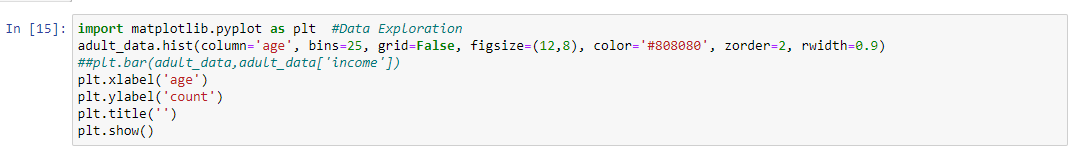


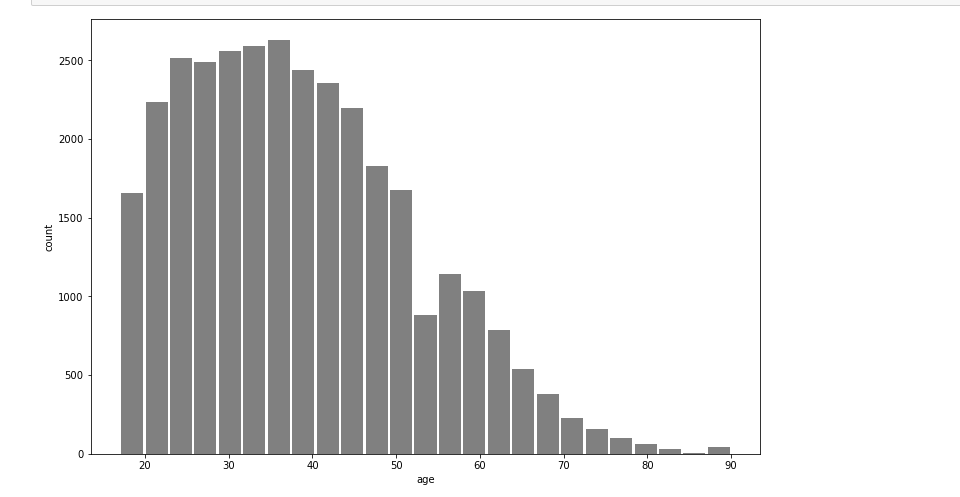


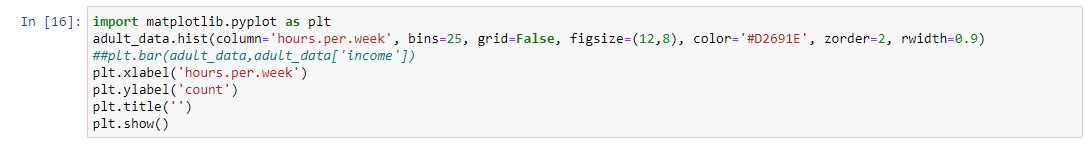


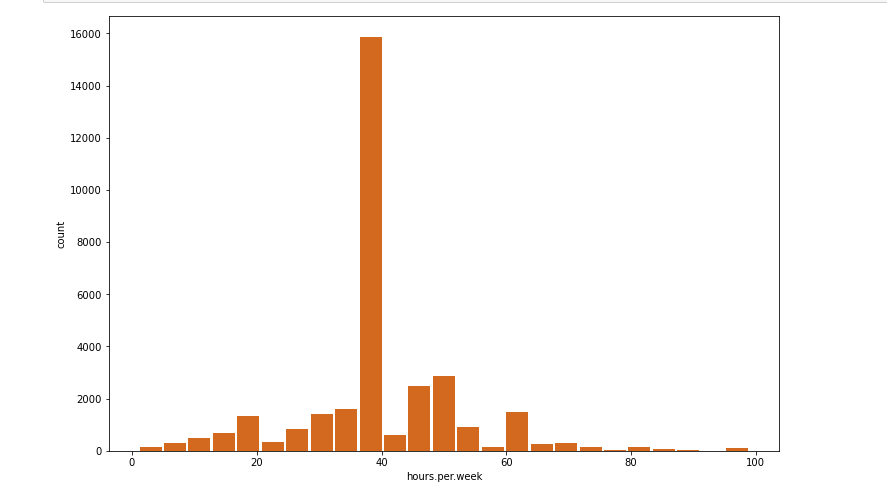


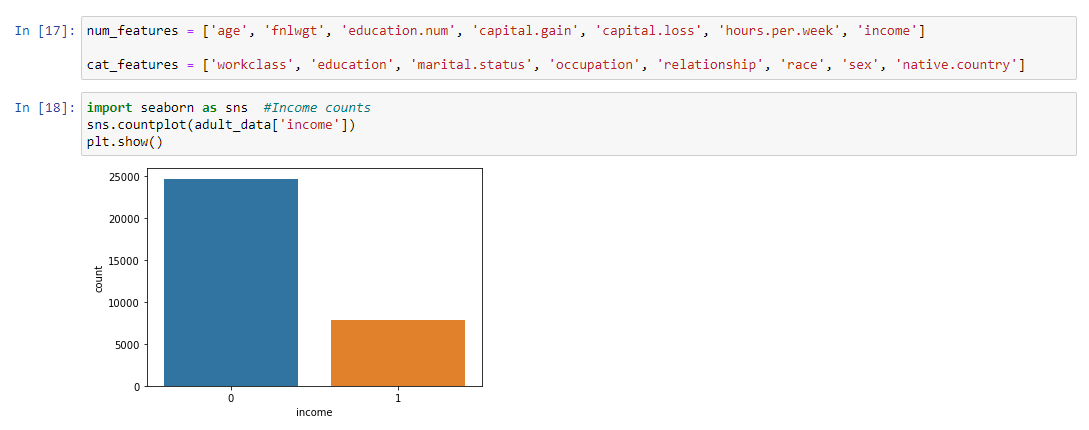
Data Exploring:

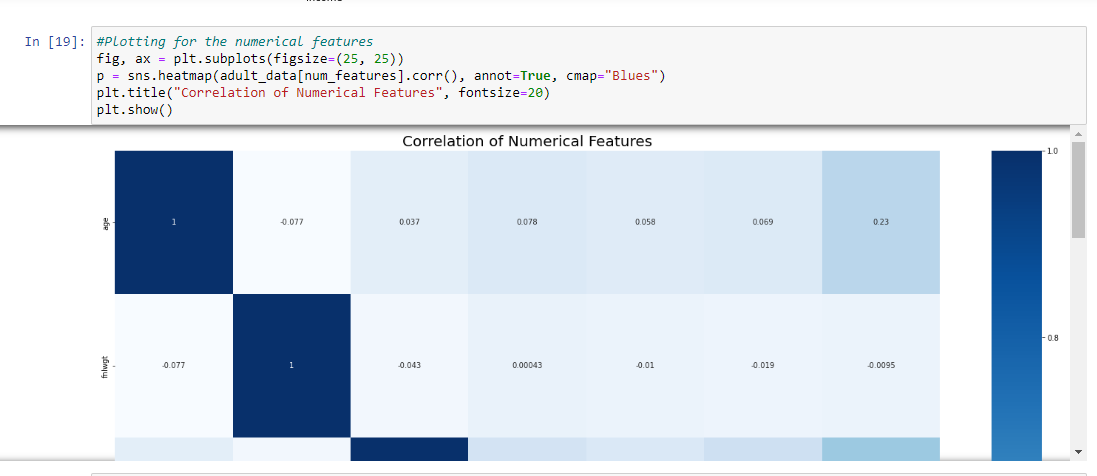


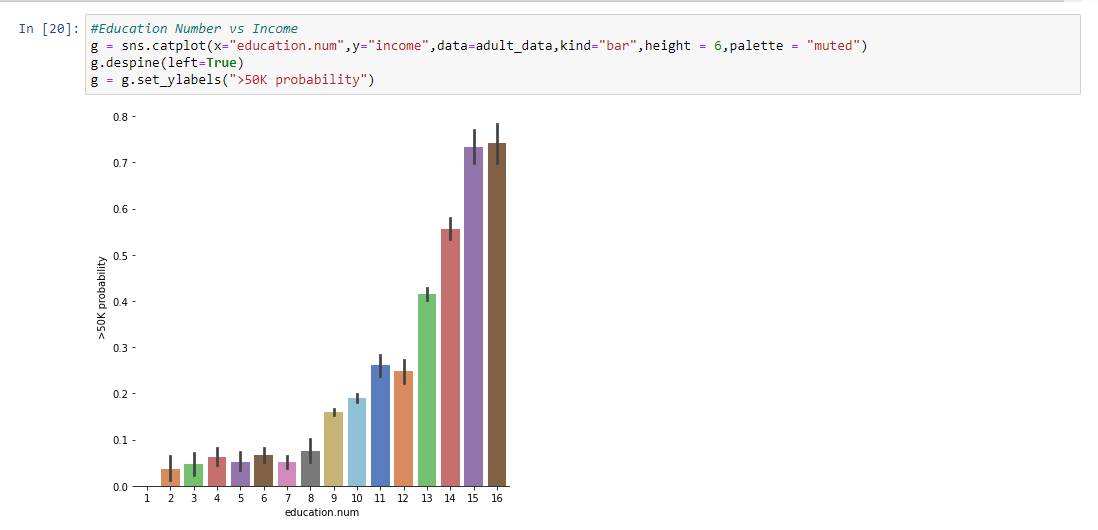






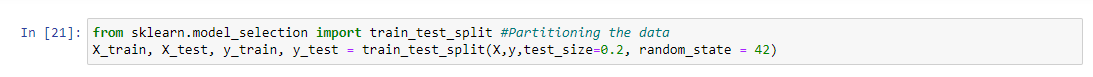




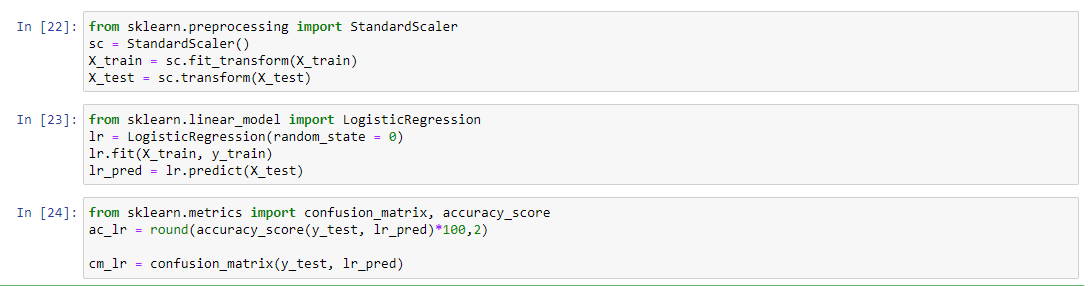


Data partitioning:

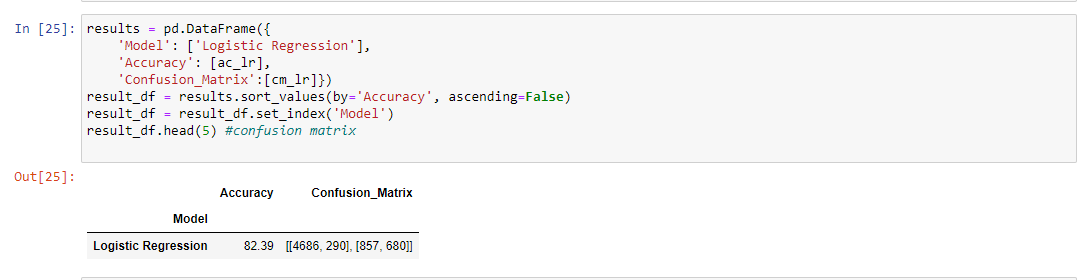
Here we split the data into 80% Training data and 20% testing data.



Model Building:



Confusion Matrix:



**Results:**

As this is a prediction problem Accuracy is the measure to find how our model is predicting the salaried class of an individual. The accuracy of the model is shown below.

Accuracy = (Correct predictions) / (Total predictions)

= (5627+722) / (5627+140+1190+722)

= 6349 / 7679

=0.8239

The accuracy of our model is 82.39% to predict the salary class of an individual by considering age, education, hours per week, occupation, and sex as categorical variables.

**Conclusions:**

From the results, we can conclude that the salary of an individual whose income is greater than $50,000 is male, whose education level is higher than a master’s degree and who works more than 40 hours per week, and whose occupation is in the private sector.

**Future Work:**

82.39 percent accuracy was obtained in the aforementioned result.

By including additional data samples in the data set and increasing the amount of training data, we can improve our accuracy.

**GitHub Link:**

https://github.com/vineeth3652/Machine-Learning-Deep-Learning.git

**References:**

<https://www.keboola.com/blog/logistic-regression-machine-learning>

<https://journalofbigdata.springeropen.com/articles/10.1186/s40537-022-00559-6>