**Course Work Project**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Semester** | **202420** | | **Division** | | CIS |  |
| **Assessment title in Syllabus** | **Project** | | **Program** | | **IT and IS** |  |
|  |  | |  | |  |  |
| **Course Code** | **CIS 2423** | | | | |  |
| **Course Title** | **Programming for Data Analytics** | | | | |  |
| **CLOs** | **All CLOs** | | **Accreditation Body** | | **CAA & CIPS** |  |
| **Course Instructor** |  | | **CRN** | |  |  |
| **Assessment Weight** | **40%** | | **Submission Date** | | **Week 14** |  |
| **For Group Work submissions an additional individual assessment will be conducted.**  **Grades for the students in one group will vary based on the individual performance in the additional assessment.** | | | | | |  |
|  | | | | | |  |
| **Student Declaration**:  **Academic Integrity Statement**  In accordance with the HCT Academic Integrity Policy  • Students are required to refrain from all forms of academic integrity breaches as defined and explained by HCT.  • A student found guilty of having committed acts of academic integrity breach(es) will be subject to the relevant sanctions as outlined by HCT.  إفادة النزاهة الأكاديمية  **وفقًا لسياسة كليات التقنية العليا للنزاهة الأكاديمية**  **• على الطلبة الإلتزام بلوائح وقواعد النزاهة الأكاديمية، كما هو مبيّن وموضح في السياسات والإجراءات الخاصة بكليات التقنية العليا.**  **• في حالة ارتكاب الطالب أي شكل من أشكال الإخلال بالنزاهة الأكاديمية، سيتعرض الى العقوبات الموضحة في السياسات ذات الصلة.**  This assignment is entirely my own work except where I have duly acknowledged other sources in the text and listed those sources at the end of the assignment. I have not previously submitted this work to the HCT, or any other entity. I understand that I may be orally examined on my submission.  **Student (s) Signature: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**  **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** | | | | | |  |
|  | | | | | |  |
| **Student Name(s):** |  |  | |  | |  |
| **Student HCT ID(s):** | H00 | H00 | | H00 | |  |

**For Examiner’s Use Only**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Group (50%)** | | | | | **Individual (50%)** |  |  |
| **C**LO | **1** | **2** | **3** | **4** | **Report Formatting** | **Oral Defense** | **Total** | **%** |
| **Marks Allocated** | 10 | 10 | 42 | 26 | 12 | **50** | **100** | **4**0 |
| **Marks Obtained** |  |  |  |  |  |  |  |  |

|  |
| --- |
|  |

# Section I

The dataset we chose is the Global AI Content Impact Dataset. The dataset contains AI-related information and includes the columns; Country, Year, Industry, AI Adoption Rate (%), AI-Generated Content Volume (TBs per year), Job Loss Due to AI (%), Revenue Increase Due to AI (%), Human-AI Collaboration Rate (%), Top AI Tools Used, Regulation Status, Consumer Trust in AI (%), Market Share of AI Companies (%).

The purpose of data analysis for this dataset is to uncover patterns and relationships between AI adoption and job loss, helping to understand the broader impact of AI technologies on employment.

The type of programming we will use for data analysis is Python, since it boasts a vast collection of libraries specifically designed for data analysis, such as Pandas, NumPy, Matplotlib, and others. Python is versatile and can be used for data preprocessing, model building, visualization, and advanced analysis tasks.

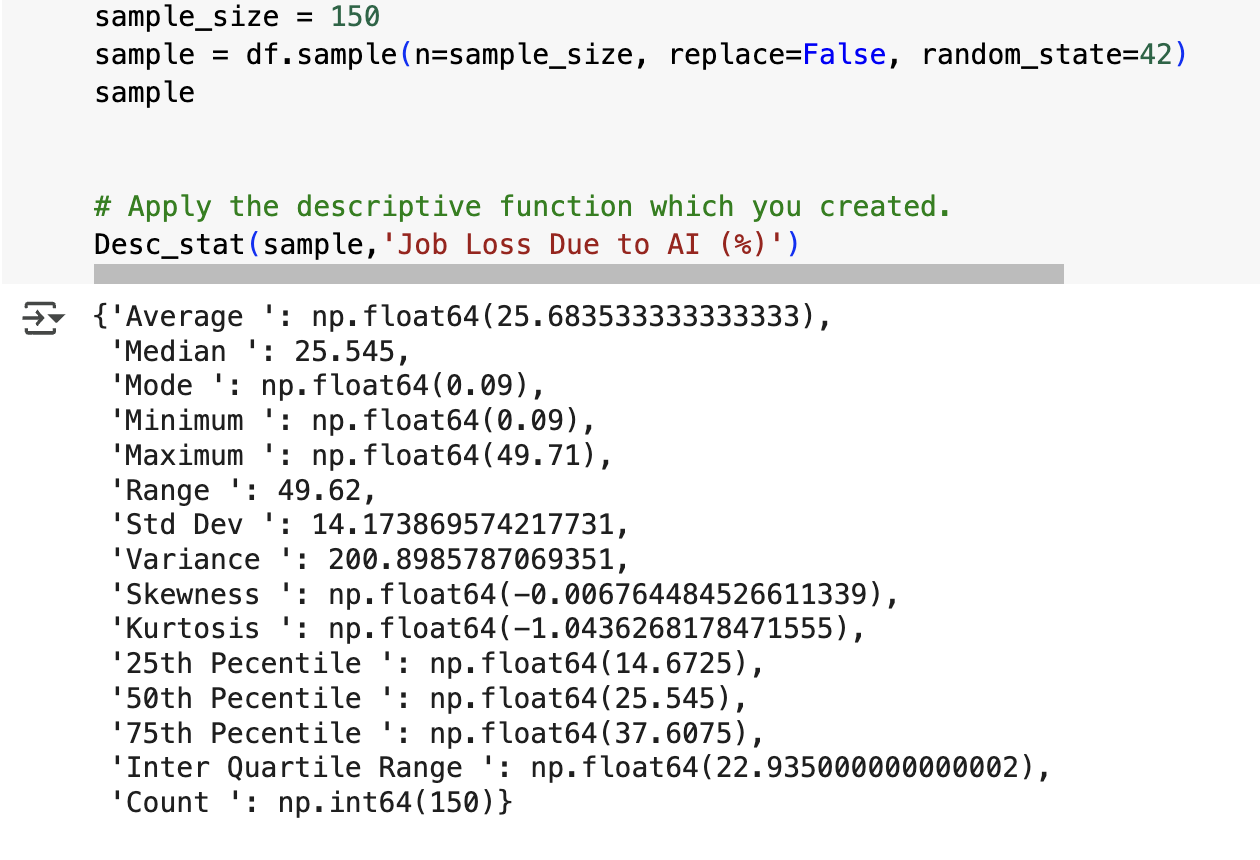
The machine learning algorithm type we plan to use is regression because our goal is to analyze and predict how changes in AI Adoption Rate (%) impact Job Loss Due to AI (%). The purpose of implementing regression is to model and quantify the relationship between these two variables to make future predictions or informed decisions.

Our research question is “How does AI Adoption Rate (%) influence Job Loss Due to AI (%) across industries and countries?” We chose AI Adoption Rate (%) as the independent variable because it shows how much AI is being used, and Job Loss Due to AI (%) as the dependent variable because it reflects the impact on employment. This makes sense, as higher AI adoption rate is expected to lead to job loss. The dataset supports this relationship across different countries and industries. Later in the report we used more than one independent variable because we had to for Multiple Linear Regression since it requires multiple input variables. So, in addition to AI Adoption Rate (%), we included Human-AI Collaboration Rate (%) and Consumer Trust in AI (%). We chose these variables because they directly relate to how AI is applied and received in the workplace. Human-AI Collaboration shows how many people are working alongside AI, which is a factor in Job Loss Due to AI (%). Consumer Trust in AI reflects how accepted and integrated AI is within society, which can also influence job loss.

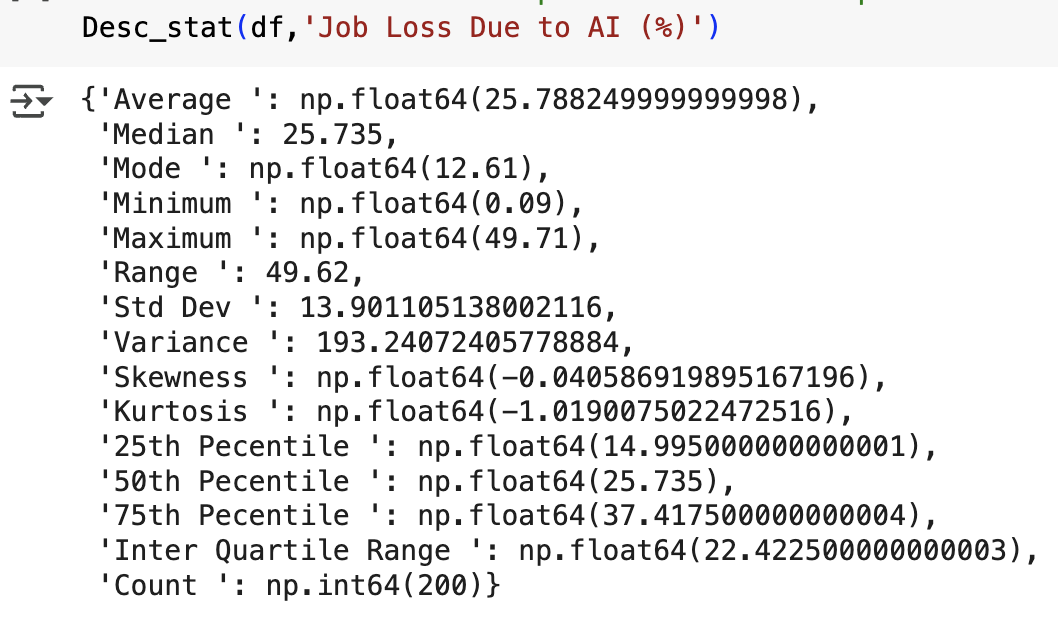
# Section II

Descriptive analysis is an essential step in understanding our dataset's structure and behavior. It makes it simpler to spot patterns, trends, and inconsistencies by helping to summarize and simplify data in a way that can be understood and interpreted visually. Understanding the distribution of AI Adoption Rate (%) and Job Loss Due to AI (%) can be achieved by examining the types of descriptive statistics like the mean, median, standard deviation, and range. As a result, descriptive analysis allows us to better understand the data, helps identify errors and outliers before proceeding further with analysis, and supports accurate interpretation in our regression model.

To begin with descriptive statistics with our dataset we created and developed a python function. From this function we found the descriptive statistics of our dependent variable ‘Job Loss Due to AI’ based on a random sample of 150 observations to which we added to our descriptive statistics function ‘Desc\_Stat’. Output:



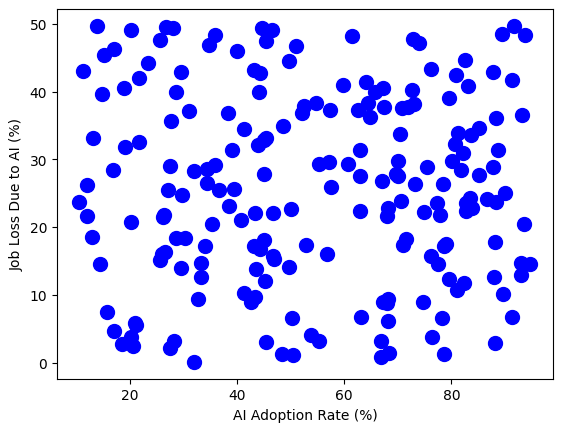
From the function ‘Desc\_stat’ we can understand our dependent variable ‘Job Loss Due to AI’ more by examining the types of descriptive statistics such as the mean, median, standard deviation, and range. Output:



From this output we know the mean is 25.79, which tells us the average percentage of Job Loss Due to AI. The median is 25.73 and is very close to the mean, saying that the data is symmetrical. The mode is 12.61, meaning it's the most frequent percentage of job loss. The standard deviation 13.90 tells us how much the data values for Job Loss Due to AI differ from the mean, since 13.90 is considered a low standard deviation it shows that the values are closer to the mean. The minimum and maximum values are 0.09 and 49.71, giving us a wide range of 49.62 and shows that Job Loss Due to AI varies a lot in the dataset. The skewness is slightly negative -0.04, meaning the distribution is nearly symmetrical with a small lean to the left. The interquartile range is 22.42 which means that the middle 50% of the Job Loss Due to AI values fall within 22.42%.

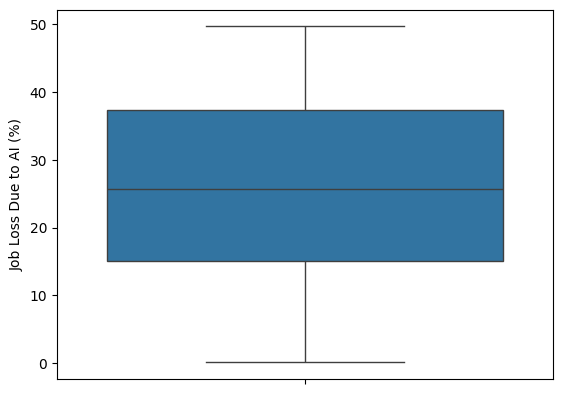
To visualize our dependent variable ‘Job Loss Due to AI’, we performed exploratory data analysis using various plots, including a scatter plot, box plot, histogram, and heatmap. Each visualization provides unique insights: the scatter plot shows potential relationships with independent variables, the box plot highlights the presence of outliers and the spread of data, the histogram reveals the distribution and skewness, and the heatmap displays correlations between variables.

#### Scatter Plot:



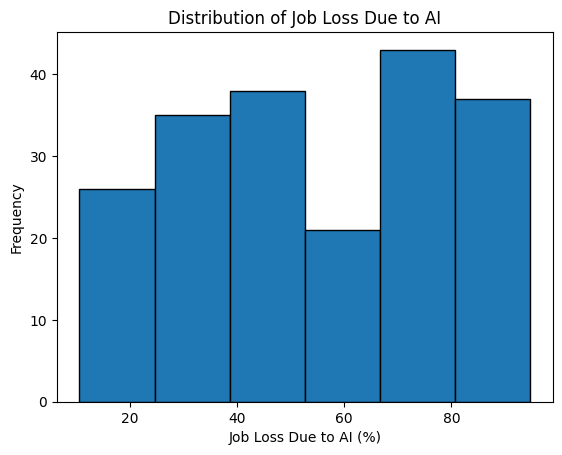
This scatter plot shows the relationship between x (AI Adoption Rate) and y (Job Loss Due to AI) variables. Each point represents one observation from the dataset. From the plot we can see that the data points are widely scattered with no clear upward or downward trend. This suggests that there is no strong linear correlation between AI adoption and job loss, suggesting that higher AI adoption Rate does not consistently lead to higher or lower job loss.

#### Box Plot:



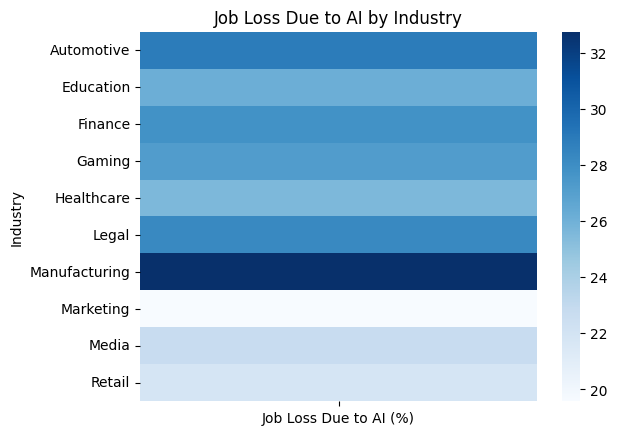
This box plot visualizes the distribution of the dependent variable Job Loss Due to AI (%). It helps give us a view of how job loss data is distributed and helps identify any potential outliers or asymmetry in the data. There are no extreme outliers in this plot. The box plot looks symmetrical and shows that the job loss values are fairly evenly distributed around the median.

#### Histogram:



This histogram shows the distribution of our dependent variable Job Loss Due to AI (%). The bars represent how often job loss values fall within certain percentage ranges. The histogram shows that Job Loss Due to AI (%) is spread across different levels, but most values fall between 40% and 90%. The highest frequency is in the 70–80% range, showing that many industries are experiencing high levels of Job Loss Due to AI.

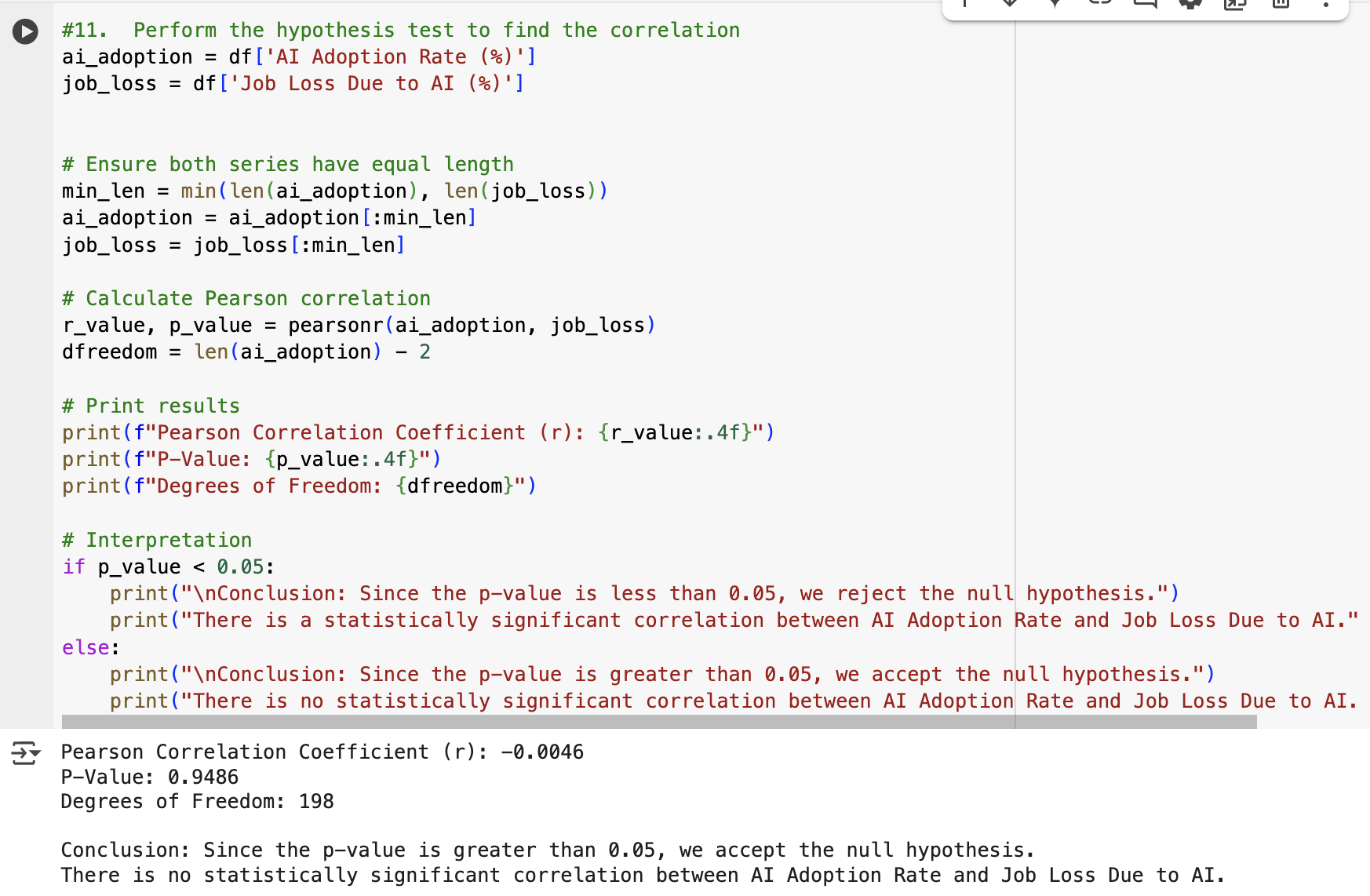
#### Heat Map:

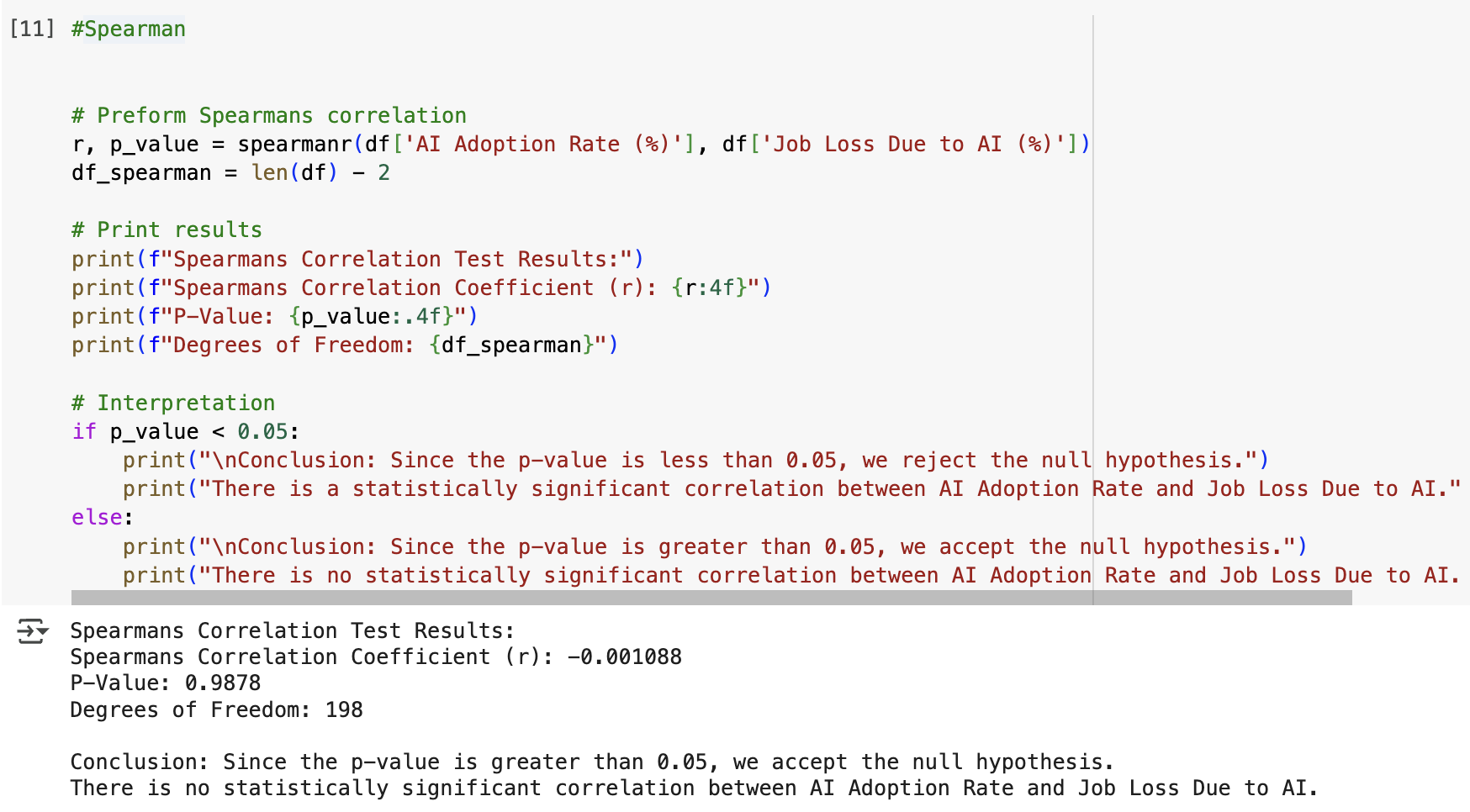


For this heatmap we needed to use a categorical variable to compare with our dependent variable, so we chose Industry as the independent variable. This heatmap shows the average Job Loss Due to AI (%) across different industries using color intensity. Darker shades represent higher percentages of job loss. From this heatmap, we can see that industries like Manufacturing and Automotive industry are experiencing higher levels of job loss, while industries like Marketing, Media, and Retail show lower percentages. This helps us quickly identify which industries are most affected by AI in terms of job loss.

#### Hypothesis Testing

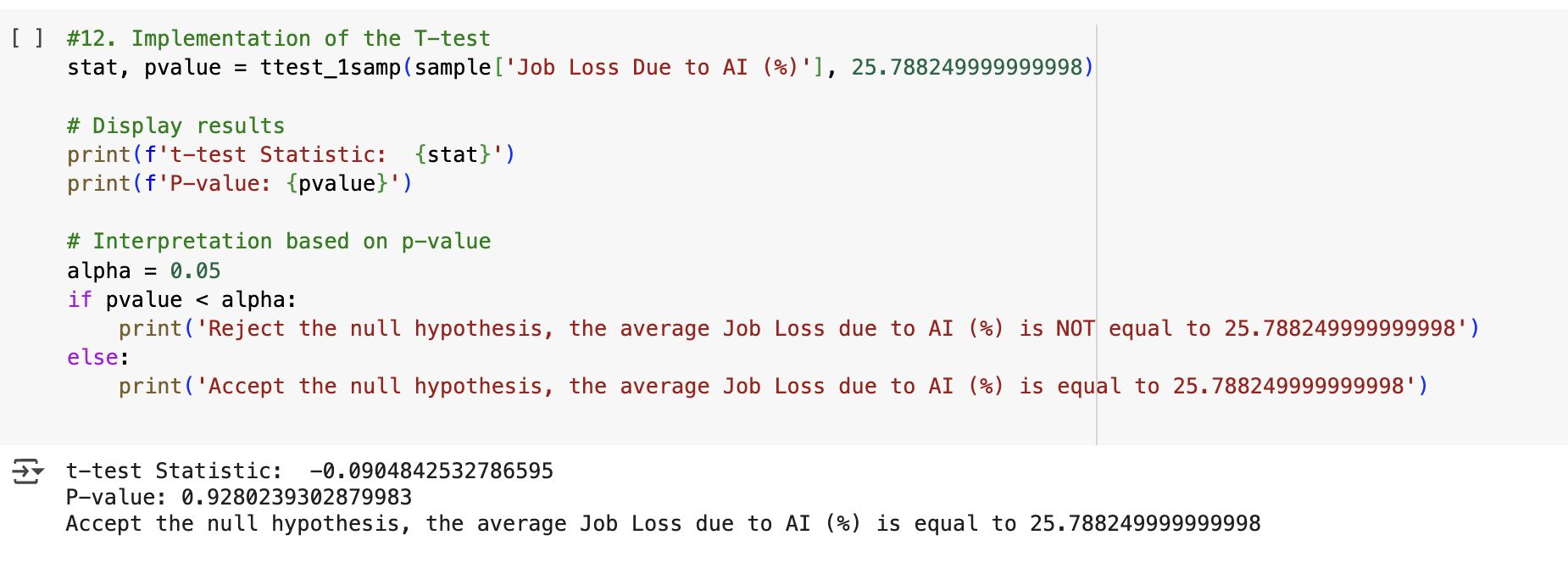
To find the correlation between our independent (AI Adoption Rate Job) and dependent (Job Loss Due to AI) variable. To test the correlation, we implemented the Pearson and Spearman tests since our variables are continuous. The null hypothesis is, there is no correlation between Job Loss Due to AI (%) and AI Adoption Rate (%). And the alternative hypothesis is there is a correlation between Job Loss Due to AI (%) and AI Adoption Rate (%) Output of the tests:





The findings show that there is no significant correlation between the independent variable (AI Adoption Rate) and the dependent variable (Job Loss Due to AI). The results from both the Pearson and Spearman correlation tests indicate that there is no statistically significant relationship between AI Adoption Rate (%) and Job Loss Due to AI (%). In the Pearson test the p-value is0.9486, and in the Spearman test the p-value is0.9878. Since both p-values are much greater than the 0.05 significance level, we accepted the null hypothesis in both cases. This means that based on our dataset there is no correlation between how much AI is adopted and the level of job loss it causes.

To assess the performance of our dependent variable we performed a one-sample t-test to compare the average Job Loss Due to AI (%) in our random sample of 150 records to the overall population mean (25.79%). Output of t test:



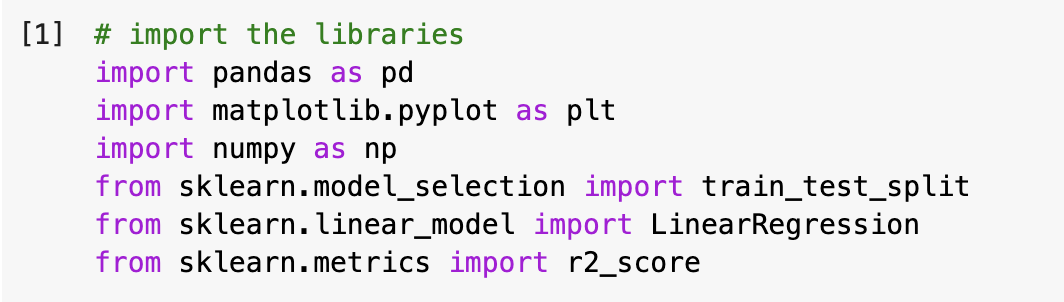
The p-value is shown to be 0.928, which is much greater than the significance level of 0.05. This means we accept the null hypothesis and conclude that there is no statistically significant difference between the sample mean and the population mean.

# Section III

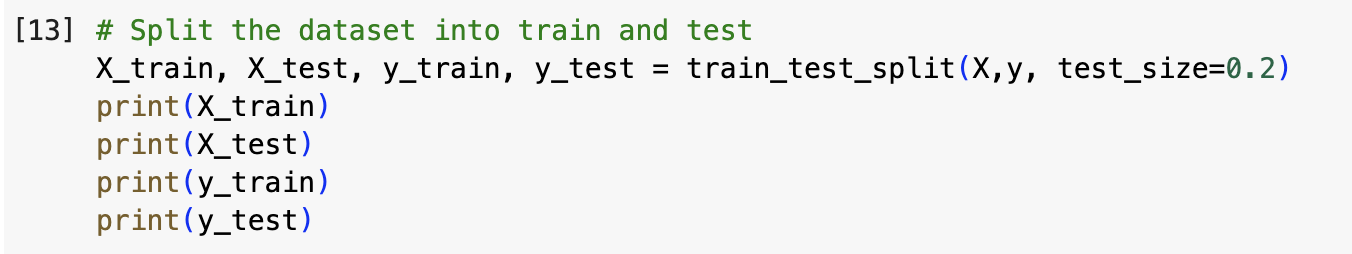
|  |
| --- |
| 1. Build, Train, Develop and Evaluate using Simple Regression for chosen dataset. |
| 1. Develop a script to forecast the value of the dependent variable from all the relevant independent variables using Multiple Linear Regression |
| 1. Predict the value of the dependent variable from the different classifier such as Logistic Regression, KNN, Naïve-Bayes and Decision Tree. |
| 1. Evaluate the performance of each model using confusion matrix and accuracy and identify the best fit classifier for the chosen dataset. |
| 1. Predict the dependent variable by using best-fit classifier. |
| 1. Perform the cluster analysis such as K-means and Horizontal for any field from the chosen dataset. |
| 1. Explain the strategy for improving the system after viewing the cluster diagram. |

##### Developing a Script Using Multiple Linear Regression

Multiple Linear regression analysis is used to predict the values of a dependent variable based on the values of more than one independent variable. To begin the process, we selected the independent (AI Adoption Rate (%), Consumer Trust in AI (%), and Human-AI Collaboration Rate (%)) and dependent (Job Loss Due to AI (%)) variables for our model. Then we imported essential Python libraries.



We divided the dataset into a training set and a test set (80% training, 20% testing).

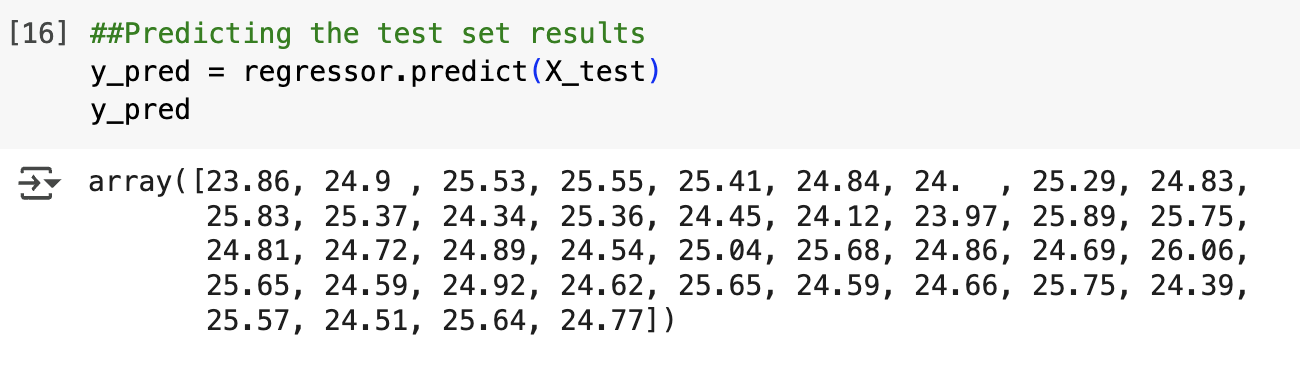


We successfully created a LinearRegression object and trained the model using the .fit() method with our training data. This allowed the model to learn the relationships between the independent variables and the dependent variable.

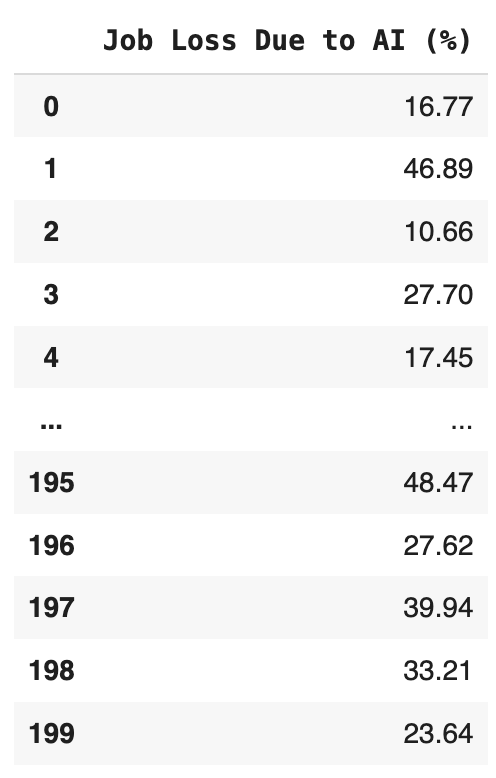
## 

After training the model, we used the .predict() method to generate job loss predictions for the test set. These predicted values were then compared with the actual values to visually assess how closely the model captured the trends in the data. The comparison shows that the model gives accurate predictions since many of the job loss values align with the actual results.

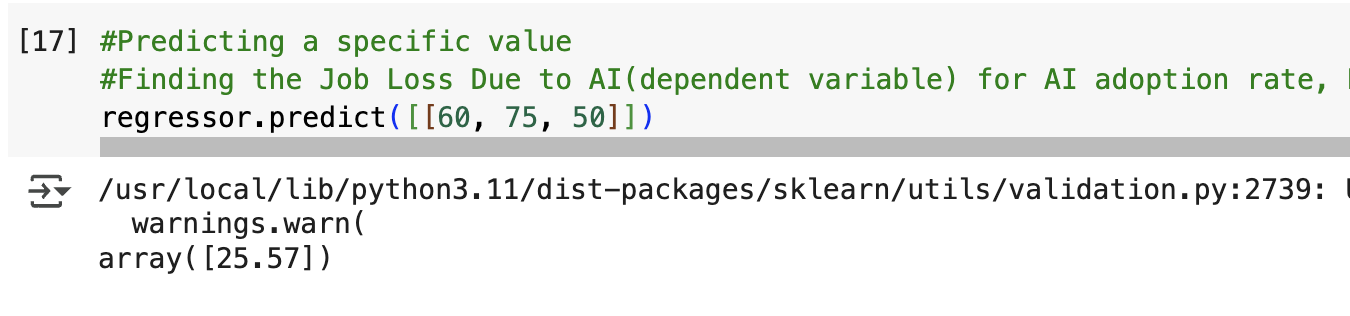
Predicted values:



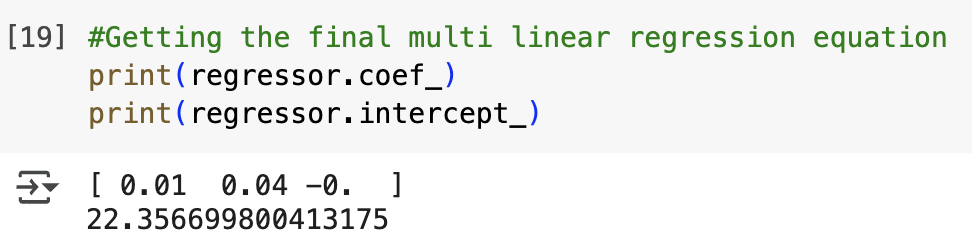
Actual Values:



To test the model we provided custom values for the independent variables, AI Adoption Rate = 60, Consumer Trust = 75, and Human-AI Collaboration = 50 and predicted the corresponding Job Loss Due to AI (%). The model returned a prediction of 25.57%.



Finally, we extracted the regression coefficients and the intercept for the final equation. The linear equation is Job Loss = 22.356699800413175 + 0.01\*(AI Adoption Rate) + 0.04\*(Consumer Trust in AI) - 0.00\*(Human-AI Collaboration Rate). Output:



This equation can now be used to estimate job loss due to AI.

##### Classification Models to Predict Job Loss Due to AI (%)

We implemented and compared four classification algorithms to predict our dependent variable (Job Loss Due to AI); Logistic Regression, KNN, Naïve-Bayes, and Decision Tree. Before applying these models, we converted the continuous values in our dependent variable Job Loss Due to AI into categorical classes (Low and High) for all models to make them suitable for classification tasks as in these algorithms the dependent variable must be categorical.

Converting job loss into categorical classes, we followed these steps and implemented them in all four classification models:

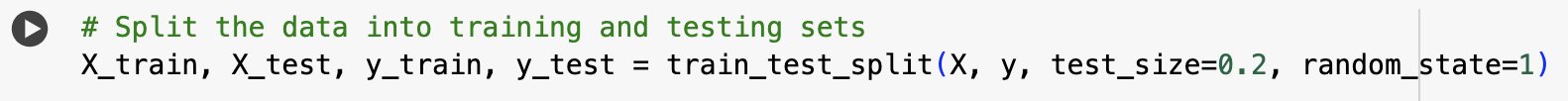


Since these algorithms require the dependent variable to be numeric, we applied the .map() function to convert the categorical labels (‘Low’, ’High’) into numeric values (0, 1), allowing the models to work with the data correctly.

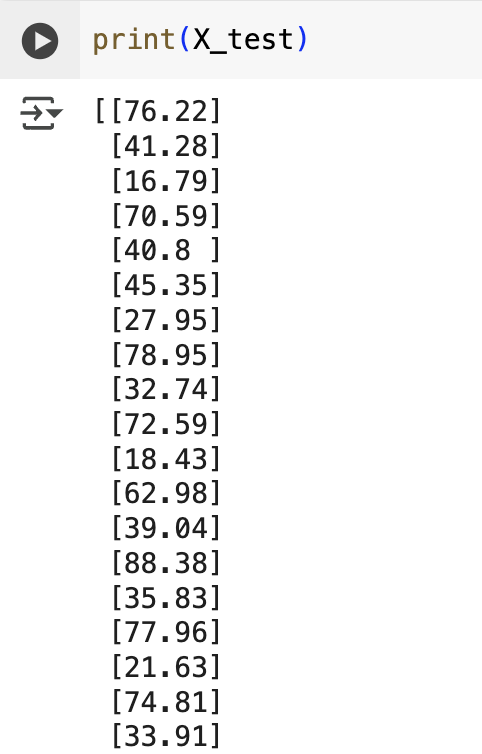
#### Classification – Logistic Regression

Logistic regression predicts a categorical dependent variable from independent variables. To implement this, we had to convert our numerical dependent variable into a categorical one by converting job loss continuous values into (Low and High), then to numeric labels as seen above.

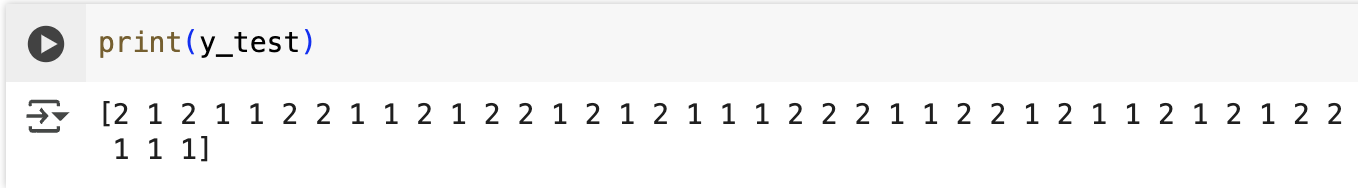
We then split the data into training and testing sets.

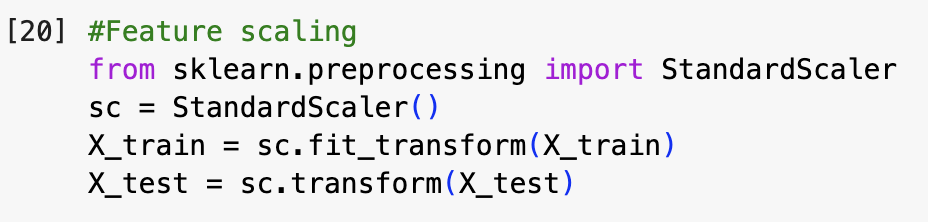


X:

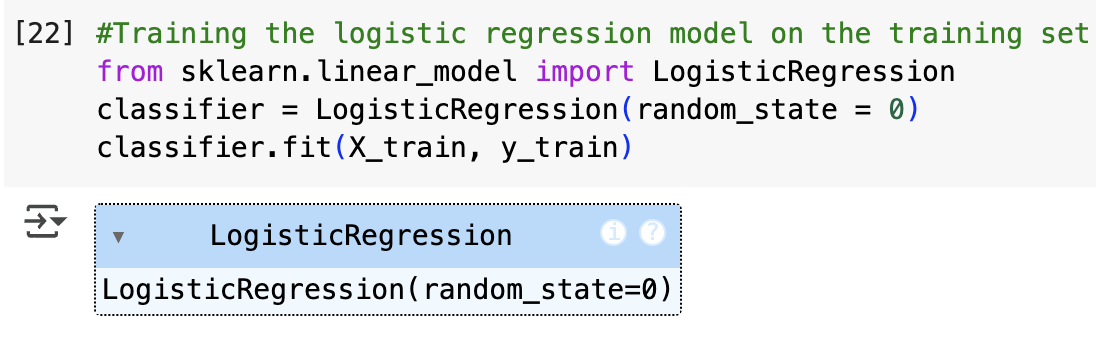


Y:

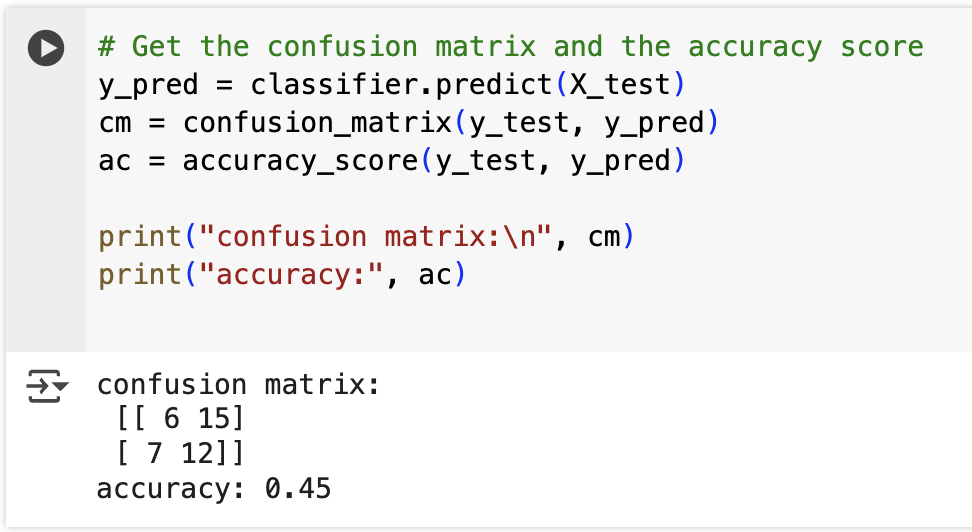




We then imported the LogisticRegression class from sklearn.linear\_model. Then, we created a model instance named classifier using LogisticRegression(random\_state=0). We attatched the classifier .fit() method to train the model on our training data (X\_train and y\_train), allowing the algorithm to learn the relationship between AI Adoption Rate (%) and Job Loss categories.



To evaluate the model’s performance, we used the confusion\_matrix and accuracy\_score functions from sklearn.metrics.

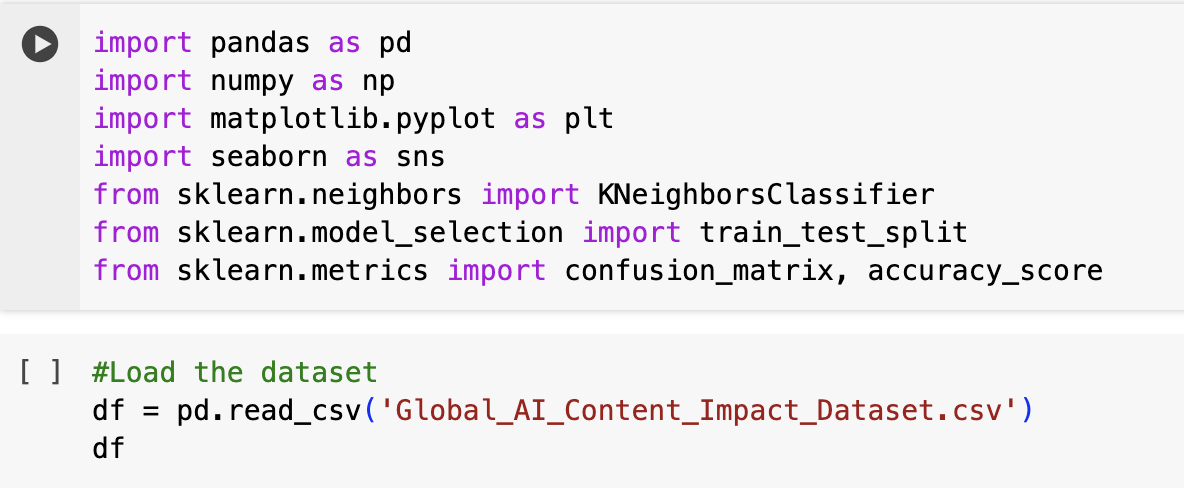


As seen from the output 45% of the predictions were correct.

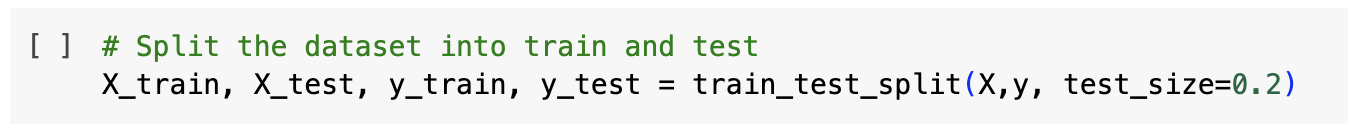
#### Classification – K-Nearest Neighbor (KNN)

K-nearest neighbors (KNN) algorithm is used for both classification and regression. It works by finding the k most similar points to a new data point.

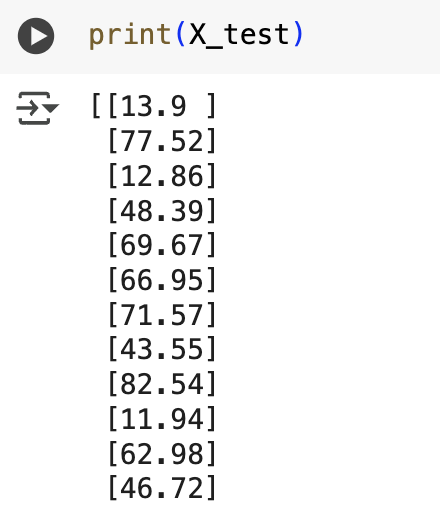
First, we imported the needed libraries and uploaded the dataset.



And we selected independent and dependent variables with the same steps as the Logistic Regression model, after that we split the dataset.



X:



Y:



Then we trained and fitted the needed model for KNN.

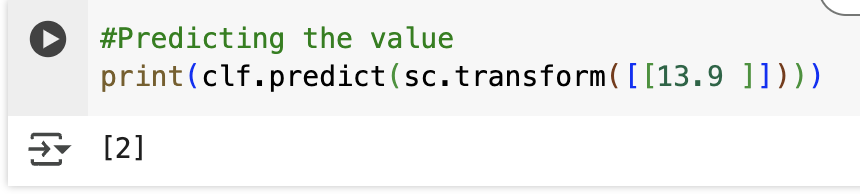
#### 

We then got the accuracy score and confusion matrix to evaluate the performance of our KNN classification model.

#### 

As seen from the output 55% of the predictions were correct.

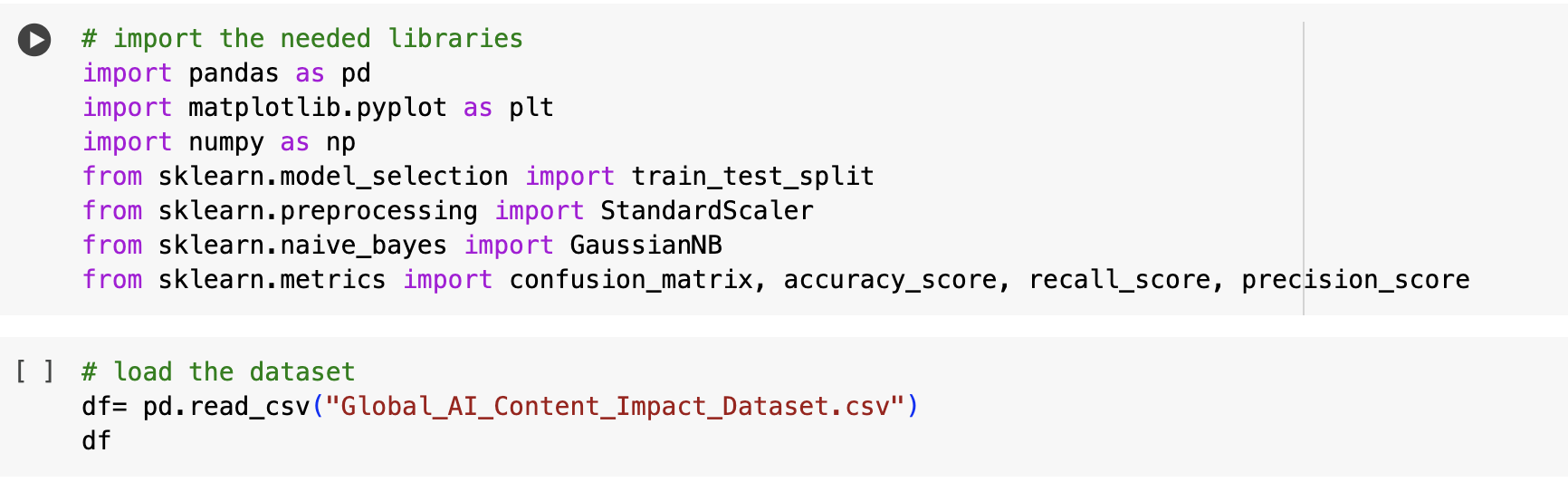
Predicting the value:



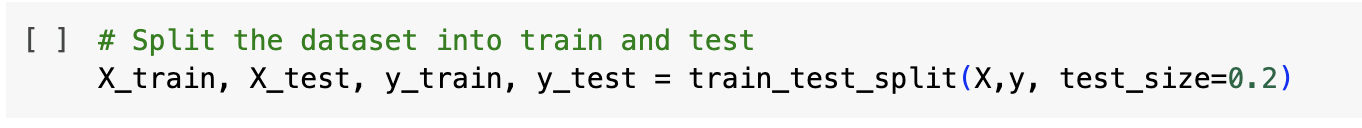
It predicted correctly that class=2 same as the y\_test output. So, the model is working fine.

#### Classification – Naïve Bayes

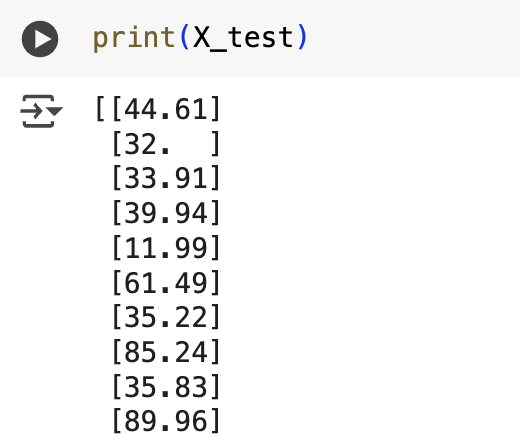
Bayes' theorem describes how to update our beliefs about an event (hypothesis) when we receive new information (evidence).



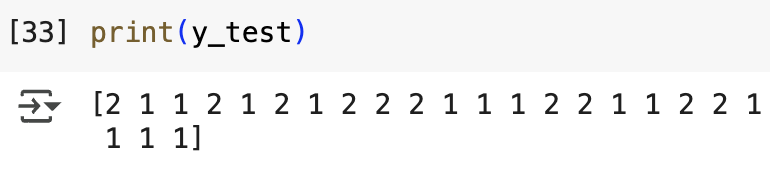
And we selected independent and dependent variables with the same steps as the Logistic Regression model, after that we split the dataset.



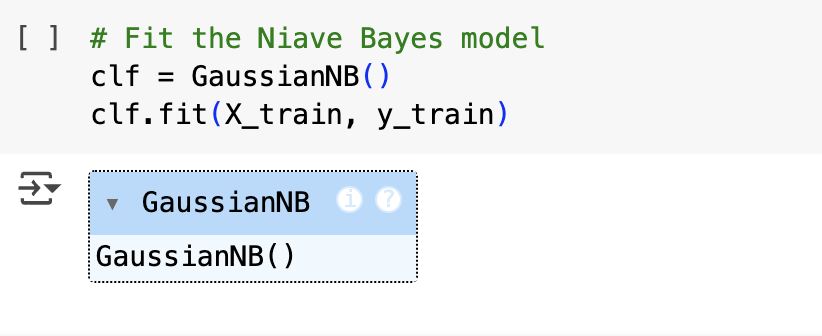
X:



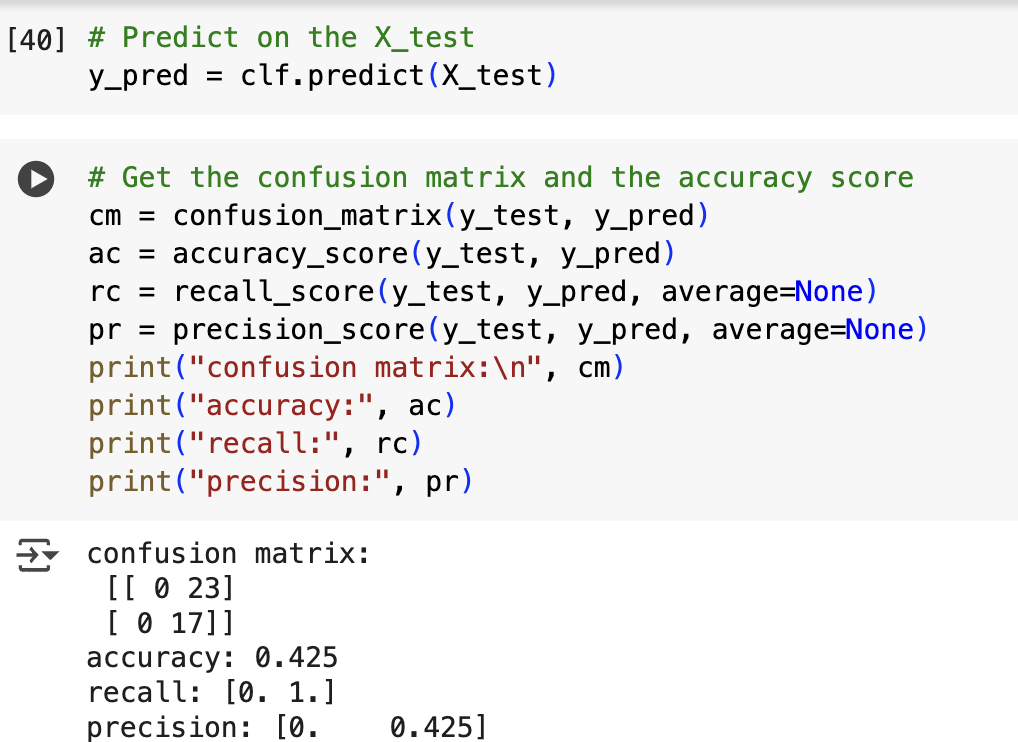
Y:



Then configured for Naïve Bayes model by mentioning the classifier is Naïve Bayes and attached the classifier to the train dataset.

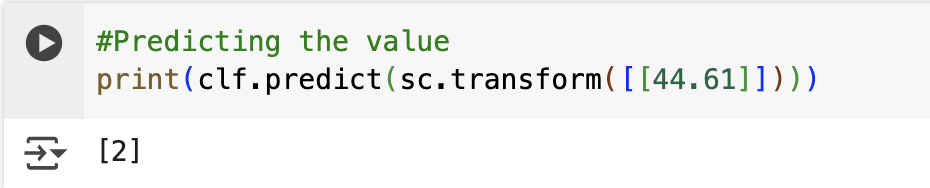


We then got the accuracy score and confusion matrix to evaluate the performance of our Naïve Bayes classification model.



As seen from the output, 42.5% of the predictions were correct.

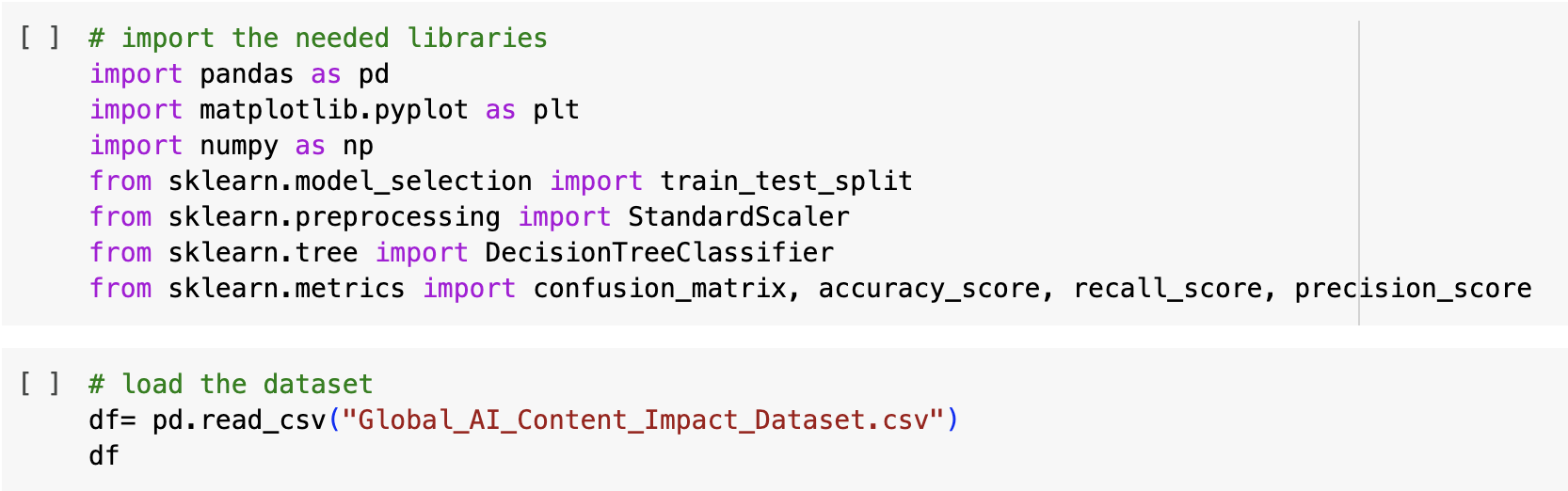
Predicting the value:

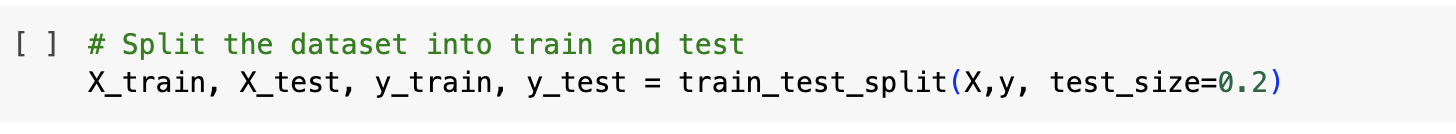


It predicted correctly that class=2 same as the y\_test output. So, the model is working fine.

#### Classification – Decision Tree

Decision trees are tree-like structures that represent a series of decisions and their possible consequences. They offer a visual and intuitive way to make decisions based on data.

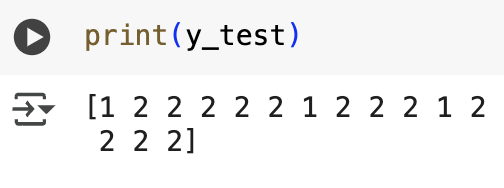


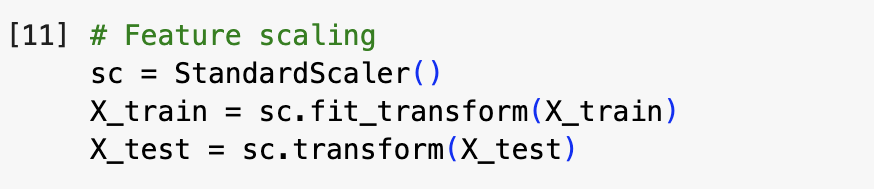


X:

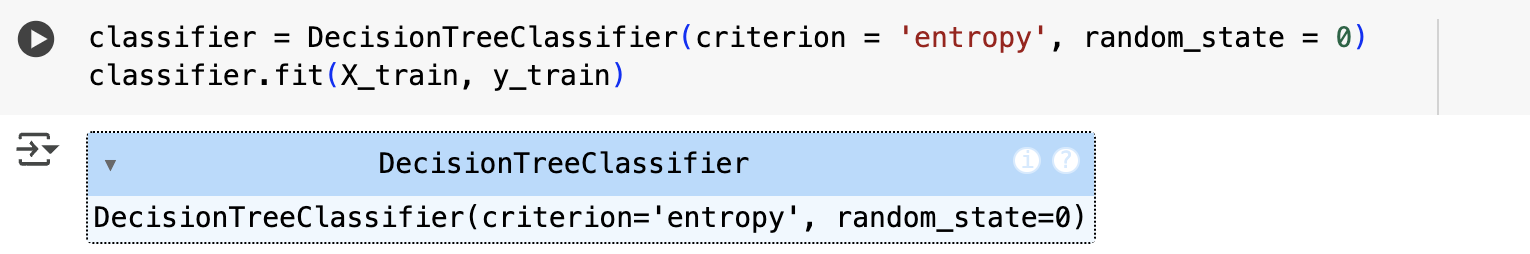
## 

Y:

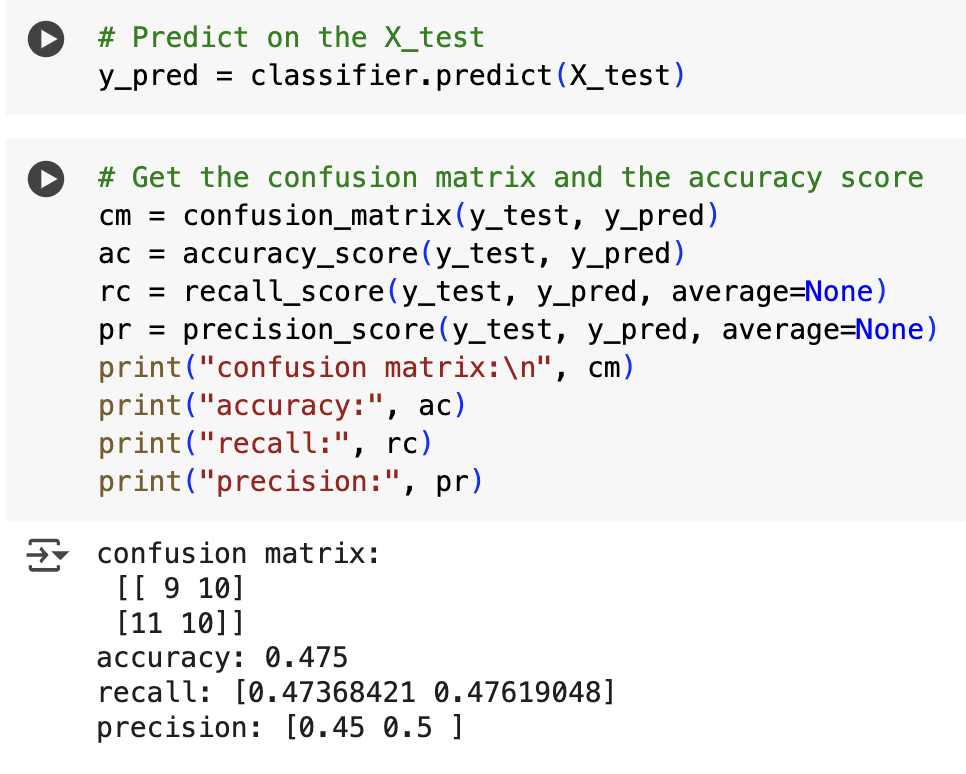




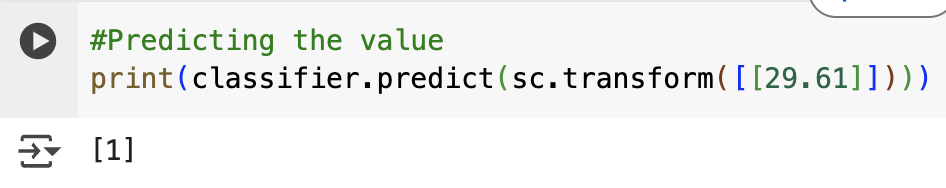
Then configured for the Decision Tree model by mentioning the classifier as Decision Tree and attached the classifier to the train dataset.



We then got the accuracy score and confusion matrix to evaluate the performance of our Decision Tree classification model.



As seen from the output, 47.5% of the predictions were correct.



It predicted correctly that class=1 same as the y\_test output. So, the model is working fine.

#### Classification – Evaluation

To evaluate each model and find what best fits our data set we will compare the four models, Logistic Regression, KNN, Naïve-Bayes, and Decision Tree, by comparing their errors and percentage of accuracy to do this we followed the formula: FP(False Positives) + FN(False Negatives).

Below is the result of performance metrics(confusion matrix) for classification models of our data set:

|  |  |  |
| --- | --- | --- |
| **Model** | **Errors** | **Accuracy** |
| **LR** | 7 + 15 = 22 | 45% |
| **KNN** | 8 + 10 = 18 | 55% |
| **NB** | 0 + 23 = 23 | 42.5% |
| **DT** | 11 + 10 = 21 | 47.5% |

The best fit model for our dataset is KNN model since it has the highest accuracy score and lowest false predictions out of all models.

## Section IIII

## Conclusions