**1. Part-01**

# Explain the dataset, conduct exploratory data analysis (EDA) and preprocess the data as needed to prepare it for modelling, discuss ethical/social issues.

**Explanation of dataset:-**

The dataset is of wifi signals measured from different locations within a building.

# Features:-

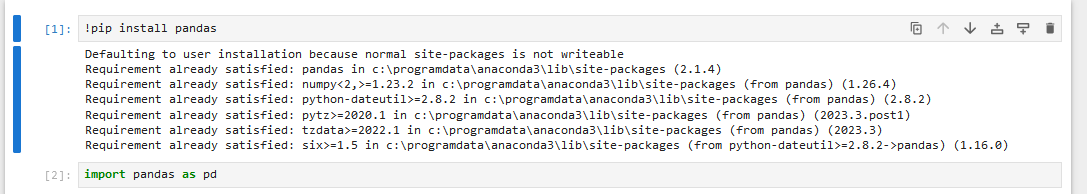
In this data set there are 7 columns named as wifi1, wifi2,wifi3…..wifi7. These columns represent the different strengths of wifi signals which are measured from different API’s.

# Target:-

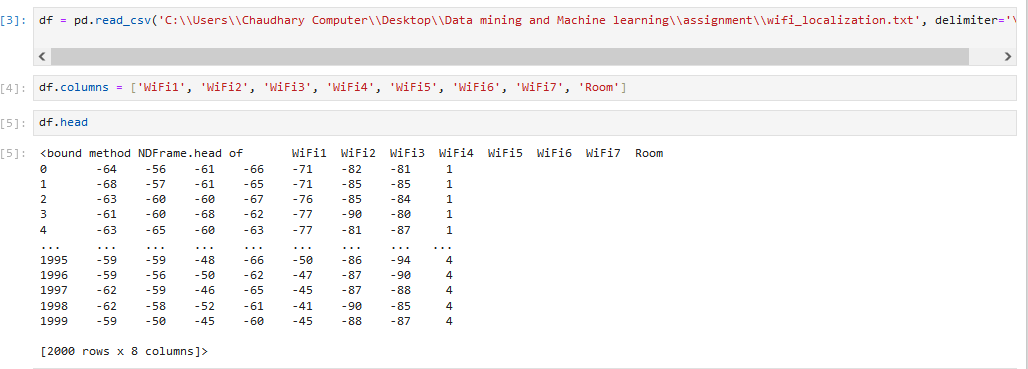
The last column which is named as Room, it represent the room number from where the measurement were taken. So in this dataset our target variable would be Room. We will apply classification model that can accurately predict the room number based on the Wifi signal strength.

**Reference:-** https://archive.ics.uci.edu/dataset/422/wireless+indoor+localization **Exploration of dataset:-**

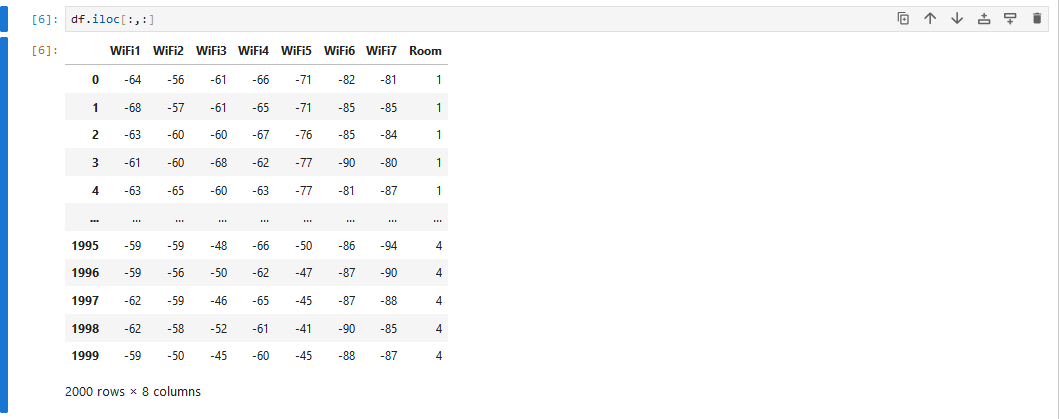
First of all install pandas packeg and import it as follow



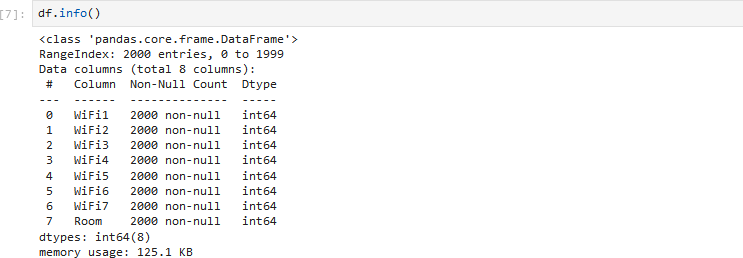
Then I create dataframe of my data using pandas as follow and assign the header manually and then see the head of data to conform that everything is right.



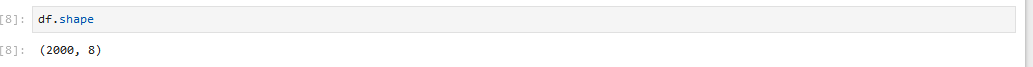
Then I show all elements using iloc



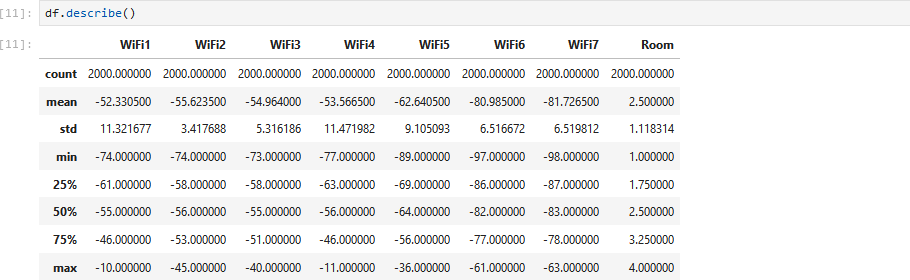
Then I use info function to see all the information regarding the dataset as follow.



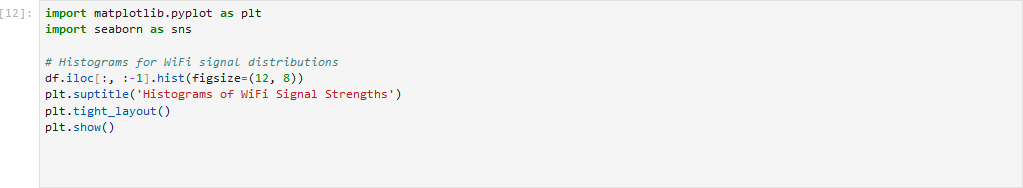
Then I write shape function to see the total number of rows and columns.



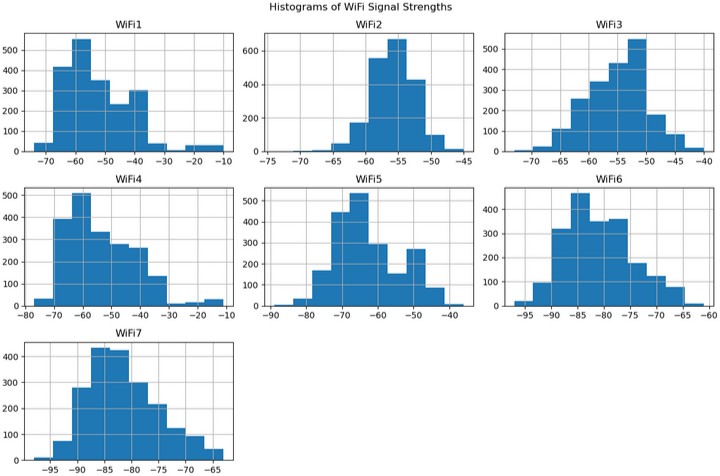
Then I use describe() function to see the descriptive statistics of my data frame as follow



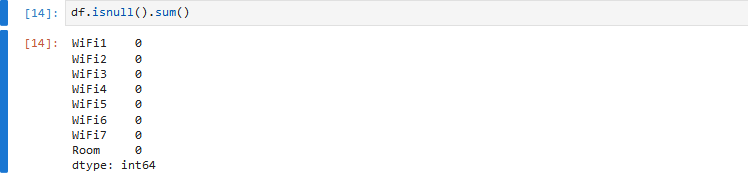
Then I plot the histograms of wifi signal strengths to see the distribution of data as follow

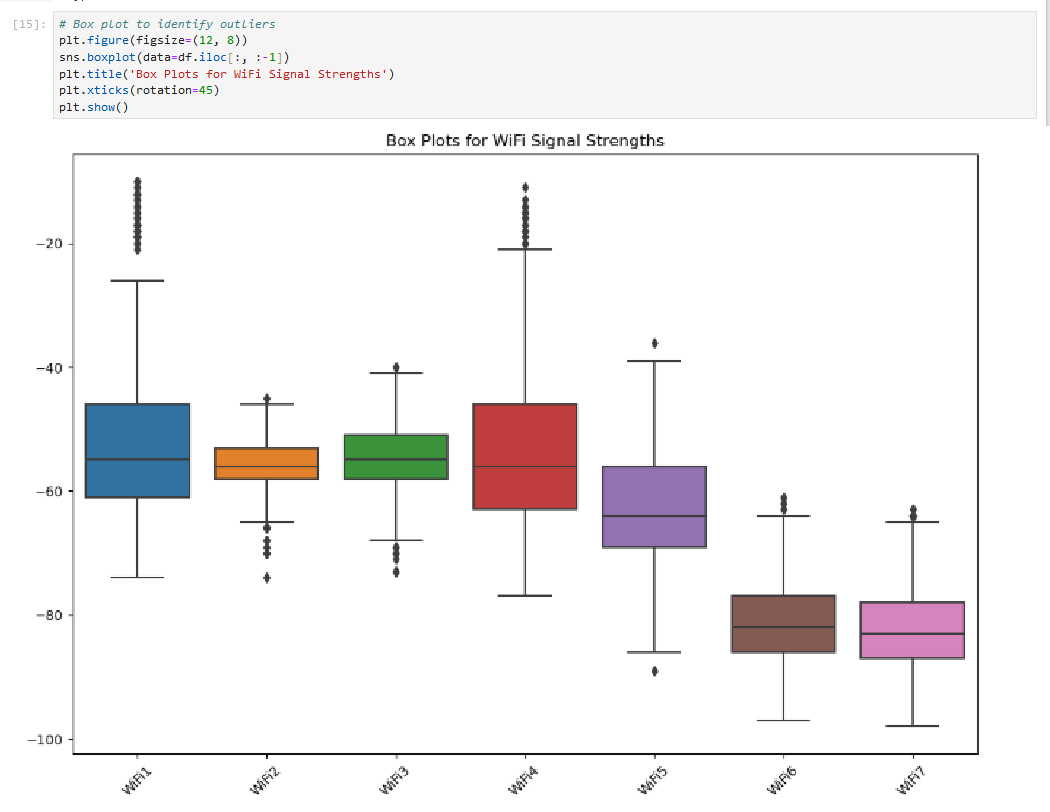


Results:-



Then I checked that is there any column which contain any null value as follow



From above you can see that there is no null value in my dataset. Then I plot the box plot to see the outliers as follow

Then I write a code to see that the wifi signal strengths have a linear relation with target variable or not. So in this way I can predict which classification algorithm may be more suited.



# Social/ethical issues:-

As the dataset is publically available on google so it is assumed that the data collection complied with ethical guide line.

The data is publically available so it is critical to ensure compliance with the lisence under which the data is shared on google.

The dataset is anonymized but I assured that it will not be misused by me.

I think wifi localization data would rise privacy concerns because we are indirectly tracking the users location without their conset. So to avoid this issue I am taking anonymous data for ethical purpose only.

# Apply two classification algorithms of your choice on your chosen dataset using Python. Provide detailed explanations of your data mining processes, including the rationale behind the choice of algorithms and parameters.

**Classification:-**

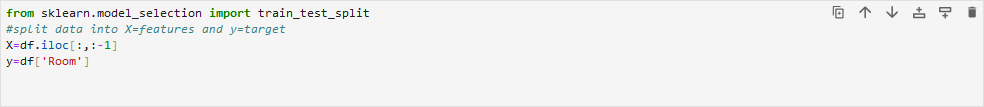
As we have seen earlier that Wifi-localization data is a classification problem. The objective is to classify the Wifi signals strenghts and predict to which Room these signals belong to. Now I am going to apply 3 different classification algorithms i.e KNN, Decision tree and logistic regression.

# KNN:-

I applied KNN because it is non-parametric algorithm and it does not make any assumption about the distribution of data. It can handle multiclass classification well and it is very suitable for data like our data where closeness in values is important.

# Determining the class feature and input features:

So to apply KNN first of all I am going to slice my dataset into input and output. The input data would be of features and the output data would be of target variable. Iam gonna represent input as X and output as Y.



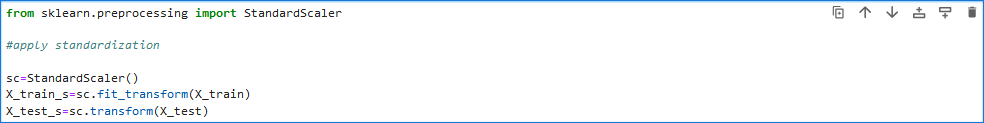
# Splitting the dataset into the Training set and the Test set:

Now I will split the dataset into traing dataset and test dataset as follow.



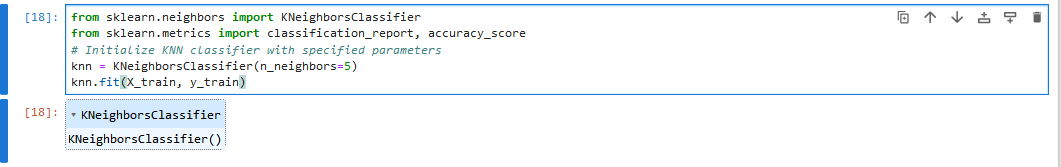
I have split the data-set into the ratio of 30 and 70, like 30% of data will be used for test and the rest data will be used for training. I have set the random\_state variable to 0, so that we can get different train and test set across different executions.

# Scaling features:

As every data-set contain unstandardize features on which we cannot apply KNN algorithms because KNN perform better when the dataset is standardize. So for this purpose I am going to standardize my dataset using fit-transform and transform method on training and test data respectively. This standardization will substract the mean and divide by the standard deviation to shift the distribution in a way that each feature have a mean of zero and standard deviation of 1.

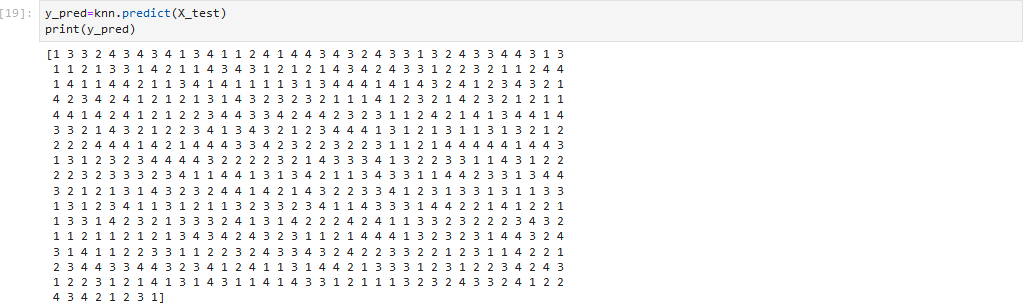
# Training the model:

Now I am going to apply K-neighbour classifier from scikit learn library.

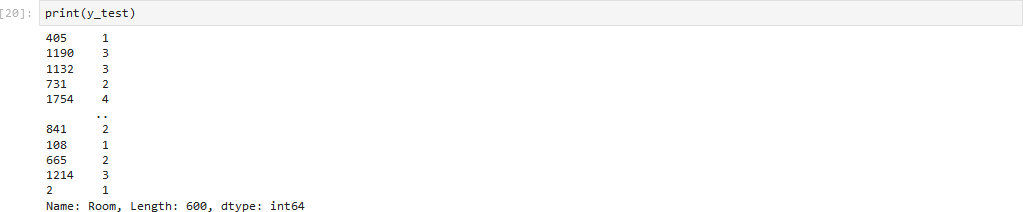


In this above code I have specified n\_neighbors=5. Now it will calculate the distance from nearest 5 elements.

# Evaluating the model:

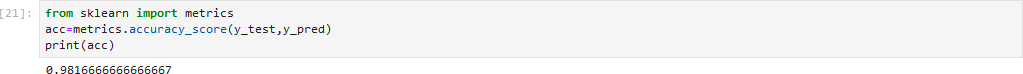
Once our model is trained I use “predict”function to see the predictions on our test data as follow.

From above you can see that these are the predicted values of target variable. Now I am going to print the real values of test data and will compare them with the predicted data.



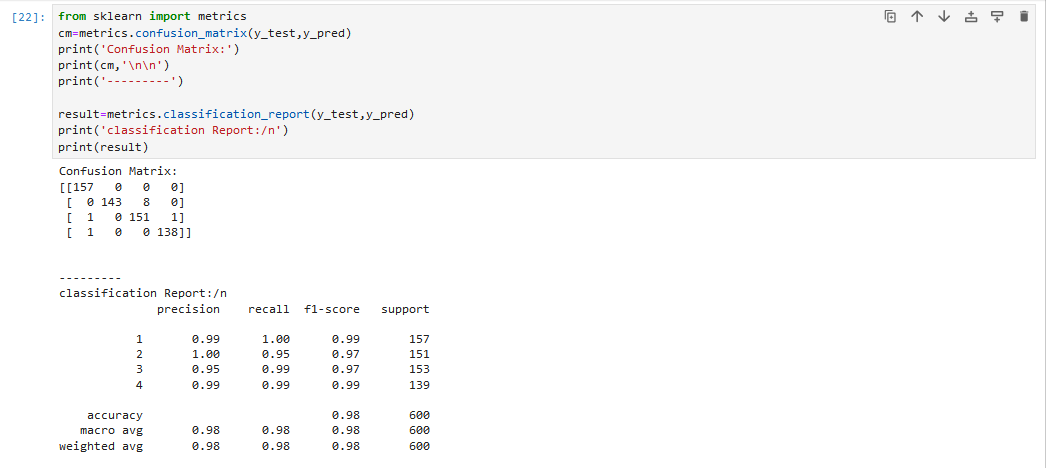
These are the value of real test data.

Now I am going to write a code that will compare both predicted and real test values and give us accuracy report.

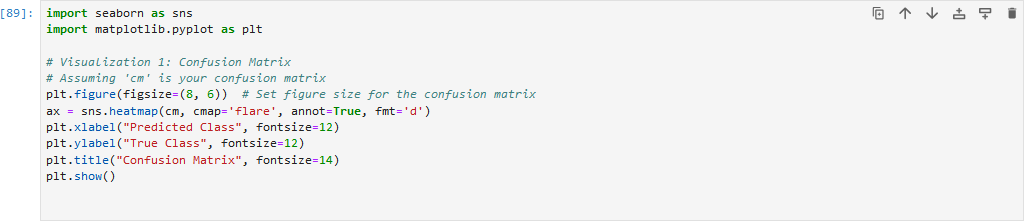


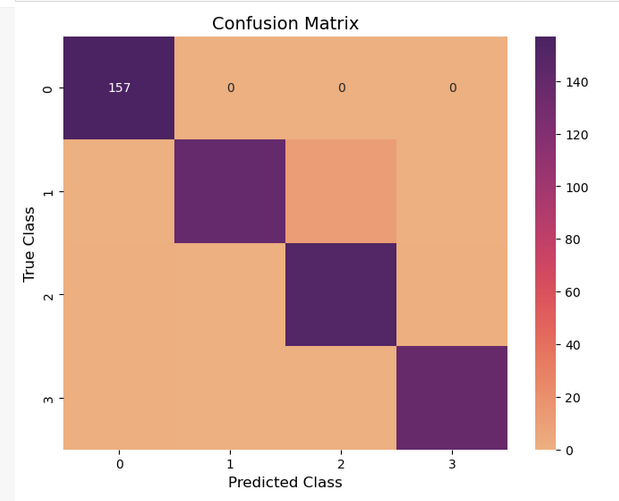
Now from this accuracy score we can see that it is 98% mean our model accuracy is 98% which is very good.

For more detailed evaluation of results I am going to make confusion matrix.



Then I plot the confusion matrix





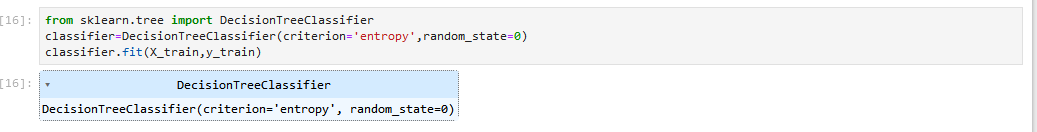
I don’t know why it is showing only 1st row I tried multiple time to fix the issue but due to shortage of time I am not gonna wait more time on it but we have proper confusion matrix in written form we will use it for further analysis.

# Decision Tree:

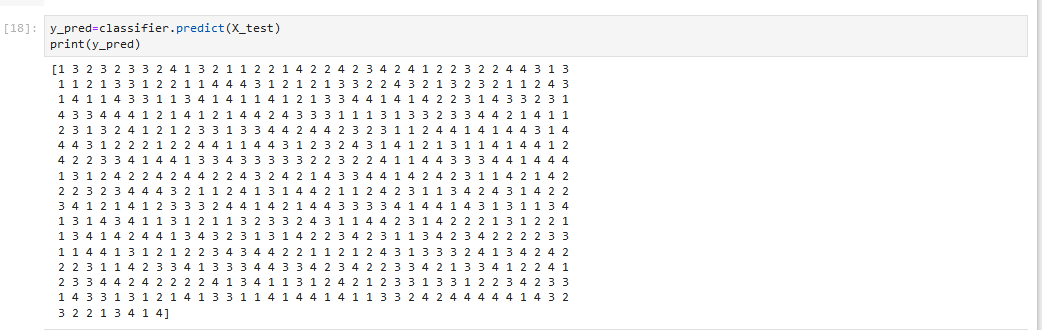
I applieddecision tree algorithm because it can handle both numerical and categorical data like my data that I have used . It is very well suited for non-linear data like there is non linear relationship between wifi strengths and rooms So it will predict better.

Now I am going to apply decision tree algorithm. I make a new notebook copy and paste the exact same procedure for KNN but with a few changes that I remove the KNN model and write decision tree model and remove the scaling part as well.

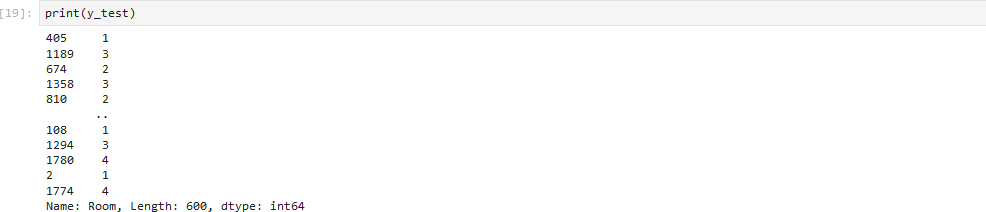
I apply the decision tree model as follow



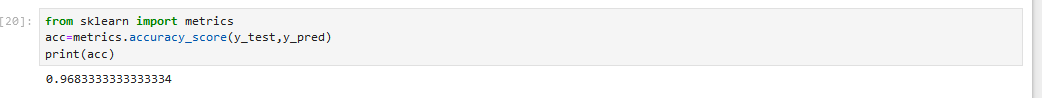
Once our model is trained I use “predict”function to see the predictions on our test data as follow.



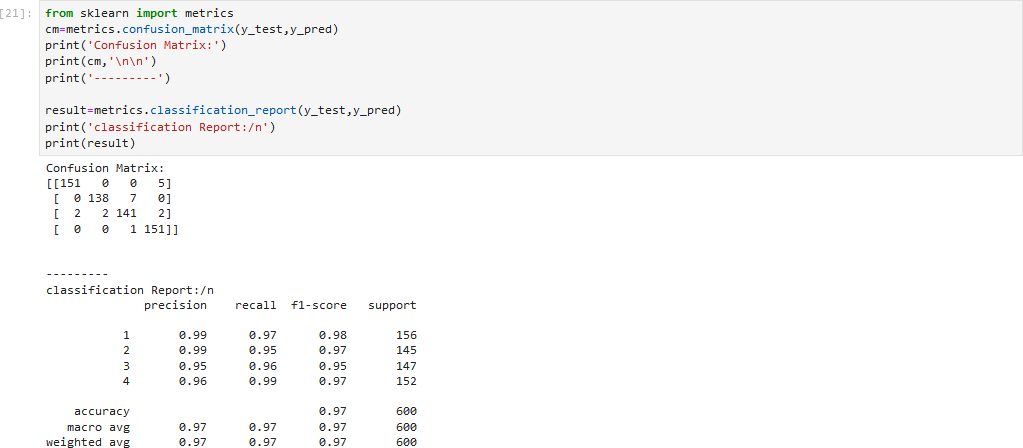
Then I print the original value of y\_test to see the difference in y\_pred and y\_test.



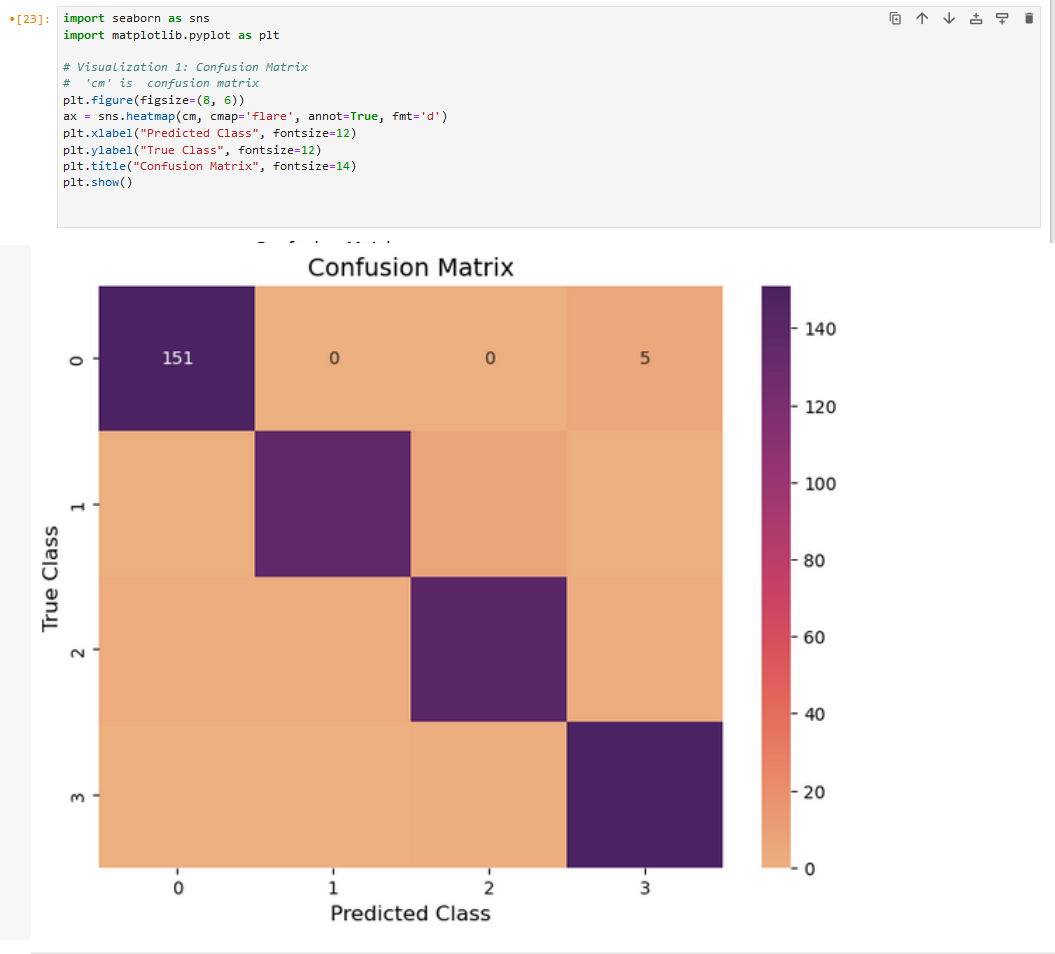
Then I use the accuracy\_score measure from sklearn to compare the difference give accuracy score as follow



so its mean our model predict 96% accuracy, that’s very good.

Now for more detail analysis of results I use the confusion matrix as follow

then I plot the confusion matrix as



I don’t know why it is showing only 1st row I tried multiple time to fix the issue but due to shortage of time I am not gonna wait more time on it but we have proper confusion matrix in written form we will use it for further analysis.

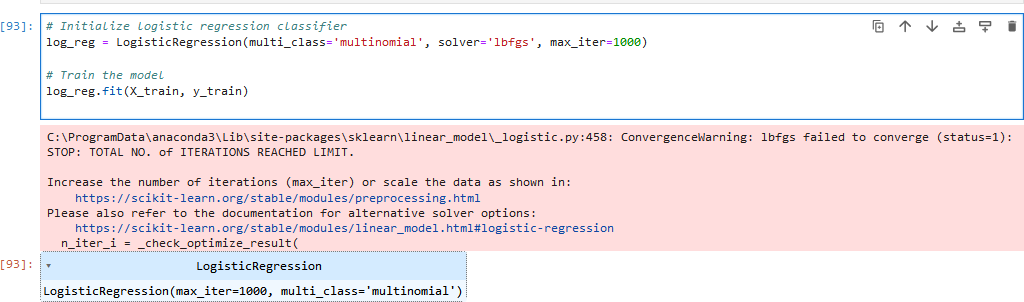
# Logistic Regression classifier:-

I apply logistic regression classifier model because I works we when the relationship between feature and target is close to linear like in our case sometimes wifi signals exhibit somehow linear separability between Room and features.

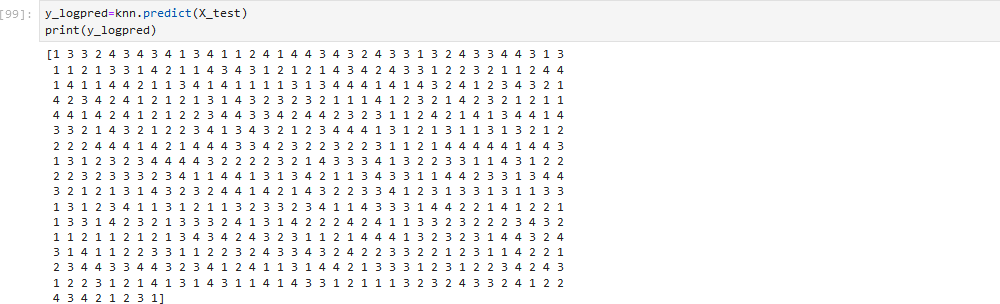
Then I apply logistic regression classifier. For this model I did not change anything from the steps of KNN. I just replace the KNN model with logistic regression classifier model. First of all I import logistic regression



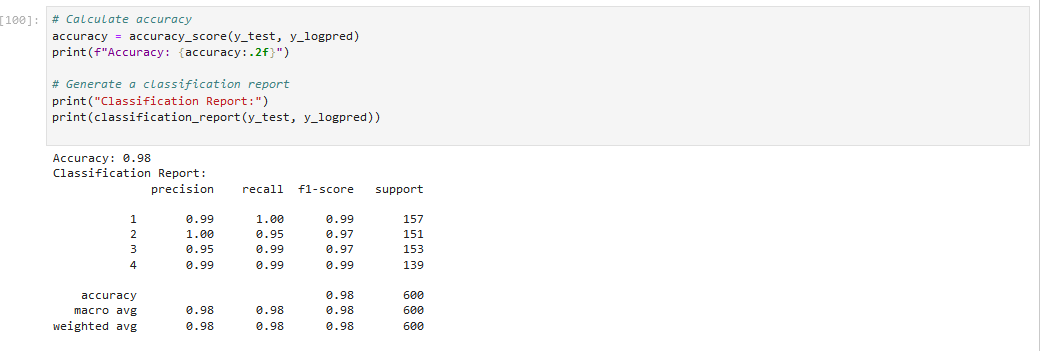
Then I replace the KNN model with logistic regression model as



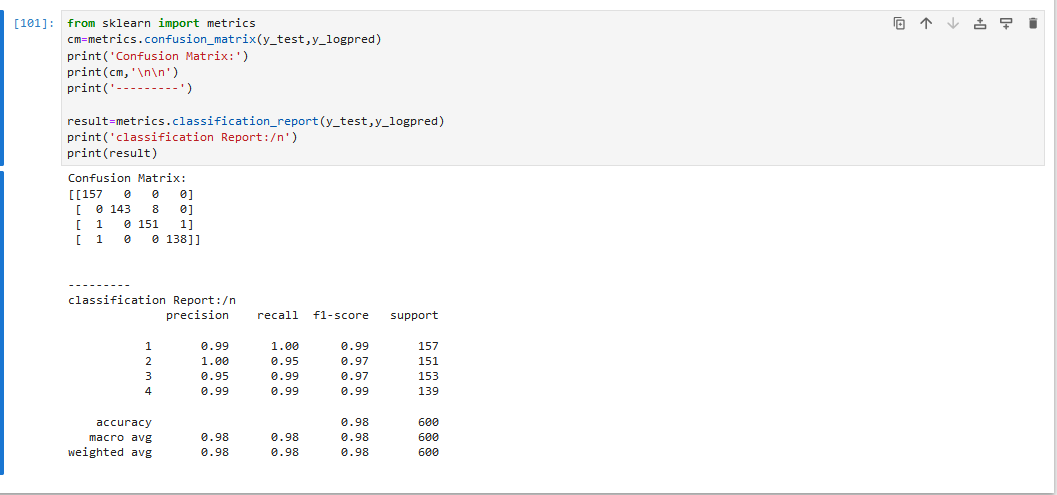
then I see the predicted values by logic regression model as



Then I see the accuracy of this model by comparing this y\_logpred values with the original values as



From above we can see that our model accuracy is 98% which is very good. Now I will write a code to see the more details of the result in the form of confusion matrix as



# Present the results, compare the performance of the two algorithms, justifying performance metrics

Confusion Matrices for Logistic Regression Classifier:-



Confusion Matrices for Decision Tree Classifier:-



# Insights:

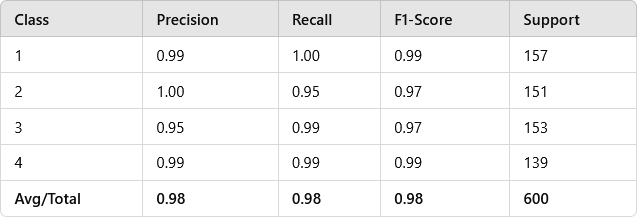
* **Logistic Regression**
  1. In logistic regression class 1 has predict all the 157 instance correctly.
  2. The class 2 contains 8 misclassification into class 3.
  3. Class 3 and class 4 contains minimal misclassification.

# Decision Tree:

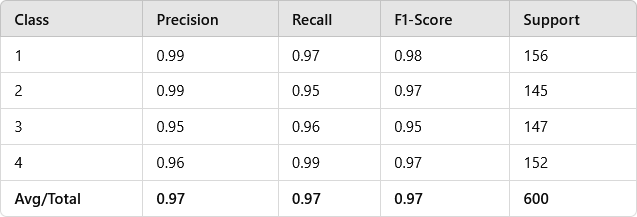
* 1. For class 1 there are 5 instances that are misclassified into class 4.
  2. For class 2 there are 7 instances that are misclassified into class 3.
  3. Class 3 shows 2 FP and 2 FN.
  4. Class 4 shows no FN.

**Classification Reports:-**

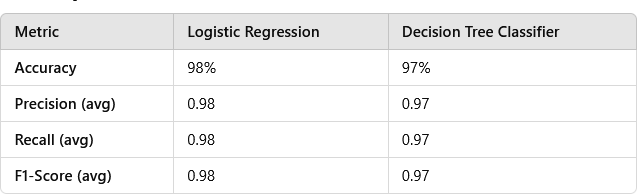
* + - **Logistic Regression:**

****

* + - **Decision tree**

****

**Comparison of Performance Metrics:-**



* + - Logistic regression achieves slightly higher accuracy (98%) when compared to decision tree classifier(97%)
    - Logistic regression acheives slightly better precision 0.98 than decision tree classifier (0.97).
    - Logistic regression achieves slightly higher recall (0.98) when compared to decision tree classifier(0.97)
    - Logistic regression also leads in F-1 score as well.

# Performance matrix:-

* + - **Accuracy:-**

Accuracy is a simple measure to understand the overall performance of the model. But it is not enough if the dataset imbalanced or if the the misclassification in certain classes are more. So that why we consider other performance matrix for judgement.

# Precision and Recall:-

These matrix provide deeper insights into model’s ability like avoiding false position(precision) and false negative(recall).

# F1-Score:-

F1 score balances precision and recall, when the class distributions are imbalanced.

# Recommended:-

By comparing the results of both algorithms I recommend logistic regression model because it is performing well than other algorithm.

# You should critically evaluate the classification output and discuss how it will benefit the related business or help solve the problem at hand. Explain how the insights gained from the model could influence decision-making, or improve processes.

**Overall performance:-**

Logistic regression achieves better overall performance than decision tree. It achieves 98% accuracy mean it correctly classifies most instances of room localization based on wifi signals.

The confusion matrix of logistic regression shows minimal misclassification than decision tree classification model.

The f1-score of logistic regression classifier model suggest better balance of precision and recall effectively.

Both the models have slight difference between them. So we can say that both models can perform in real world.

# Business Applications

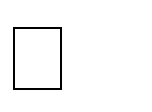
The primary goal of this project is to identify the location of the user within specific room of the building building based on their Wifi-signal strength. So there are many business applications for this model but I make a business application of this model as **Automated Rooms:-**

This model can be used to automate the room. Like this classification model is just predicting the room based on the Wifi signal. Once the Room is predicted, the system triggers specific automation tasks for that room.

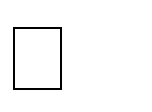
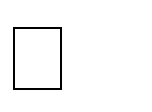
How it works:-

This room localizer will be integrate with IOT devices like( smart lights, thermostate and speaker) and a central control system(a home automation hub).

When this classification model predict the room in which the user is present using their wifi signal strength it will transmit this information to IOT devices. So these IOT device will be pre-programmed for each room like for example

**Room 1**: Turn on overhead lights and adjust the thermostat to 22°C.

**Room 2**: Play relaxing music and dim the lights.



**Room 3**: Switch on a projector for presentations. Then these IOT devices will perform their function. **Benefits:-**

Lights automatically turn on when the person enter into room and turn off when he leaves the room.

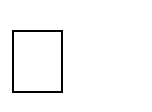
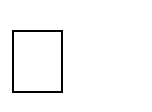
User can control the climate of his room. This process is energy efficient.

# Actionable Recommendations to Leverage Insights:-

* + - In this above discuss project if these classification model is used than it can predict the placement of user based on their wifi signal strength.
    - By analyzing confusion matrix, business can identify the problematic areas and solve the issue.
    - Devices and applications will only be active when needed.
    - It will enhance user experience.
    - It would be energy efficient.

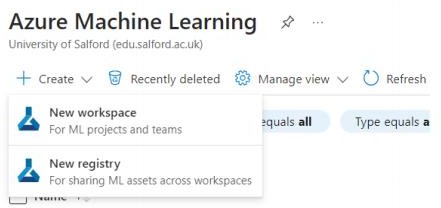
# b) Use Azure Machine Learning Designer to preprocess and apply two classification algorithms to the same dataset as you used for part a.

**Evaluate and compare the results**

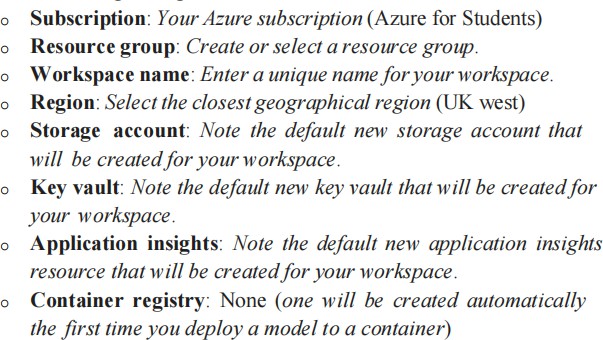


For applying azure machine learning first of all I sign into my azure Microsoft account Then I click on +create a resource and set a new workspace

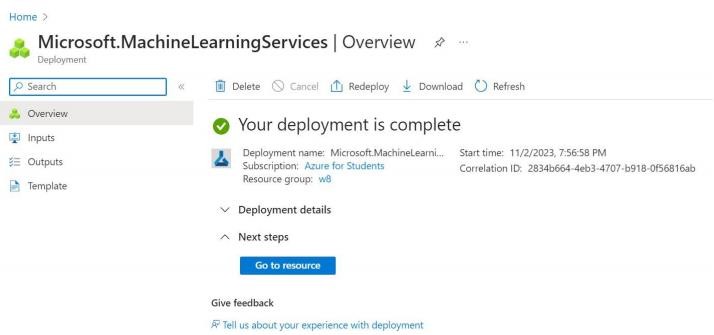




then I use the following settings



Then I click on Review +create button and select create and then click on “Go to resource”.

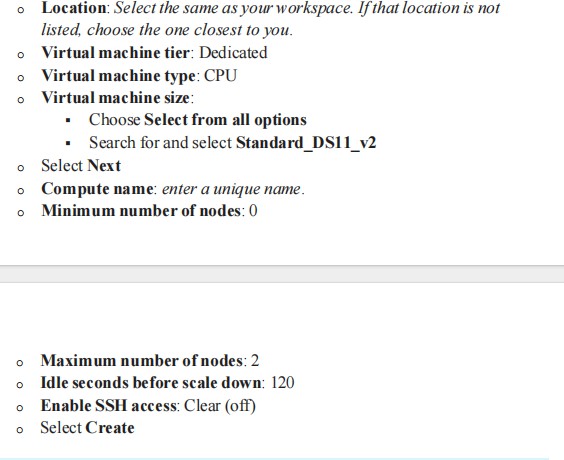


Then I launch studio

Then I choose my workspace for this project.

**Create compute:-**

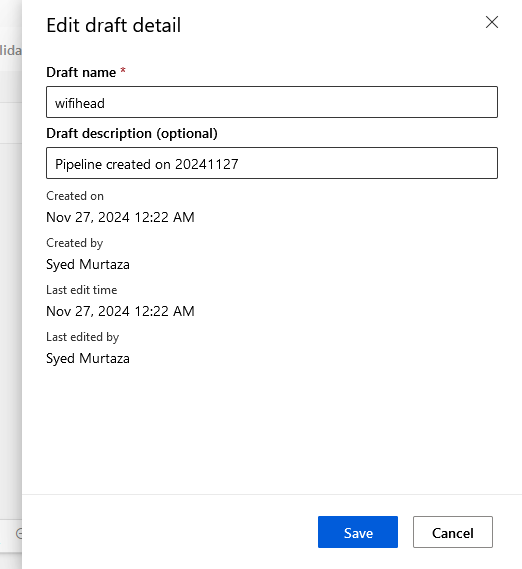
On the compute page I select compute cluster tab and add a new cluster with the following settings.



**Create a pipeline in Designer:-**

In azure studio I expended the left hand pane than click on Designer and select + to create a new pipeline

Then I click on pen icon near the name of pipeline and change the draft name and click on Save.

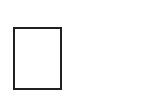


**Create a dataset:-**

Click the data page and create a dataset from this page by uploading the data from my laptop as follow

First of all I click on +create to create a dataset using following settings.

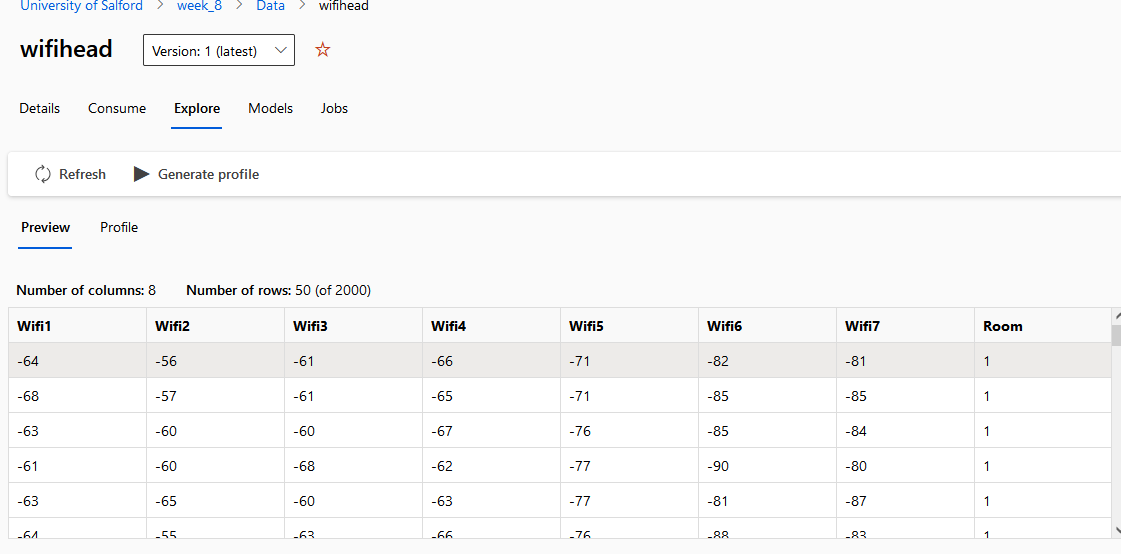
**Basic Info**:

**Name**: wifihead **Description**: wifi data **Dataset type**: Tabular

Then I select Next and click on upload to upload some data from my laptop and select Next

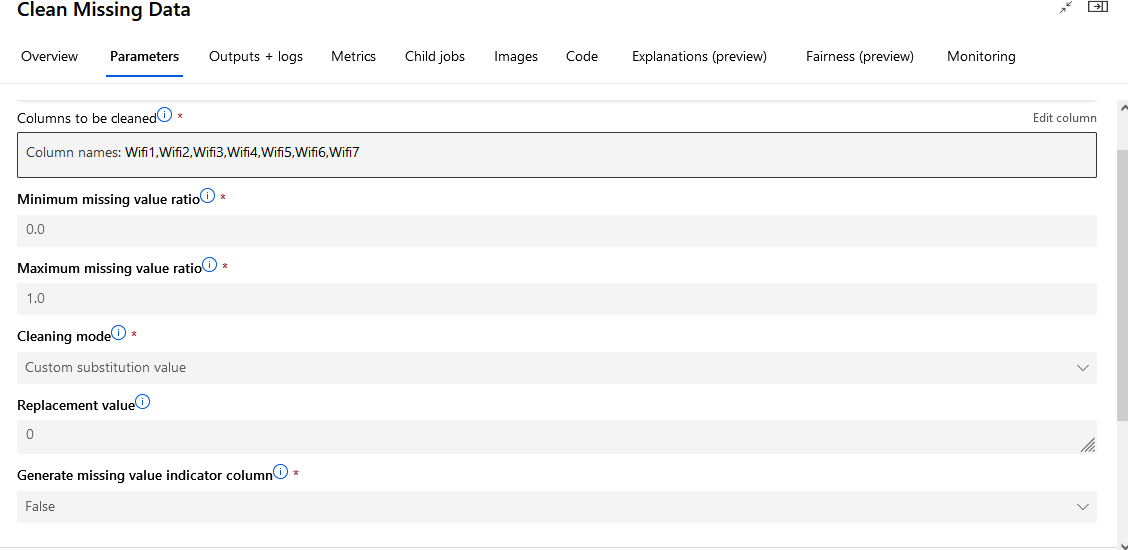
Then I do the setting and preview things by consulting from lab work and Click on Create.

After the data has uploaded the I view the explore tab to see a sample of data.

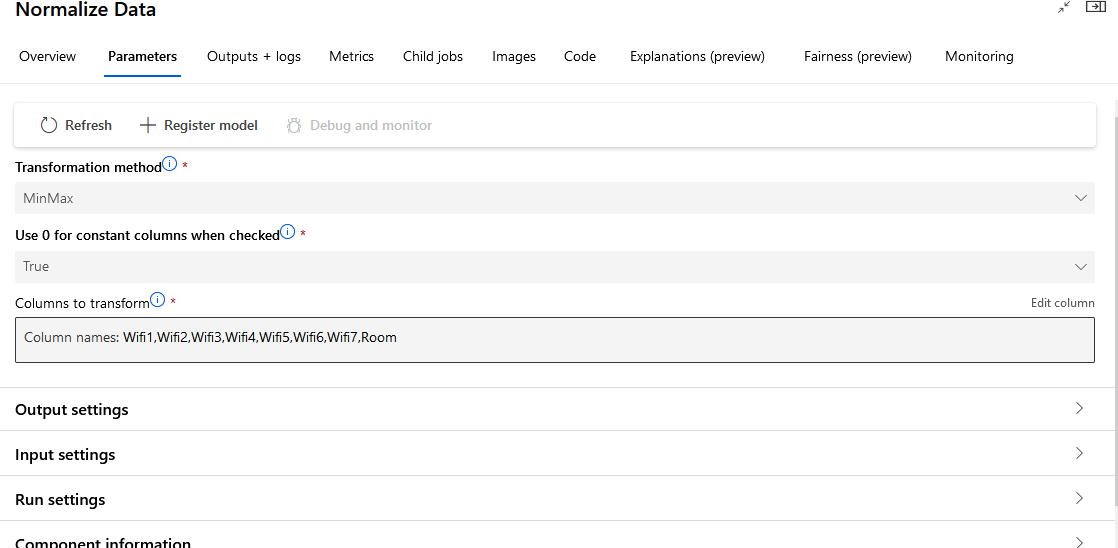


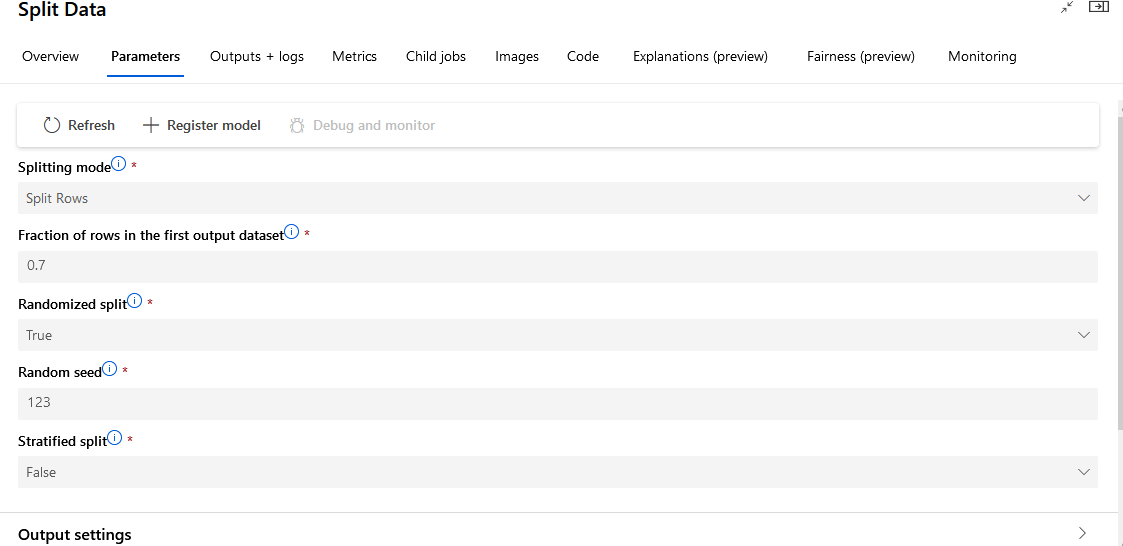
**Load data to canvas:-**

* First of all I click on data and search my my dataset (wifihead) and drag it on to the canvas.
* Then I click on the component and search for Clean missing data module and place it to the canvas and join it with the data. I used the following parameters for Clean missing value module

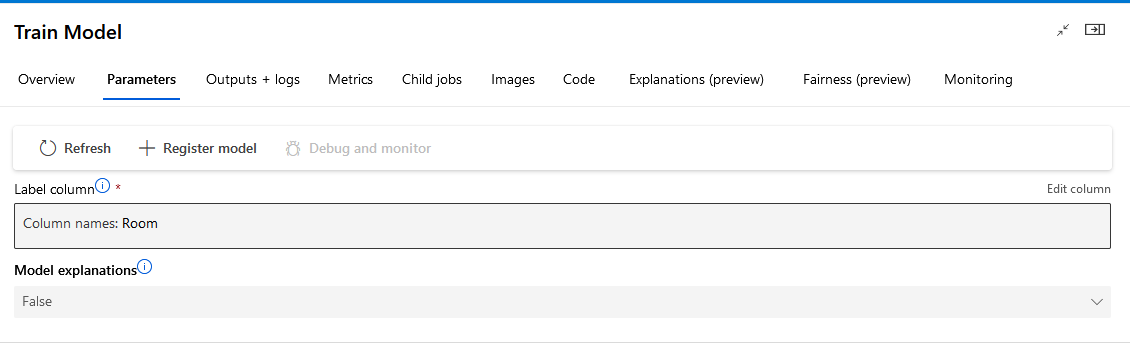


* From components I search for Normalize data module and drag it to the canvas and connect it with clean missing data module. I used the following parameters in Normalize data.



* Then I search for split data module and drag it to the canvas and connect the left side (transformed data) of the normalized module to the head of split\_data module.
* 
* Then I search for train data module and drag it to canvas and connect bottom left

side of split\_data to right side of Train Model.

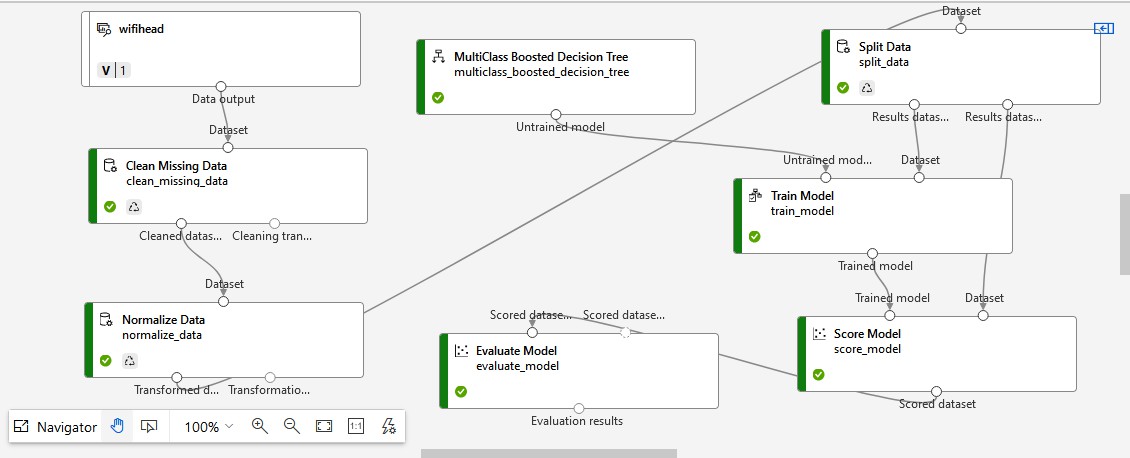
* ​
* Then I search for Score Model and drag it to canvas and connect bottom of train

model to upper left side of Score model.

* Then I search for Evaluate Model and connect bottom of Score Model to the upper left of Evaluate model.
* At the end I searched for multiclass boosted decision tree module and drag it to canvas and connect upper left side of train model to the bottom of multiclass boosted decision tree module.

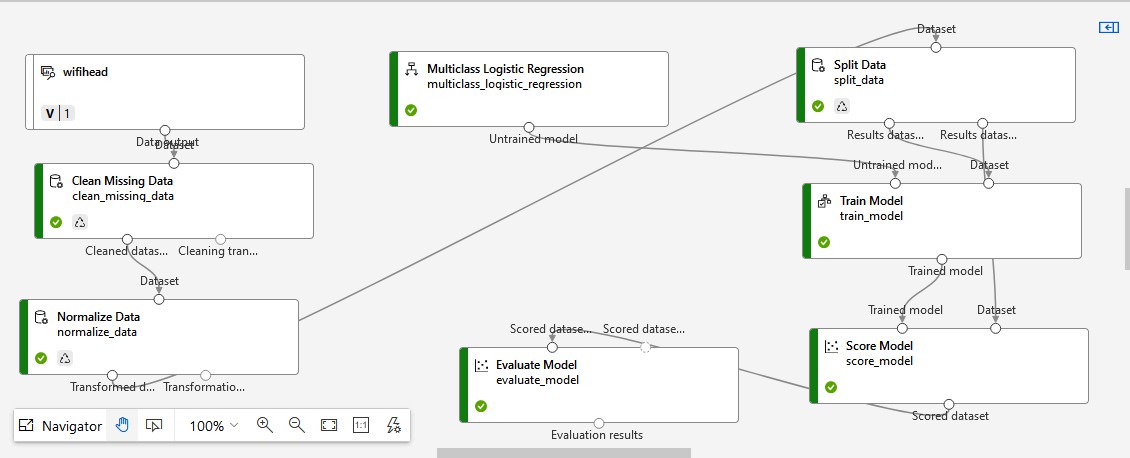
**Run the pipeline:-**

* + Then I click on configure and Submit.
  + For runtime I attached that cluster that I made before and simple naming the experiment name I click on Submit.

**Overall pipeline:-**

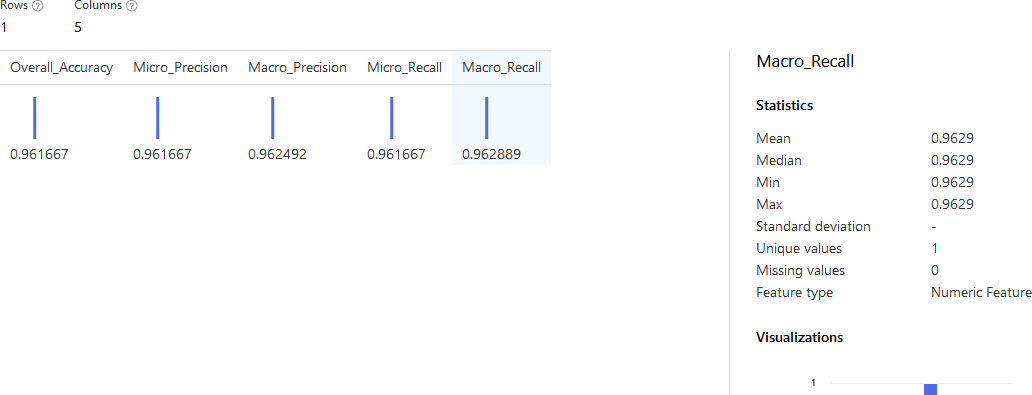
**2 pipeline:-**

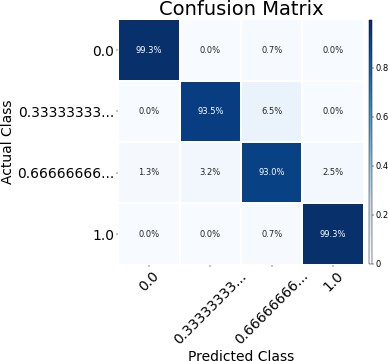
For the second pipeline I did all steps as I did in 1 pipeline but I changed the multiclass bossted decision tree with logistic regression classifier module.



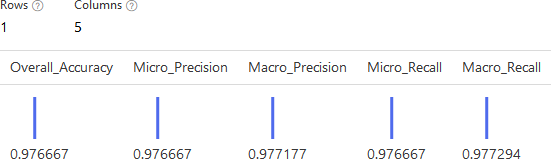
**Results:-**

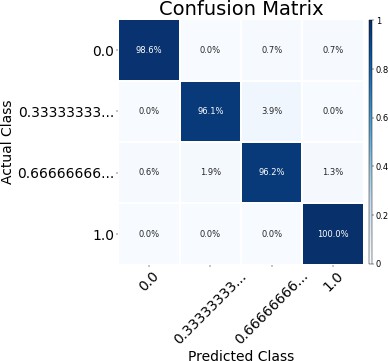
The result for multiclass logistic regression is



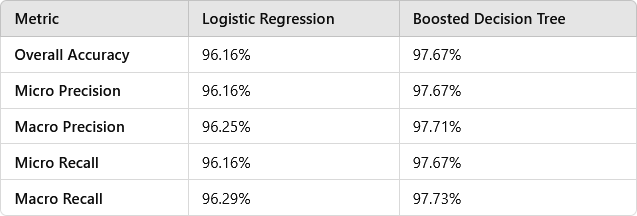


Result set for multiclass boosted decision tree is





**Summary:-**



**Conclusion:-**

From above summery we can visualize that the boosted decision tree performs better than the logistic regression across all the metrics.