# Technical Documentation

# Introduction

## Type of Machine Learning

* + Explain the main differences between supervised and unsupervised learning.

|  |  |  |
| --- | --- | --- |
|  | **Supervised Learning** | **Unsupervised Learning** |
| **Definition** | The supervised learning method in machine learning makes use of labeled datasets to teach computers to accurately categorize input or predict outcomes. The model evaluates the significance of several features to gradually enhance the model fit to the known result using the labeled data. (alteryx, 2022) | Algorithms are used to investigate and classify unlabeled datasets in unsupervised learning. Without human oversight, these algorithms may find hidden patterns in data. (alteryx, 2022) |
| **Applications (use)** | 1-speech recognition  --by training it on voice  -Google Assistant  - Siri  2-Bioinformatics  -Fingerprints  -earlobes  (Naukri, 2022)  3-Price prediction | 1-Medical imaging  2-Computer vision  - object recognition  3-Prepare data for the supervised learning |
| **Strengths** | -often used to manage a wide range of issues  -Gain knowledge from little pieces of data.  -Make reliable predictions  (Naukri, 2022) | 1-It can perceive things that the human mind cannot.  2- being able to spot hidden patterns |
| **Limitations** | The model could function well with the training data but badly with fresh, untried data. | You cannot obtain exact data on the data sorting. |
| **Common Algorithms** | -Decision trees  -Random forest  -Naive Baves | -K-Means  -Gaussian Mixture Models  -Frequent Pattern (FP) Growth |

|  |  |
| --- | --- |
| **Supervised** | **Unsupervised** |
| The input data is labeled | The input data is Unlabeled |
| Depends on training Dataset | Depends just on input dataset |
| A training dataset is used to classify the data. | classifies supplied data using its properties. |
| Used for prediction | Used for Analysis |
| split into two categories Classification & Regression | split into two categories Clustering and Association |

## Supervised Machine Learning

* + Explain the main differences between regression and classification techniques.

|  |  |  |
| --- | --- | --- |
|  | **Classification** | **Regression** |
| **Type of Learning** | Supervised Learning | Supervised Learning |
| **Applications (use)** | Identifying which group a new observation belongs to, such as determining whether an email is spam or not. | Predicting a continuous outcome, such as the price of a stock or the temperature tomorrow. |
| **Strengths** | Good at handling categorical variables and non-linear decision boundaries. | Can handle a large number of input variables and can predict continuous outcomes. |
| **Limitations** | Can be sensitive to small fluctuations in the training data and may not be able to predict continuous values. | Can be sensitive to outliers and may not be able to handle categorical variables well. a |
| **Common Algorithms** | Logistic Regression  Naive Bayes  K-Nearest Neighbors  Decision Tree  Support Vector Machines | Linear Regression  Decision Tree  Support Vector Regression |

|  |  |
| --- | --- |
| **Regression Algorithm** | **Classification Algorithm** |
| The output variable in regression must have a continuous or real value. | The output variable for classification must have a discrete value. |
| The regression algorithm's job is to map the continuous output variable to the input value (x) (y). | The classification algorithm's job is to associate the discrete output variable with the input value (x) (y). |
| Continuous data are employed with regression algorithms. | With discrete data, classification algorithms are applied. |
| We look for the best fit line in regression in order to more precisely forecast the result. | When classifying data, we seek out the decision boundary that may categorize the dataset. |
| May be used to resolve regression issues like weather forecasting and house price prediction, among others. | May be used to identify spam emails, recognize speech, identify cancer cells, and other classification-related issues. |

## Computing systems

Investigate the computing systems used in cloud computing services (such as AWS and Azure), and compare them in terms of structure, GPU, CPU, type of storage devices used, and other related hardware aspects.

Leading cloud computing services AWS (Amazon Web Services) and Azure (Microsoft Azure) both provide a variety of computing resources, such as virtual machines, storage, and networking. Although both platforms employ a range of hardware platforms to offer their services, there are some significant distinctions between them. (BasuMallick, 2022) (Kinsta, 2022)

**Structure**

The infrastructure used by AWS is distributed globally into regions and availability zones. There are several availability zones, which are independent data centers that offer low-latency network connections, inside each region, which is a distinct geographic location. In contrast, Azure makes use of a worldwide network of data centers called regions, each of which is composed of a number of fault-tolerant clusters called availability sets. (BasuMallick, 2022) (Kinsta, 2022)

**GPU**

For machine learning and other high-performance computing applications, AWS provides GPU-powered instances. These instances come in a range of sizes and employ NVIDIA GPUs. Additionally, Azure provides GPU-powered instances for high-performance tasks like machine learning. These instances come in different sizes and employ AMD and NVIDIA GPUs. (amazon, 2020) (Alpaydin, 2021)

**CPU**

For its virtual machines, AWS provides a wide selection of CPU alternatives, including Intel and AMD CPUs. Additionally, Azure provides a range of CPU choices, including as Intel, AMD, and ARM CPUs. (amazon, 2020) (Alpaydin, 2021)

**Type of storage devices used:**

AWS provides a range of storage choices, including Simple Storage Service (S3) for object storage and Elastic Block Storage (EBS) for permanent block storage. Azure provides several different storage choices, such as Azure Blob Storage for object storage and Azure Disk Storage for durable block storage. (Kinsta, 2022)

**Other related hardware aspects.**

Other hardware choices including load balancers, network interfaces, and firewalls are provided by both AWS and Azure. Additionally, both systems provide a wide range of software choices, including databases, analytics, and developer tools. (Kinsta, 2022)

# Methodology

## data science life cycle

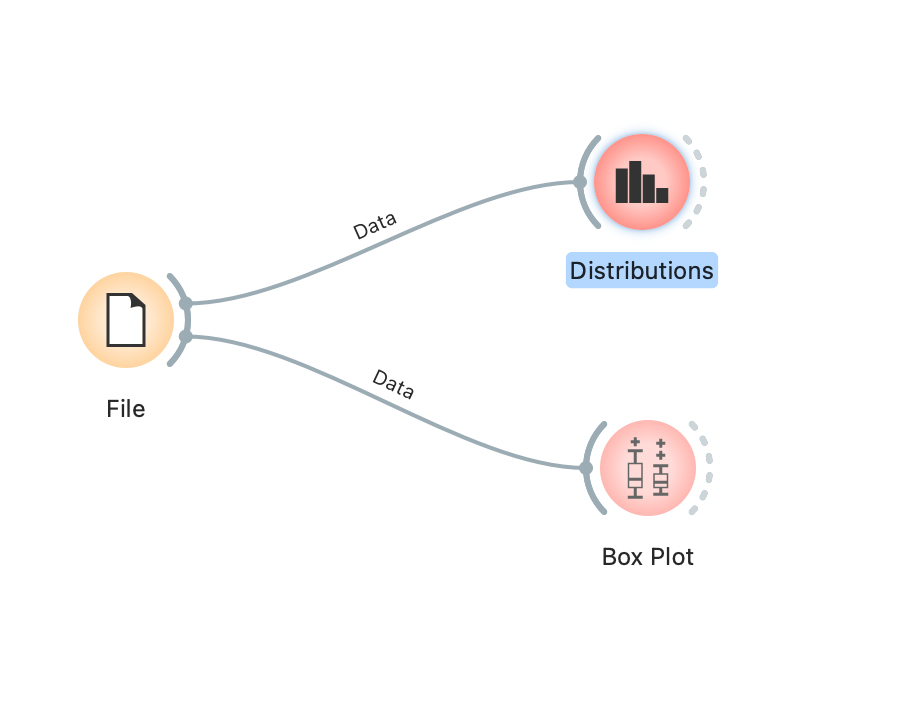
* + Discuss and explain the data science life cycle **of your developed prediction model** (any of them).

Taking the KNN model

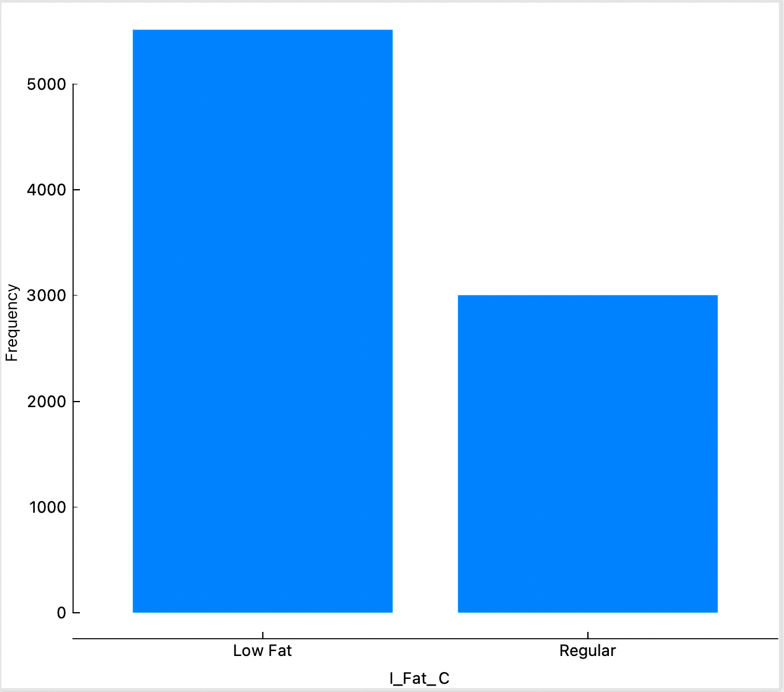
|  |  |
| --- | --- |
| **The Data Life Cycle Stage** | **The importance of the given Stage toward your project lifecycle** |
| 1. Understanding the Problem | This phase entails comprehending the problem in question and setting the project's goals.  The problem in my project was to predict the store from which the product was bought  The importance of this stage: will make me know my goal |
| 2. Gathering Data | The data was gathered by the organization which means I did not gathered any data I got it as a CSV file  The importance of this stage: because it will be used to train the model on it |
| 3. Cleaning Data | The data must be in a format that makes it simple to use them for the KNN model, therefore this step is crucial. It entails locating and erasing any missing, not correct values, or inconsistent data.  So I filled all the null values in the column that contains nulls and there were in a column the same values but in different spelling so I managed them to be the same spelling.  The importance of this stage: the accuracy and the model will be better depending on how much the data is clean and realistic |
| 4. Exploring Data | In this stage, I oversee all the data to if it is discrete or continuous to know what model I will use for the classification or the regression  The importance of this stage: tell us which model to use based on the type of values we have if it was discrete of continuous |
| 5. Feature Engineering | In this stage I oversee all the data and I delete Every value and feature is not necessary for the prediction of the results, the columns that are not necessary and to avoid too many unrelated variables. Even if there were a coloration between two columns dose not mean that we need the columns like the I\_MRP\_JD  (Maximum retail price of the product in Jordanian dinar column I know that were a coloration between this column and other columns bs I don’t need it is just not needed information  The importance is to improve the result when I apply my model because it minimize the dimension of the data set. |
| 6. Modelling Data | This step involves training a model that can generate predictions using the data. This is significant since it forms the basis of the project and impacts how accurately the forecasts turn out.  I choose the KNN model and then I split the data to x and why I have trained the model on 0.20 of the values of the column O\_T |
| 7. Interpreting Data | This stage is very important because it allows as to understand the data very good by doing visualize the data for better understanding and by Various plots by this stage I will be able to know every detail about my data set and also I can know how much is good or I have to improve it by calculating the accuracy and the precision and recall and the f1score |

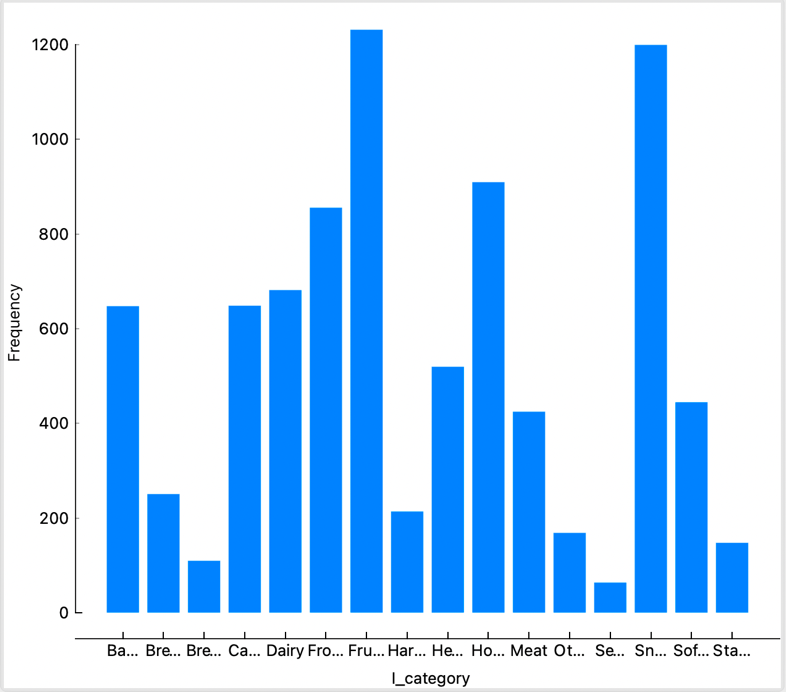
Note that one of life cycle stages requires visual plot(s) for each feature in the dataset.

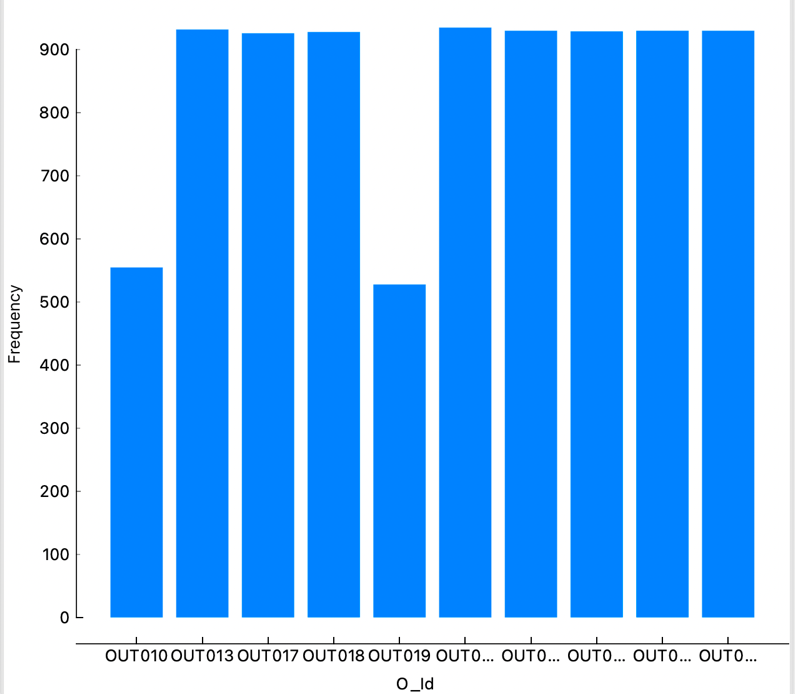
**4. Exploring Data: I used the orange application to make a visualization**

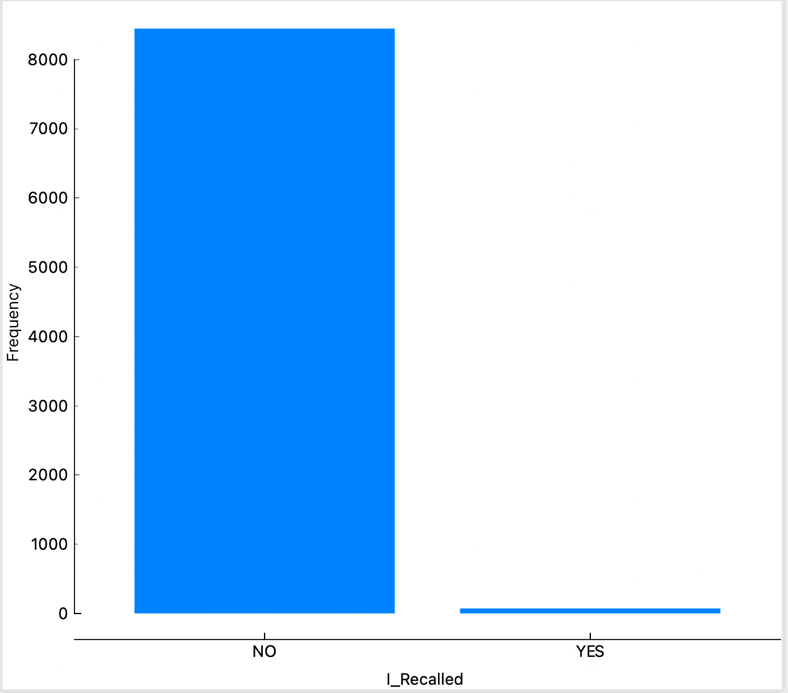
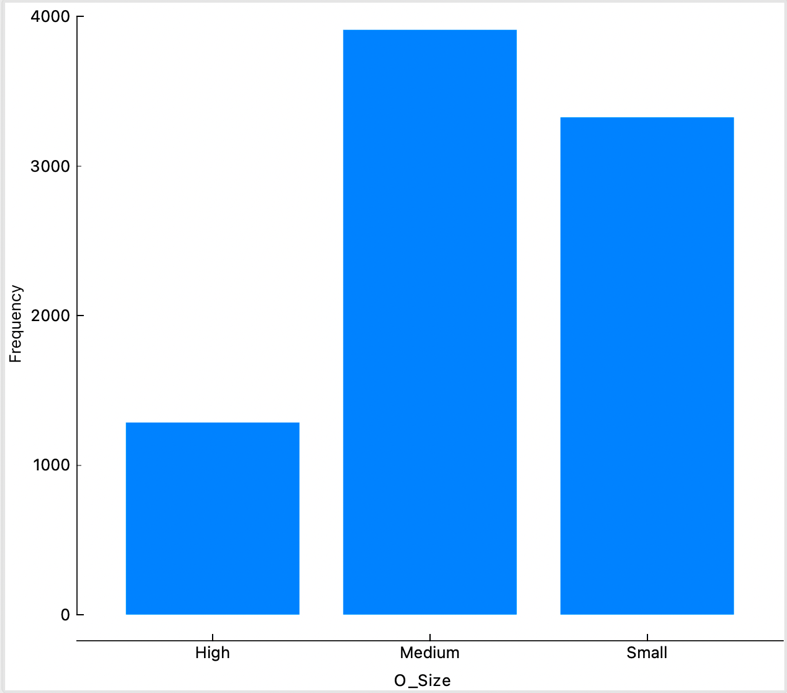


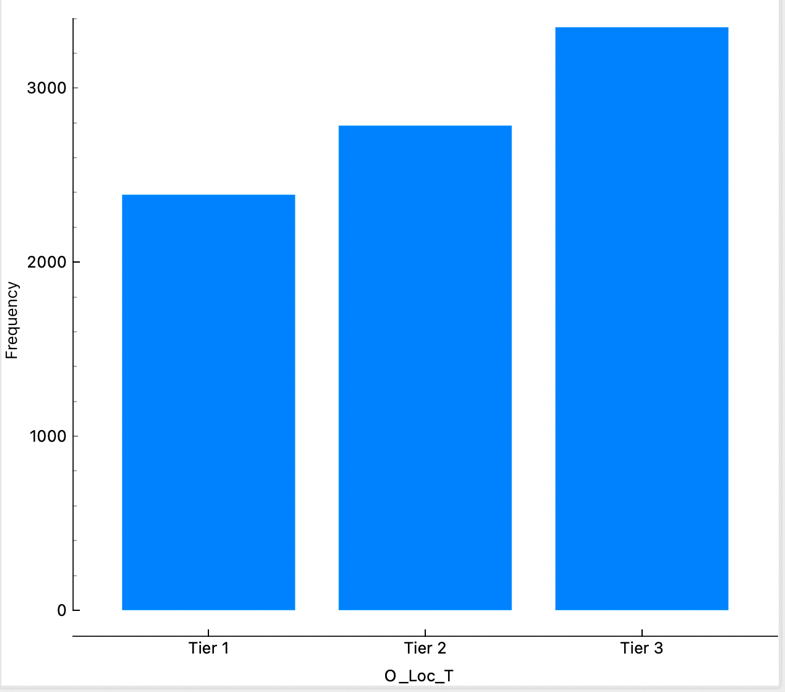
**Starting with the distribution of the categorical values I use this kind of visualization because the are categorical values**

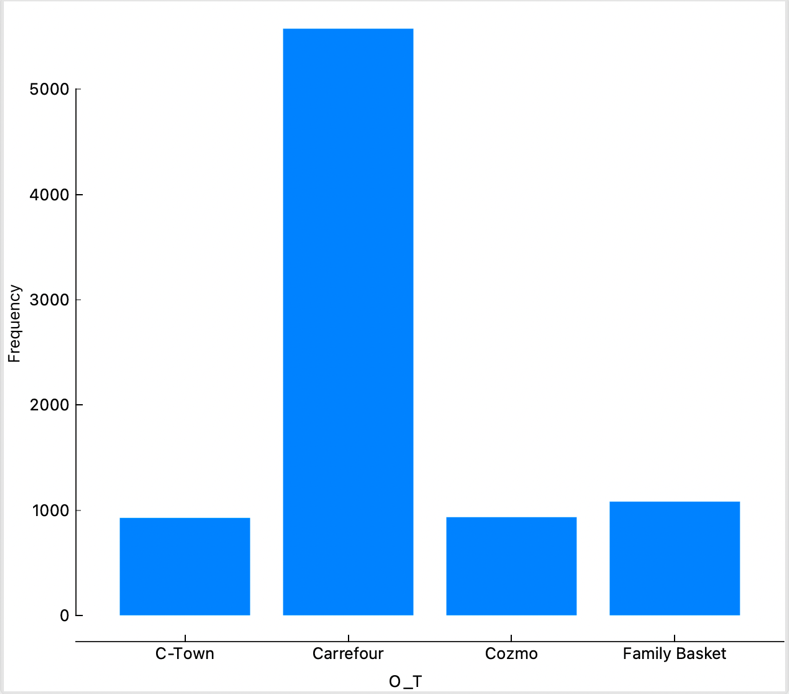






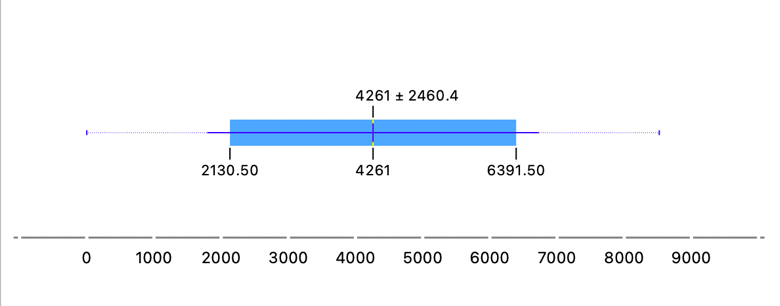




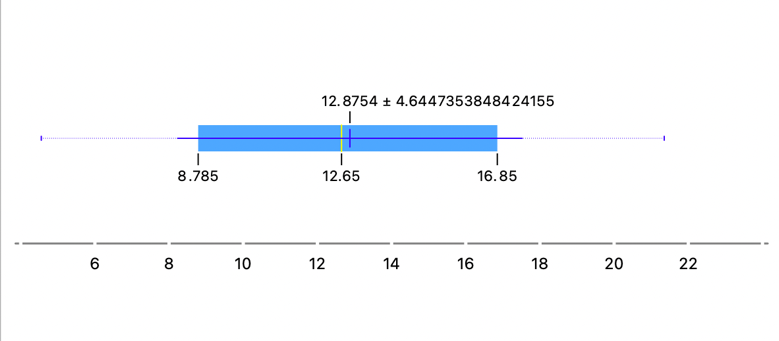


**Moving on to the box plot for the are numerical values values I use this kind of visualization because the are numerical values**

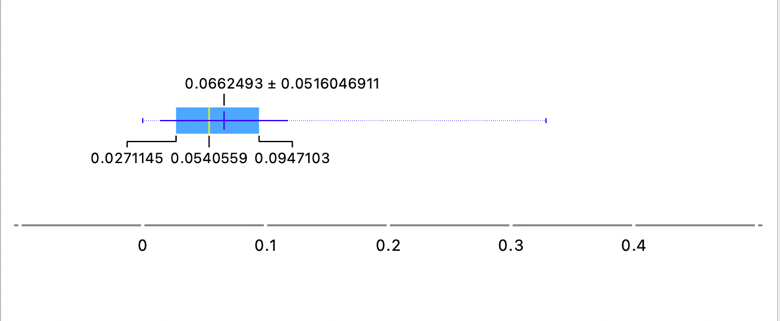
Feature 1



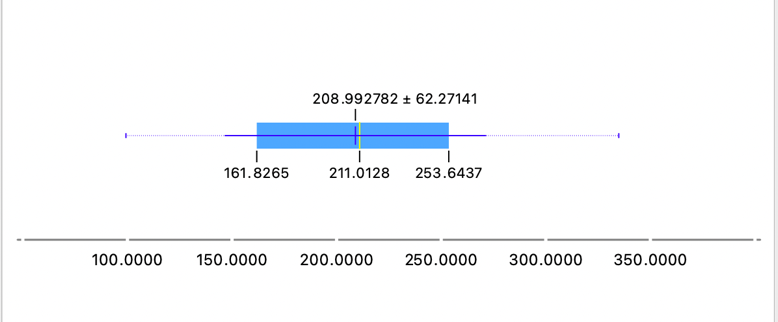
I\_W



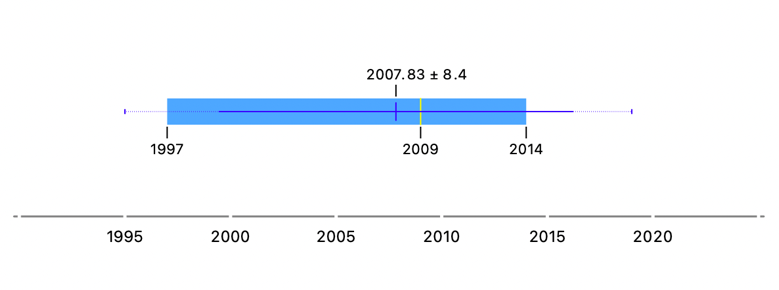
I\_Vis



I\_MRP\_US



O\_Establ\_Y



I\_O\_sales



## data preprocessing

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Step** | **data pre-processing step** | **Column name** | **Description** | **Justification** |
|  | Reading the CSV file |  | read the CSV file and convert it into a DataFrame by this command  (data=pd.read\_csv("dataset.csv")) | So I can see the content of the data, and because dealing with DataFarme is simple and easy by using the NumPy and Pandas library for DataFrames |
|  | Handling missing data | I\_W  (Weight of the product) | I put a piece of code that will fill every null value with its weight in this column based on the first column which is I\_Id (Item\_Id)  First thing this code will put the products that have the same Id in the Same group which means every group will contain just the product with the same id after this it will check the column of the weight and when a null is found the code will check the Id of this null and go to its group and take the mode of the known weight from the group that it belongs to .  A conclusion it is like mapping between tow columns | I filled it in this way because when I put a filter on the data set from A to Z I found that the I\_Id column has a lot of Ids that are the same which means we have a lot of the same product and when I saw the weight I found all the similar Ids have the same weight which makes sense, so I want to fill the nulls with its real weight and we can found it by getting the weight from the know ones.  Because I want the nulls to be filled in a logical way to save in order to maintain accuracy |
| I\_W  (Weight of the product)  (just four items ) | I filled these four nulls with the mean. | Because these four product does not have a similar product with the same Id which means from all the product these four does not Match any other product so the first code cannot fill them because the first code get the weight of the null product from the know similar id product |
| O\_Size  (representing the Size of the store) | I put a code which will help mee to fill the nulls, first the code checks the probability of each value, and then this part of code will convert each probability to a percentage value, then the nulls values will be filled randomly based on the percentage | Because I don’t want fill the nulls with the mode because I don’t want all the nulls to be filled with the same value because it is not logical like this and I want the nulls to be filled with the three values (small, medium, large ) but each value has different count based on its probability |
| **…** | Data cleaning  Also we can consider this step as ( **Handling duplicate data**) | I\_MRP\_JD  (Maximum retail price of the product in Jordanian dinar | I deleted this column | I deleted this column because I don’t think I need it because I have the same values but in dollar and also I have I\_O\_Sales also in dollars but I don’t have a column of the I\_O\_Sales in dinar so in stead of adding a new column for the I\_O\_Sales in dinar in will deleted because I will have 4 column with the same values but in different currency which  means making the data bigger which will effect the efficiency of my code |
| I\_Fat\_C  (whether the product is low fat or not) | I replaced the value “LF” with the value “Low Fat” | Because “LF” is the same as “Low Fat” but they are different spelling.  And I want just one value for the Low Fat not three values but with different spelling |
| I\_Fat\_C  (whether the product is low fat or not) | I replaced the value “low fat” with the value “Low Fat” | Because “low fat” is the same as “Low Fat” but the first one with lower case and the other with upper case, And I want just one value for the Low Fat not three values but with different spelling |
| I\_Fat\_C  (whether the product is low fat or not) | I replaced the value “reg” with the value “Regular” | Because “reg” is the same as “Regular” but they are different spelling. And I want just one value for the Regular not three values but with different spelling |
|  | Data transformation | I\_W  (Weight of the product) | I did a factorize to change the data type in this column from object to float | Because I want to encode all the columns that contains objects data type so I can apply my model but the I\_W values are already encoded and are in a float form but their data type in object and I can’t encode numbers so I have to change their data type to float before doing the encoder |
|  |  | All the column that their data type id object | Encode all the columns that their values data type are objects in to numeric values | KNN ma bst8bel object  Because the models that I want to use do not accept objects values type |
|  | Splitting data | O\_T  (the store from which the product was brought ) | Splitting the data in to two parts Y and X | Y dependant but x not dependant and I want to predict y |

## Justifications for data preprocessing steps

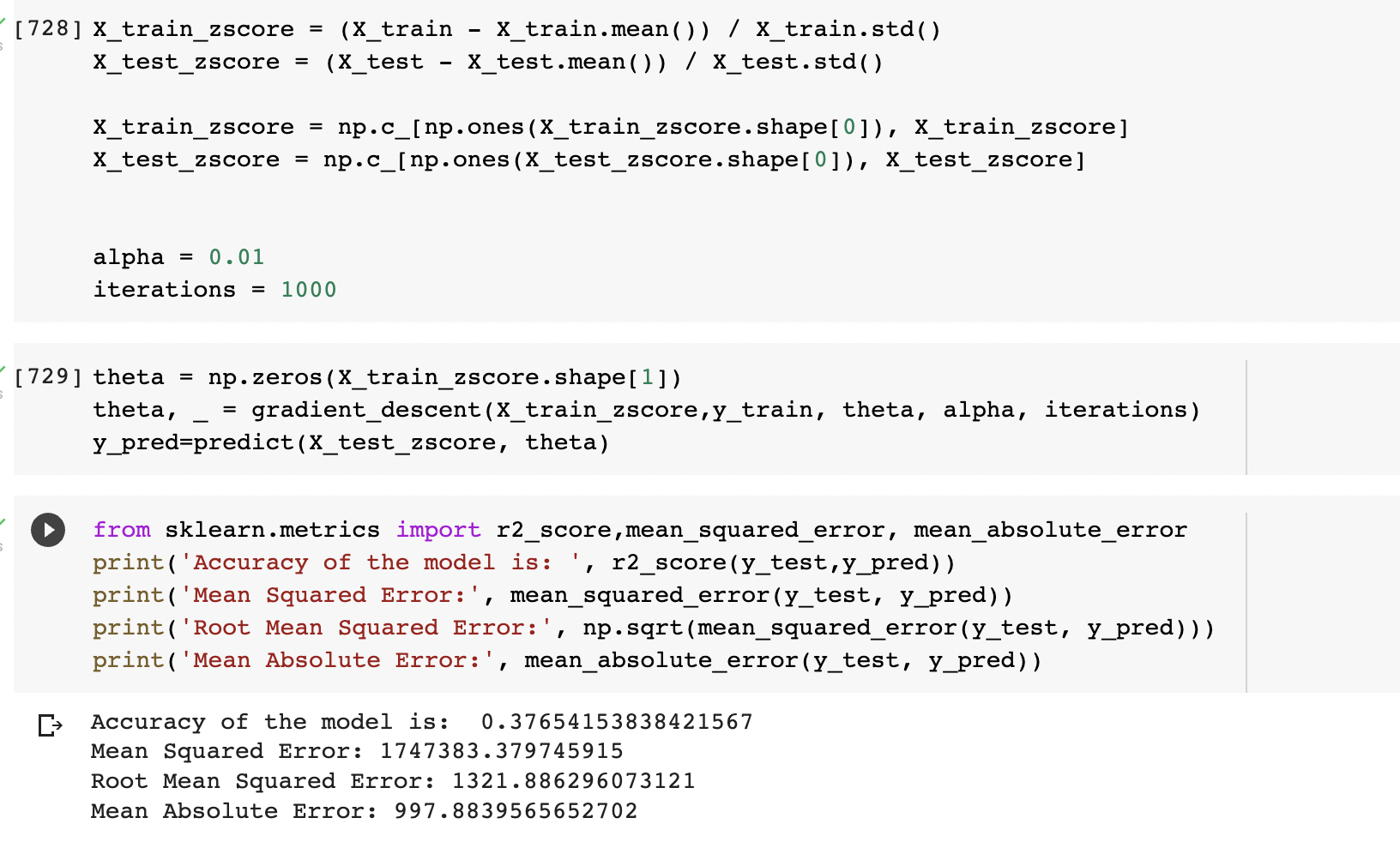
# Experiments

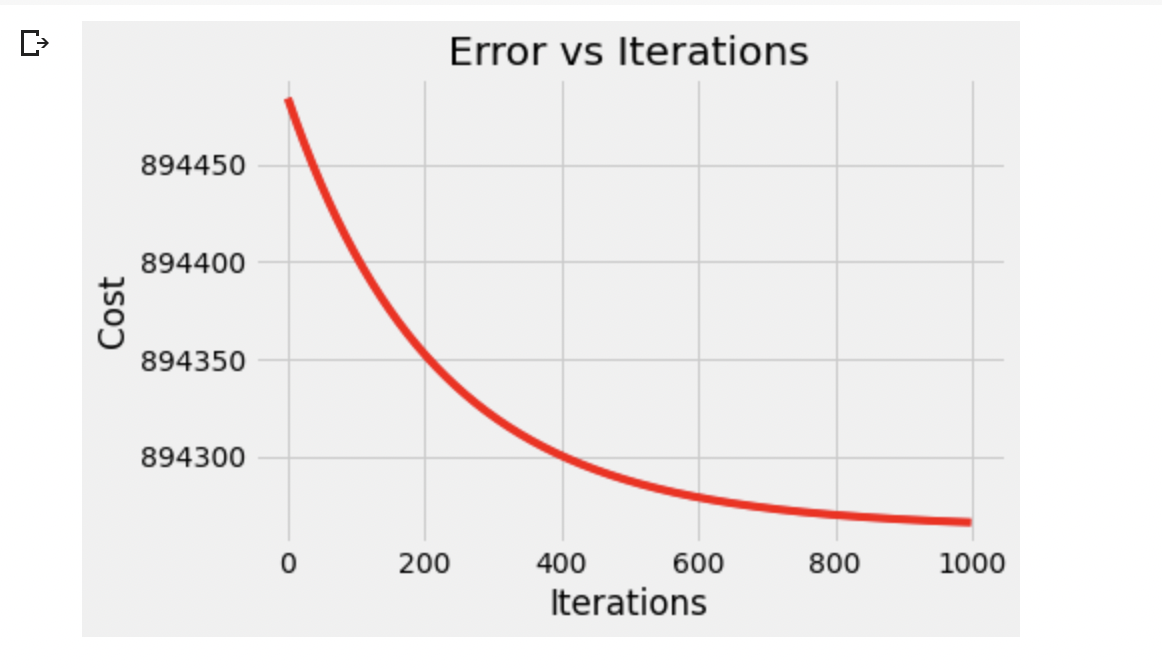
## Linear Regression Model

a statistical technique used to forecast a continuous dependent variable from one or more independent variables. Finding the optimal line that most closely matches the data points is the main goal of linear regression. The equation Y = aX + b, where Y is the dependent variable, X is the independent variable, an is the slope, and b is the y-intercept, is used to depict a line. Finding the values of a and b that reduce the discrepancy between expected and actual values is the objective. There are two types of linear regression: multiple linear regression and basic linear regression (one independent variable) (multiple independent variables). (MATLAB & Simulink, 2022)

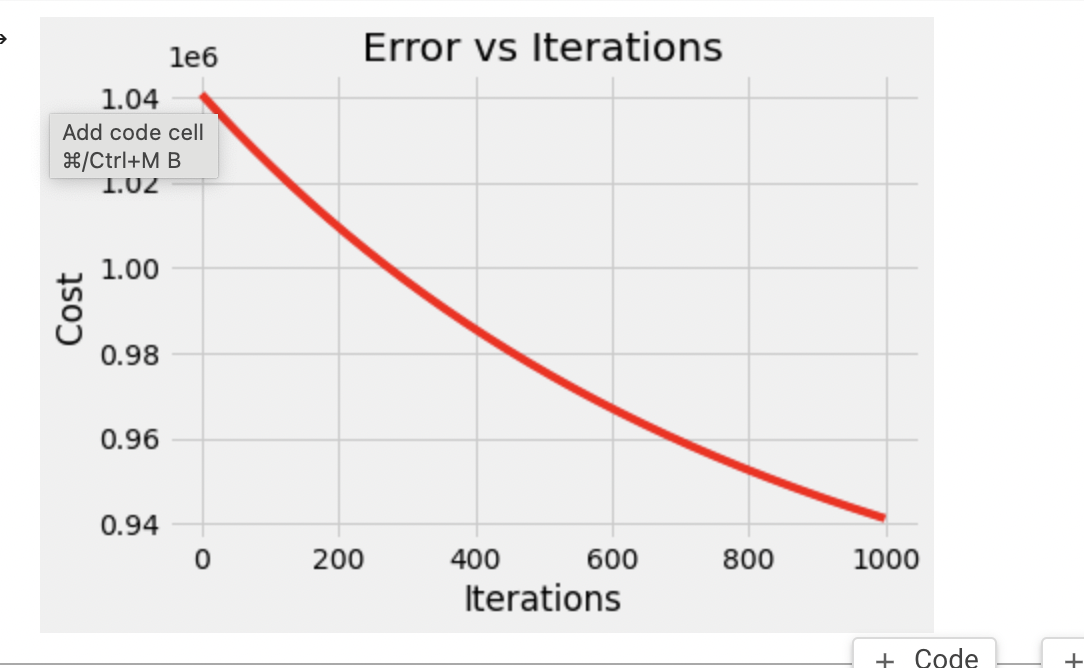
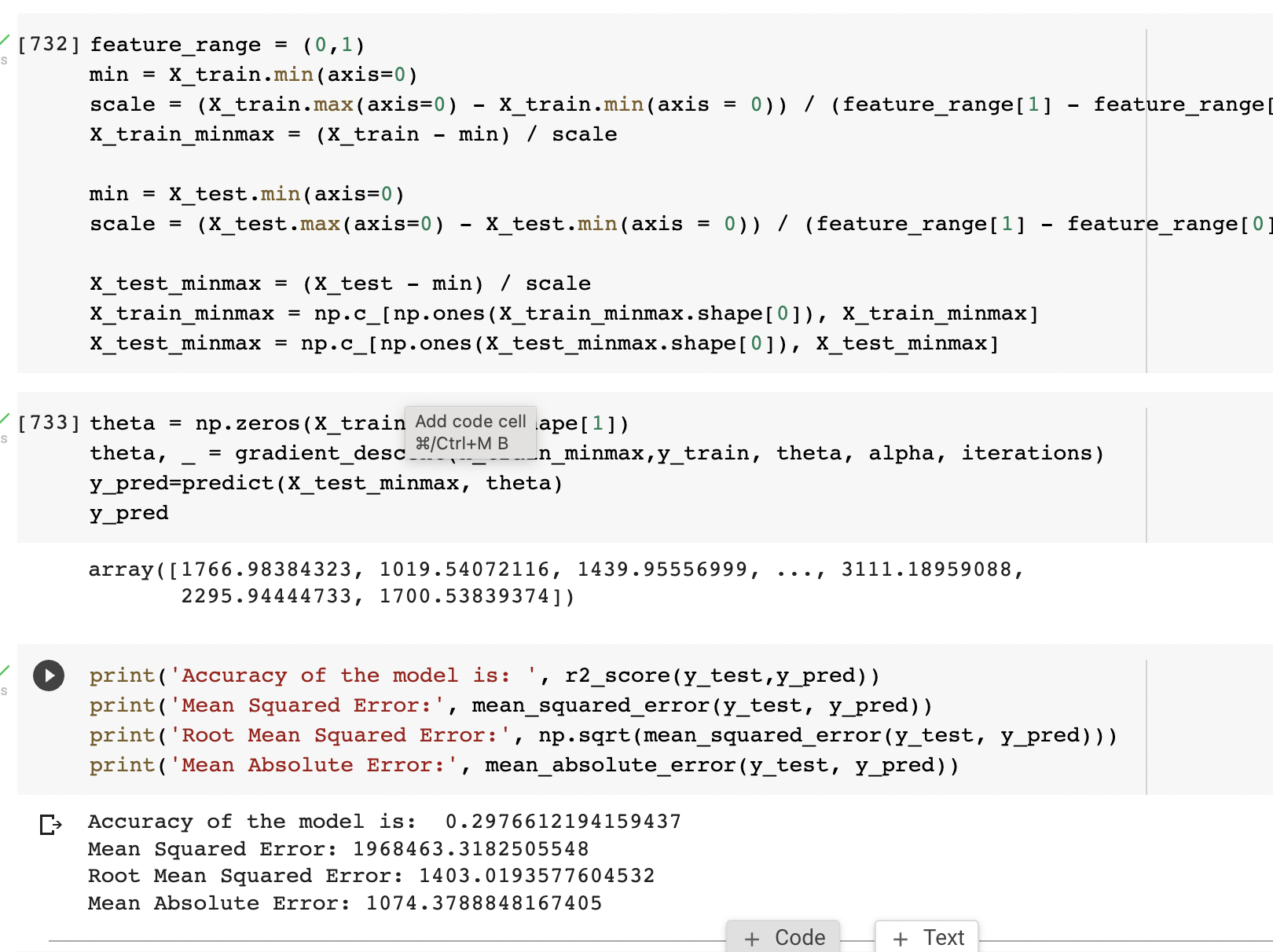
I applied the linear regression code on the data set but with different types of Normalization (Z-score, MinMax , and without Normalization)

Z-score normalization code

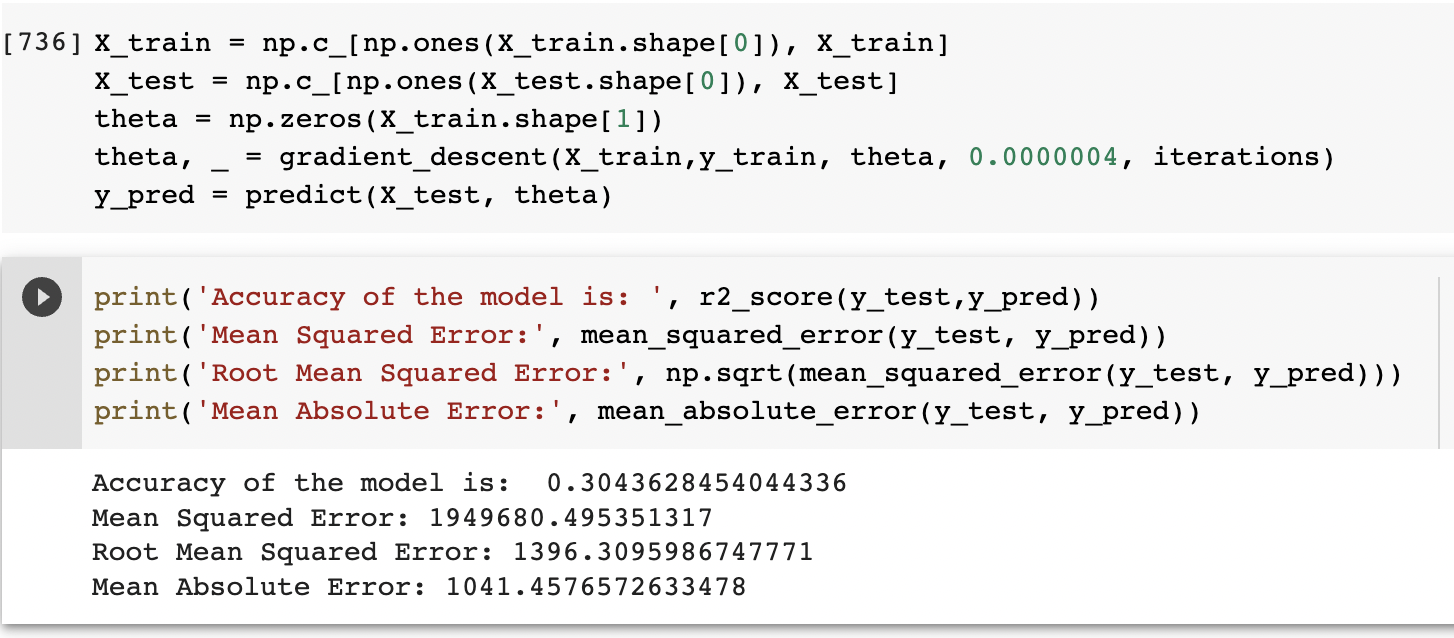


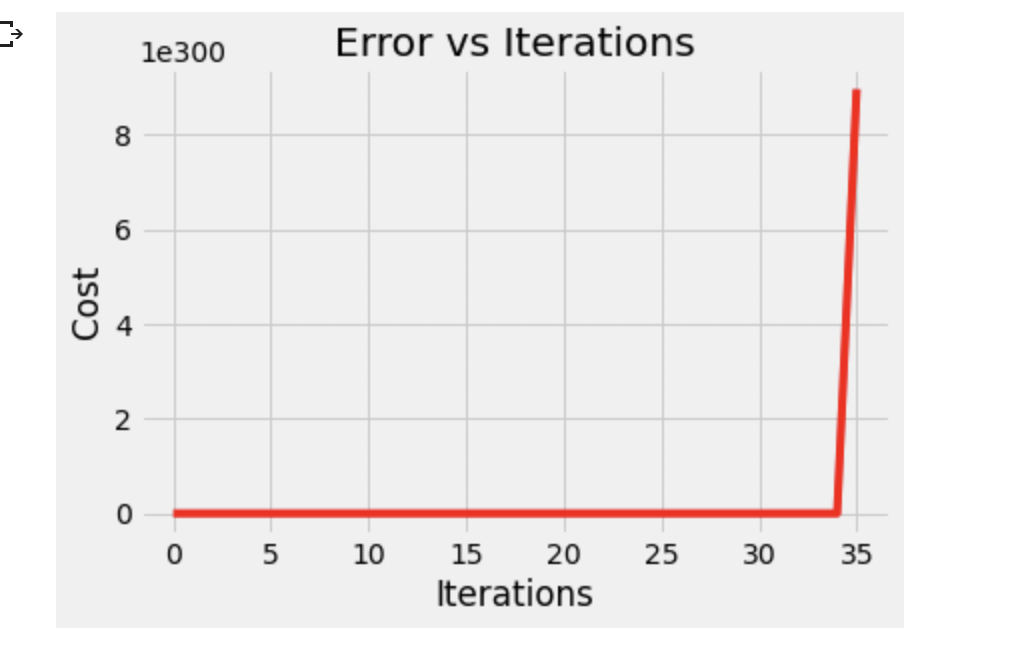


Min Max normalization



Without normalization



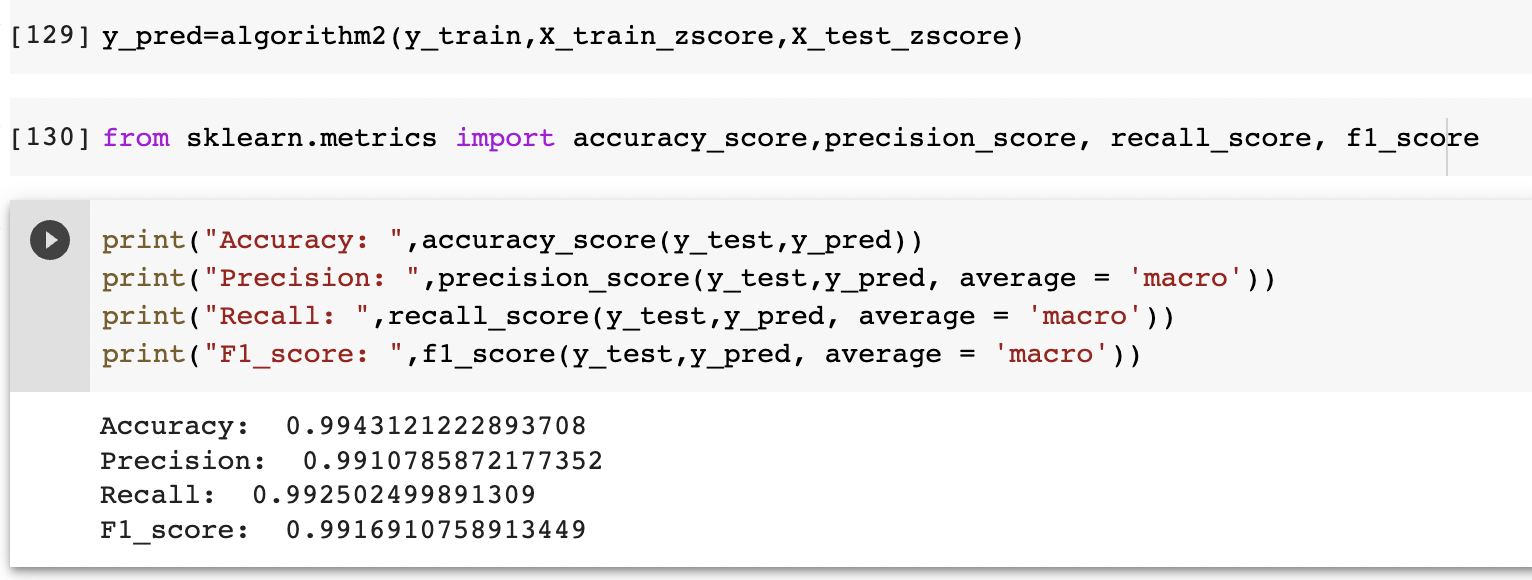


## Classification Model

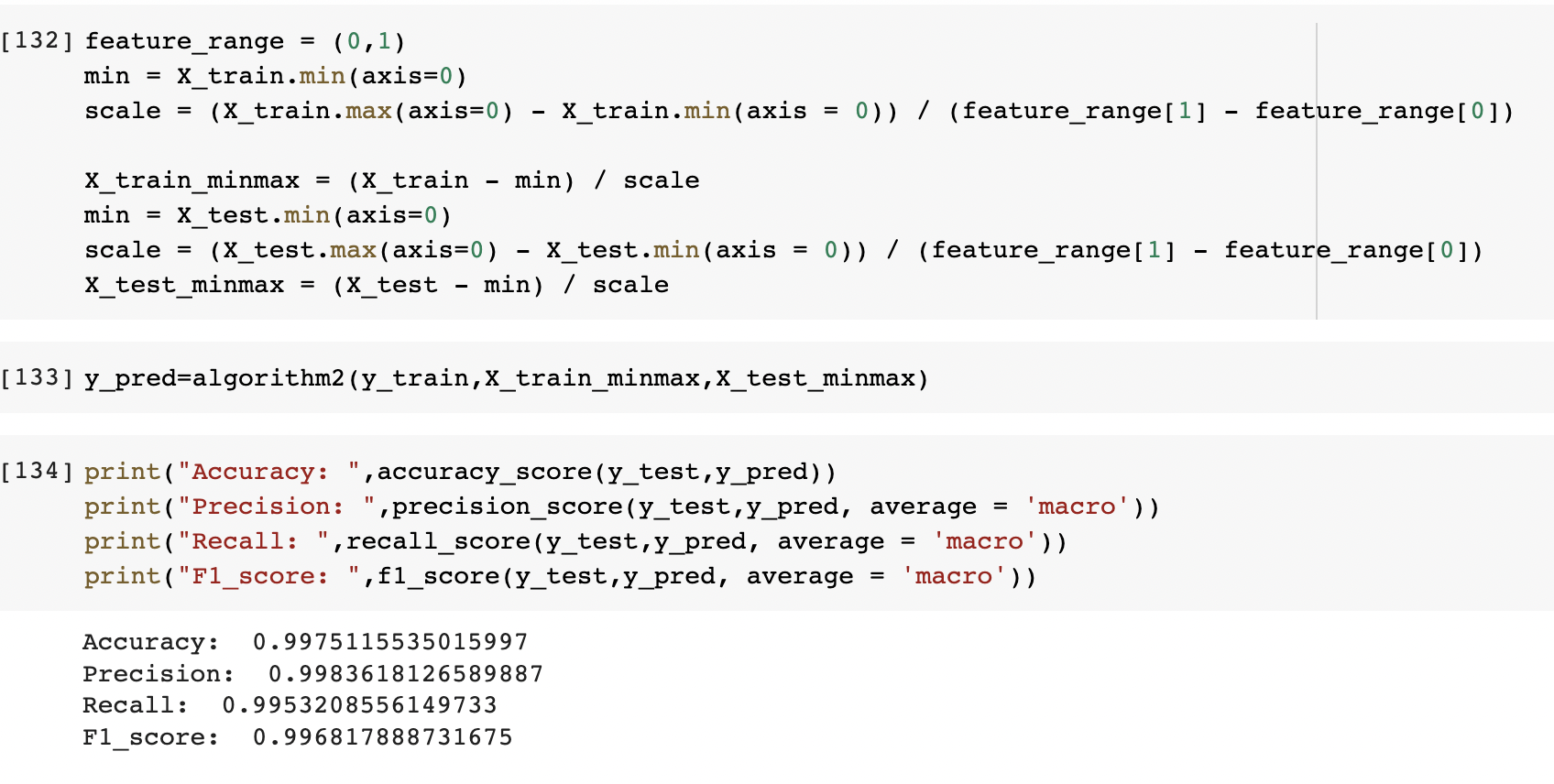
forecast a categorical label (class) based on input characteristics is a particular kind of machine learning issue. Classification models forecast discrete outputs as opposed to continuous outputs, which are forecast via linear regression. Different categories of categorization models exist. (Fanous, 2022)

I applied the classification code on the data set but with different types of Normalization (Z-score, MinMax , and without Normalization)

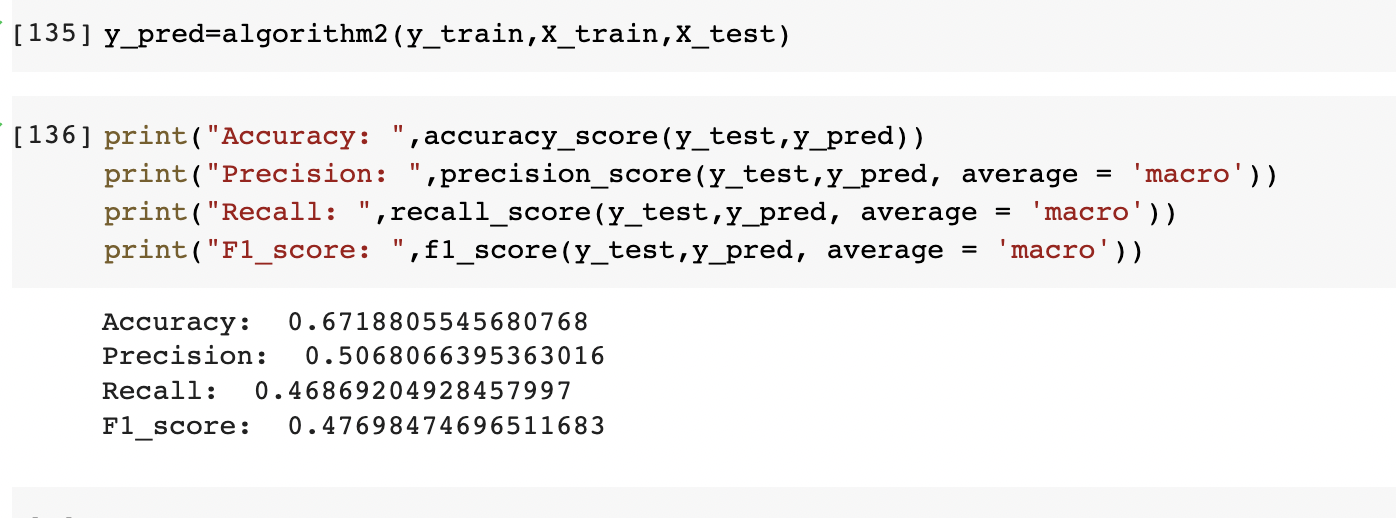
Z-score normalization



Min Max normalization



Without normalization



## Compare the different models

While LR is a parametric model, KNN is a non-parametric model.

In contrast to LR, which can quickly extract output from the tuned coefficients, KNN is sluggish in real time because it must maintain track of all training data and locate neighbor nodes.

LR can function effectively even with minimal training data, whereas neural networks require vast amounts of training data.

NN will move more slowly than LR.

With neural networks, average accuracy will always be higher.

The LR depends on the value of alfa and iterations

The KNN model depends on the value of K

The measures of the KNN model are Accuracy, Precision, Recall, and F1\_score

The measures of the LR model are (R^2 and MSE and RMSE and MAE).

(Varghese, 2019)

* + 1. Linear Regression Model with different learning rate

I will use the Z-score Normalization

iterations = 1000

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Alpha (LR)** | **R2** | **MSE** | **RMSE** | **MAE** |
| **0.01** | 0.3762776147805582 | 1748123.085350934 | 1322.1660581602198 | 994.394468884721 |
| **0.001** | 0.11156357259918359 | 2490044.0731490767 | 1577.9873488558383 | 1135.7578824457198 |
| **0.02** | 0.3784367006478554 | 1742071.7587714829 | 1319.875660345126 | 998.3028996362887 |

iterations = 10000

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Alpha (LR)** | **R2** | **MSE** | **RMSE** | **MAE** |
| **0.01** | 0.3802895582137289 | 1736878.7062828648 | 1317.906941435117 | 996.3886728534595 |
| **0.001** | 0.10611229196344629 | 2505322.520339498 | 1582.821063904413 | 1141.6669143924805 |
| **0.02** | 0.3910906775971329 | 1706606.1257418972 | 1306.3713582828955 | 985.078400192143 |

* + 1. kNN Model with different k values

I will use the Z-score Normalization

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **K value** | **Accuracy** | **Precision** | **Recall** | **F1-Score** |
| **K=3** | 0.9946676146462851 | 0.9912828045846993 | 0.9940520516519067 | 0.9925846434557711 |
| **k=10** | 0.9907571987202275 | 0.9891294329601992 | 0.9875814590489117 | 0.9882326821032602 |
| **K=20** | 9889797369356559 | 0.987184180986098 | 0.9853051087453017 | 0.9860622635431473 |

* + 1. Linear Regression Model with normalization

alpha=0.01

iterations = 1000

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Approach** | **R2** | **MSE** | **RMSE** | **MAE** |
| without normalization | 0.3043528332827793 | 1949708.556587232 | 1396.3196469960708 | 1041.4634611249535 |
| Min-max normalization | 0.28484462498335184 | 2004384.7235650546 | 1415.7629475180704 | 1081.3323420333488 |
| Z score Normalization | 0.3945275234167762 | 1696973.5878364367 | 1302.6793879679055 | 983.1774859572749 |

* + 1. kNN Model with with normalization

k=3

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Approach** | **Accuracy** | **Precision** | **Recall** | **F1-Score** |
| without normalization | 0.6718805545680768 | 0.5068066395363016 | 0.46869204928457997 | 0.47698474696511683 |
| Min-max normalization | 0.9975115535015997 | 0.9983618126589887 | 0.9953208556149733 | 0.996817888731675 |
| Z score Normalization | 0.9943121222893708 | 0.9910785872177352 | 0.992502499891309 | 0.9916910758913449 |

## Analysis of the results.

## The effectiveness of different models

As we can see I have used two models in my code: the KNN and linear regression models

Starting with the linear regression model: I have applied three types of normalization (Z score and Min-Max and without normalization) the I put 4 measures so I can evaluate the model (R^2 and MSE and RMSE and MAE). The first thing I found is that if we take a constant value of iterations and we change the alfa value we can see that when you use a big value of alfa we can get higher accuracy and lower MSE and lower RMSE and lower MAE than others lower values of alfa as we can see in this example the iterations was 1000 when the alfa value was 0.02 the values of the measures were (R^2=0.3784367006478554 / MSE= 1742071.7587714829 / RMSE= 1319.875660345126 / MAE= 998.3028996362887 ) but when the value of alfa was 0.01 the measures were values (R^2=0. 0.3762776147805582 / MSE= 1748123.085350934/ RMSE= 1322.1660581602198/ MAE= 994.394468884721) so we can see that the when the alfa= 0.02 the R^2 was bigger than when alfa=0.01 and the MSE RMSE MAE values were lower than when alfa=0.01 which mean when alfa value is big the results will be better

Also if we change the iterations value we can see that when we use a big value of iterations we can get higher accuracy and lower MSE and lower RMSE and lower MAE than other lower values of iterations. As we can see in this example the iterations =1000 and alfa =0.02 the results of the measures were (R^2=0.3784367006478554 / MSE= 1742071.7587714829 / RMSE= 1319.875660345126 / MAE= 998.3028996362887 ) but when the iterations =10000 and alfa =0.02 the results of the measures were (R^2=0.3910906775971329/ MSE= 1706606.1257418972/ RMSE= 1306.3713582828955/ MAE= 985.078400192143) so we can see that the when the iterations =10000 and alfa=0.02 the R^2 was bigger than when the iterations=1000 alfa=0.02 and the MSE RMSE MAE values were lower than when iterations=1000 which means when iterations value is big the results will be better.

**Moving on to the linear regression with different normalization ways:**

My alfa was = 0.01 and iterations = 1000 I found that the Z-score normalization gave me the highest R2 =(0.3945275234167762) and the lowest MSE= 1696973.5878364367/ RMSE= 1302.6793879679055/ MAE= 983.1774859572749) which mean it is the best and also cost error is 594930.2296042482 which is the lowest cost from others. The without normalization was in the middle between the Z-Score and the Min-Max, If we want to arrange them descending from the best to the worst

Z-score the best > without normalization > Min-Max so we can see from the results in the table that the Z-score normalization type gave mee the best results.

This is cost error for the Min-Max type is 627523.9150459112 which is bigger than the Z-score

Moving on to the KNN model I found that the smaller K value, the higher Accuracy, Precision, Recall, and F1\_score. When the k= 3 the (Accuracy= 0.9946676146462851 /Precision=0.9912828045846993 / recall= 0.9940520516519067/ F1\_score=.0.9925846434557711)

but when K was =10 the (Accuracy= 0.9907571987202275/Precision=0.9891294329601992/ recall= 0.9875814590489117 / F1\_score=.0.9882326821032602)

so we can say that when K value is small the results will be better.

**Moving on to the KNN with different normalization ways:**

My KNN was =3 I found that KNN model with Min-Max normalization gave me the highest values for the measures (Accuracy=0.9975115535015997/Precision=0.9983618126589887/ recall= 0.9953208556149733/ F1\_score=0.996817888731675)

And the Z-score came after it the (Accuracy= 0.9946676146462851 /Precision=0.9912828045846993 / recall= 0.9940520516519067/ F1\_score=.0.9925846434557711) and I got the least results from the model without normalization (Accuracy=0.6718805545680768/Precision=0.5068066395363016/ recall= 0.46869204928457997/ F1\_score=0.47698474696511683)

which means that the Min-Max normalization is the best, The Z-score normalization was in the middle between the Min-Max and the without normalization, If we want to arrange them descending from the best to the worst

Min-Max the best > Z-score > without normalization so we can see from the results in the table that the Min-Max normalization type gave me the best results

The effectiveness of KNN we can see that the least value of k gives me the best results so I will use the KNN with small values of K plus I will use the Min-Max normalization on the data set when I use the KNN model because it will give me the highest results, not like the other types of normalization which is what we need so it will be so effective KNN is sluggish in real time because it must maintain track of all training data and locate neighbor nodes.

The effectiveness of linear regression we can see that when the alfa is a large value we get the best results and when the iterations are also high number also we will get high results also I will use the Z-Score normalization because the error cost is less than the other cost of the other normalization types also because the run time for it is less than others and because the Z-score normalization gives mee the highest results, which is what we want.

So both models are effective so much but depend on how we use it

# **Introduction**

## **Type of Machine Learning**

alteryx (2022) *Supervised vs. unsupervised learning; which is best?*, *Alteryx*. Available at: https://www.alteryx.com/glossary/supervised-vs-unsupervised-learning#:~:text=Supervised%20and%20unsupervised%20learning%20have,tagged%20with%20the%20right%20answer.&text=A%20classification%20problem%20uses%20algorithms%20to%20classify%20data%20into%20particular%20segments. (Accessed: January 22, 2023).

Naukri (2022) *Naukri.com*. Available at: https://www.naukri.com/learning/articles/supervised-learning-real-life-analogy-and-applications/ (Accessed: January 23, 2023).

BasuMallick, C. (2022) *AWS vs. azure: Understanding the key differences*, *Differences Between AWS and Azure*. Available at: https://www.spiceworks.com/tech/cloud/articles/aws-vs-azure/ (Accessed: January 23, 2023).

Kinsta (2022) *AWS vs azure: Comparing the cloud computing giants*, *Kinsta®*. Available at: https://kinsta.com/blog/aws-vs-azure/ (Accessed: January 23, 2023).

amazon (2020) *Improve the CPU performance of an Amazon EC2 linux instance*. Available at: https://aws.amazon.com/premiumsupport/knowledge-center/ec2-linux-performance/ (Accessed: January 23, 2023).

Alpaydin, E. (2021) *Machine learning*, *Amazon*. The MIT Press. Available at: https://aws.amazon.com/machine-learning/accelerate-machine-learning-P3/ (Accessed: January 23, 2023).

alteryx (2022) *Supervised vs. unsupervised learning; which is best?*, *Alteryx*. Available at: https://www.alteryx.com/glossary/supervised-vs-unsupervised-learning#:~:text=Supervised%20and%20unsupervised%20learning%20have,tagged%20with%20the%20right%20answer.&text=A%20classification%20problem%20uses%20algorithms%20to%20classify%20data%20into%20particular%20segments. (Accessed: January 22, 2023).

About Sakshi Gupta Sakshi is a Senior Associate Editor at Springboard. She is a technology enthusiast who loves to read and write about emerging tech. She is a content marketer and has experience working in the Indian and US markets. (2022) *Regression vs. classification in machine learning: What's the difference?*, *Springboard Blog*. Available at: https://www.springboard.com/blog/data-science/regression-vs-classification/#:~:text=The%20most%20significant%20difference%20between,types%20of%20machine%20learning%20algorithms. (Accessed: January 22, 2023).

www.javatpoint.com. (n.d.). *Regression vs Classification in Machine Learning - Javatpoint*. [online] Available at: https://www.javatpoint.com/regression-vs-classification-in-machine-learning.

‌ *Regression vs Classification in Machine Learning - Javatpoint*. (n.d.). www.javatpoint.com. <https://www.javatpoint.com/regression-vs-classification-in-machine-learning>

MATLAB & Simulink (2022) *Linearmodel*, *What Is a Linear Regression Model? - MATLAB & Simulink*. Available at: https://www.mathworks.com/help/stats/what-is-linear-regression.html (Accessed: January 30, 2023).

Fanous, K. (2022) *Classification models and Thresholds*, *Medium*. Towards Data Science. Available at: https://towardsdatascience.com/classification-models-and-thresholds-97821aa5760f (Accessed: January 30, 2023).

Varghese, D. (2019) *Comparative study on classic machine learning algorithms*, *Medium*. Towards Data Science. Available at: https://towardsdatascience.com/comparative-study-on-classic-machine-learning-algorithms-24f9ff6ab222#:~:text=KNN%20is%20a%20non%20%2Dparametric,from%20the%20tuned%20%CE%B8%20coefficients. (Accessed: January 30, 2023).