**Principles of data science and computing systems**

**Final report**

**Sara Aljamal**

**Section 4**

Contents

[Technical Documentation 3](#_Toc126012479)

[1 Introduction 3](#_Toc126012480)

[1.1 Type of Machine Learning 3](#_Toc126012481)

[1.2 Supervised Machine Learning 4](#_Toc126012482)

[1.3 Computing systems 4](#_Toc126012483)

[2 Methodology 5](#_Toc126012484)

[2.1 data science life cycle 5](#_Toc126012485)

[2.2 data pre-processing 9](#_Toc126012486)

[2.3 Justifications for data pre-processing steps 9](#_Toc126012487)

[3 Experiments 13](#_Toc126012488)

[3.1 Linear Regression Model 13](#_Toc126012489)

[3.2 Classification Model 13](#_Toc126012490)

[3.3 Compare the different models. 13](#_Toc126012491)

[3.4 Analysis of the results 14](#_Toc126012492)

[3.5 The effectiveness of different models 15](#_Toc126012493)

[4 References 17](#_Toc126012494)

# Technical Documentation

# Introduction

## Type of Machine Learning

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

|  |  |  |
| --- | --- | --- |
|  | **Supervised Learning** | **Unsupervised Learning** |
| **Definition** | It’s that type of ML that works with labeled data (input data), train on the input data, and classify or make predictions on Unseen data (Test data). | It's that type of ML that works with unlabeled data to find pattens and groups data according to some features. |
| **Applications (use)** | 1. Bioinformatics 2. Spam detections | 1. Natural language processing 2. Anomaly detection |
| **Strengths** | 1. Simple process to understand, unlike the unsupervised learning. 2. We can specify classes and know the exact number of classes. | 1. Doesn’t require to split the data into training and testing. 2. Doesn’t require labeled data. |
| **Limitations** | 1. Supervised learning doesn’t discover relations. 2. Training takes a lot of time. | 1. Less accurate than supervised learning. 2. The results may be ineffective as it doesn’t rely on a clear label. |
| **Common Algorithms** | 1. Linear regression 2. Decision trees | 1. K- means. 2. Hierarchical cluster analysis |

## Supervised Machine Learning

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

|  |  |  |
| --- | --- | --- |
|  | **Classification** | **Regression** |
| **Type of Learning** | Supervised learning | Supervised learning |
| **Applications (use)** | 1. Image classification 2. Customer segmentation | 1. Time series forecasting 2. Predictive modelling |
| **Strengths** | 1. Deals with non- linear relations. 2. Takes huge datasets as input. | 1. The output is easy to understand as it is an equation. 2. It can understand and define the relations between the variables in a good way. |
| **Limitations** | 1. Some models can be complex to implement and understand. 2. It makes the assumption that features are independent. | 1. Sensitive to outliers 2. Doesn’t work properly if the number of variables is large. |
| **Common Algorithms** | 1. KNN 2. Naive Bayes | 1. Linear regression 2. Support Vector Regression |

## 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:

A computing system is a combination of hardware and software to achieve a goal, so to run an application we need the program (software), and the needed hardware such as input and output devices, CPU, and other things. Computing systems are used in cloud computing services. Cloud computing is a term that means having access through the internet to computing resources, applications, servers, data storage and a lot more. These services are provided to customers or users on demand, which means it's available any time. These services may be free or paid according to the service itself and the service provider. Examples of cloud services are AWS and Azure. Both AWS and Azure share the same structure, which is having different datacenters in different locations around the world. This is considered a strength for both services as they can avoid any failures they may face when having only one datacenter.

**AWS:**

**GPU:** AWS is powered by NVIDIA GPUs

**CPU:** AWS offers so many CPU-optimized EC2 instance families, including the C5, R5, M5, and T3 instance families.

**Type of storage devices used:**

* Glacier is a limited, long-term data archiving solution.
* Storage Gateway: A hybrid storage suggestion that combines on-premises and AWS cloud storage.

**Related hardware:**

 EC2 (Elastic Compute Cloud) instances

 S3 (Simple Storage Service)

 EBS (Elastic Block Store) volumes

**Azure:**

**GPU:** Azure's needs are fulfilled using the NVIDIA Tesla K80 GPU.

**CPU:** Azures use of a variety of CPU types. Such as: Intel Xeon processors, AMD Epyc, ARM processors

**Type of storage devices used:**

* Azure Archive Storage: A long-term, moderate solution for data storage.
* StorSimple: A blended cloud storage solution that combines on-premises and Azure storage.

**Related hardware:**

* Azure load balancer
* Azure Disks: provides virtual machines with continual data storage.
* Azure ExpressRoute: offers a secure connection to Azure.

# Methodology

## data science life cycle

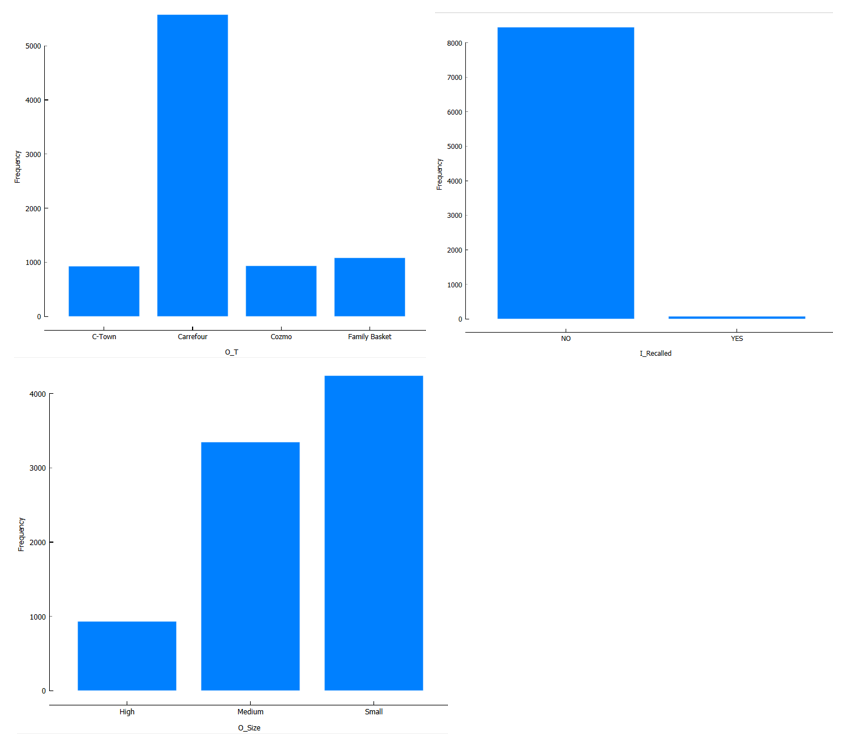
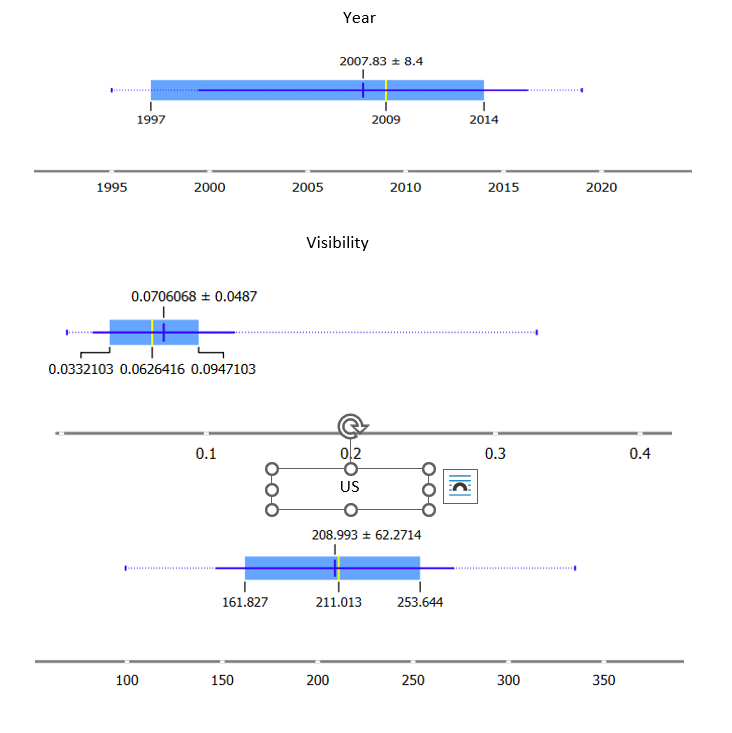
* + Discuss and explain the data science life cycle **of your developed prediction model** (any of them).

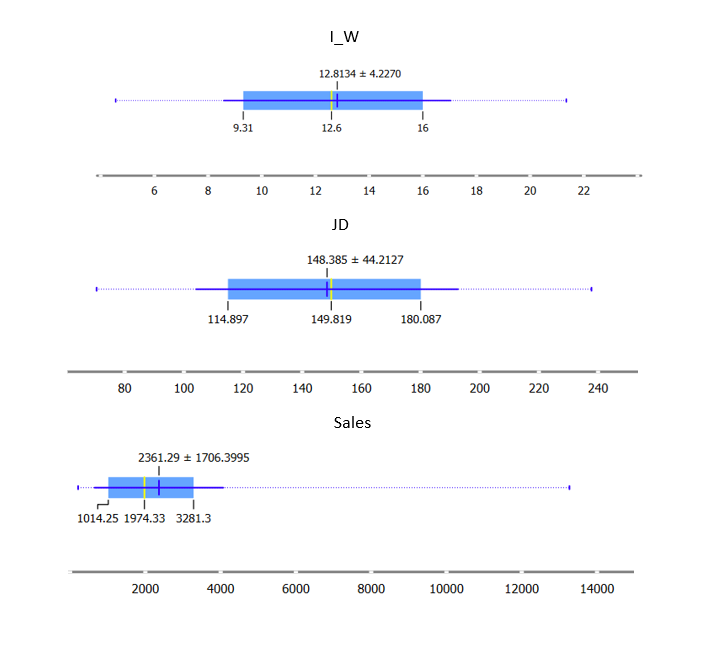
|  |  |
| --- | --- |
| **The Data Life Cycle Stage** | **The importance of the given Stage toward your project lifecycle** |
| 1. Understanding the problem | Understanding the problem we need to solve, is important to define and set the plan that our project will follow, it gives us the rules and the limitations we should take into consideration in our project. In our project, we were required to build two ML models, one for predicting the sales and one for predicting from which store the product was bought. This stage is also important because all other stages rely on the objectives defined in this stage. |
| 1. Gathering Data | Collecting data related to our problem from different resources is important, in our project we used data in a form of CSV file. Normally gathering data is important to have enough data that support our problem and can be used in training and predictions. |
| 1. Cleaning Data | This stage is about cleaning the data and pre process it, by this I mean replacing null values, normalizing data, a lot more. This stage is important because if the data wasn’t cleaned and prepared in a good way, it will affect the prediction of our model. So if we have bad prediction results, we have to consider the pre processing stage and sometimes change on it. |
| 1. Exploring Data | After cleaning data, we need to understand it, and know every column type, so categorical columns need encoding. Also, this stage is important because when we understand the data, we can uncover hidden relations and corelations. One of the ways to understand the data is presenting every column value using histograms to see the values and their frequency, this was very helpful in our project, the histograms for the columns can be found under the table. |
| 1. Feature Engineering | This stage is important as we take the information and relations found in the previous stage to determine which column should be the model’s input and which columns to drop depending on its relation with our target column. It’s also important because we leave only the important features, so the model prediction won’t be affected by features that are not important. Also, in this stage we remove highly correlated features as they can be considered as duplicated features. |
| 1. Modelling Data | This stage is very important to our project as it’s the goal of our project to build models and predict. In our project we used two modelling algorithms, KNN and linear regression. In addition, in this stage we get the results of our prediction such as the accuracy and try to decrease the model drift that may happen to our models |
| 1. Interpreting Data | This stage can be called data visualization, which is transferring the results into forms that anyone can understand. This stage is important because data scientists are required to explain the results to normal people in the company or the place they are working for, so they can’t understand results as codes but they can understand them using histograms, box plot, and bar charts. |

**-Features:**

Chart, bar chart

Description automatically generated**The categorical features were presented using histograms, while the numeric features were presented using box plots.**





## data pre-processing

## Justifications for data pre-processing steps

|  |  |  |  |
| --- | --- | --- | --- |
| **Step** | **data pre-processing step** | **Description** | **Justification** |
|  | Correlation matrix | Using the function corr () to find the correlation between columns | I started the pre-processing with this step because it defines and clarify the correlation between columns, so it helped me deciding what to do with every column |
|  | Dropping columns | Removing the unwanted columns, such as: I\_Id, I\_MRP\_JD, I\_category, O\_Id, I\_Recalled | According to the correlation I found using the correlation matrix, I decided to delete this column, also some of them are not important such as the ID, which has a lot of unique vales and it’s a column that needs one hot encoder so if I kept it then I would have thousands of columns. The JD column was dropped because we have the price in the us dollar and the sales are also using us dollar, so it wasn’t important to keep. For the category and recalled columns I dropped them because the correlation was low between them and the two columns we were supposed to work on. |
|  | Checking duplicated | I used the duplicated function to check if there is any duplicated values and drop them if any. | Checking for duplicates is important because having duplicated rows can cause problems with the prediction. |
|  | Using describe function | It’s a function that shows statistical information about the numeric columns in our dataset such as the min and max, mean, and the quartiles. | Describe function helps us understand the numeric columns we have. Besides, if there are any values =0 that they shouldn’t =0 then we would know using this function, this happened in the I\_Vis column. |
|  | Replacing 0 with null in some columns | If there are some values that =0 that doesn’t make sense then I should change them to null values, such as the values in I\_Vis columns. | As mentioned, 0 doesn’t make sense in this case, because I\_Vis is a column that describes display area for the product. |
|  | Searching the null values | I used isnull.sum () function to check the number of null values every column has. | Handling missing values is one of the most important steps in data pre processing as we fill the null values with appropriate values. In order to fill the null values in a good way, we need to see how many null values we have in each column to determine the right method. |
|  | Filling null values | After knowing which columns has null values and the number of them. We should fill these null values with good values and not random values. The column that had null values were (I\_W, I\_Vis, O\_Size) | As mentioned earlier, filling these values is important for the prediction, so I filled the I\_W column with the median, which is the value in the middle, the I\_Vis column with the mean, and the sales column with the mode but after finding relation between it and the size column. |
|  | Ensuring data integrity | In this step we ensure that the values in some columns are written in the same way, for example, the\_Fat\_C column had 2 possible values, regular and low fat, but it had several values that represents the same value, so I changed the ’LF’ and ‘low fat’ to ‘Low fat’ and ‘reg’ to ‘Regular’ | Having different unique values for the same value isn’t efficient and affects our model’s prediction, so it’s better to ensure the integrity of the data and change them. |
|  | Data transformation | This step is about encoding the categorical columns, this means transforming these columns to numeric values. These is two types for encoding, the first one using label encoder and the second one using one hot encoder. | The main reason for transforming data is that ML models doesn’t take categorical values so we need to change them. We use label encoder for the ordinal values, and we use one hot encoder for other values. |
|  | Normalization | Normalizing data means to normalize the numeric values to a specific scale, this is done before encoding. There is two types for normalization that I used, min max normalization which normalize all values in the column in the range (0,1), and z score normalization which normalize all values by (value – mean/standard deviation) | Normalization is important to give equal chance for every column to effect on the prediction. I used two types for normalization to compare them and see what the results are when using each one of them. |

# Experiments

## Linear Regression Model

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Alpha (LR)** | **R2** | **MSE** | **RMSE** | **MAE** |
| **0.1** | 0.5622054801092438 | 1333120.3241386137 | 1154.608299008202 | 862.6229402975844 |
| **0.001** | 0.5329292997288514 | 1422268.7015280519 | 1192.5890748820618 | 893.8721181407125 |
| **0.0001** | 0.19047290406307926 | 2465076.595302546 | 1570.0562395349239 | 1218.750227055512 |

## Classification Model

## Compare the different models.

* + 1. Linear Regression Model with different learning rate

(With min-max normalization)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **K value** | **Accuracy** | **Precision** | **Recall** | **F1-Score** |
| **1** | 0.7519061583577713 | 0.7519061583577713 | 0.7519061583577713 | 0.7519061583577713 |
| **3** | 0.7653958944281525 | 0.7653958944281525 | 0.7653958944281525 | 0.7653958944281525 |
| **7** | 0.7788856304985338 | 0.7788856304985338 | 0.7788856304985338 | 0.7788856304985338 |

* + 1. KNN Model with different k values

(Without normalization)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Approach** | **R2** | **MSE** | **RMSE** | **MAE** |
| without normalization | 1.0 | nan | nan | nan |
| Min-max normalization | 0.5622054801092438 | 1333120.3241386137 | 1154.608299008202 | 862.6229402975844 |
| Z score Normalization | 0.5622054801103229 | 1333120.3241353272 | 1333120.3241353272 | 862.6229402965713 |

* + 1. Linear Regression Model with normalization

(Learning rate= 0.1, number of iterations= 10000)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Approach** | **Accuracy** | **Precision** | **Recall** | **F1-Score** |
| without normalization | 0.7653958944281525 | 0.7653958944281525 | 0.7653958944281525 | 0.7653958944281525 |
| Min-max normalization | 1.0 | 1.0 | 1.0 | 1.0 |
| Z score Normalization | 1.0 | 1.0 | 1.0 | 1.0 |

* + 1. KNN Model with normalization

(K=3)

## Analysis of the results

In this part, we will analyse the results from the four tables above in this section. It is important to analyse the results so that we can better understand the results, how our model works, and how we can improve it. I used two machine learning models for this project: linear regression to predict the sales column and KNN to predict the I T column, which is the store from which the product was purchased. Each model's evaluation measures differ, so in linear regression, we have R^2, MSE, RMSE, and MAE. In KNN, we can evaluate the model's effectiveness using the accuracy, precision, recall, and F1 score. The results are shown in the four tables above for:

1. Linear regression with minmax normalization using different learning rates.
2. KNN without normalization using different K values.
3. Linear regression with alpha (learning rate) =0.1, but without normalization, with minmax normalization, and with z score normalization.
4. KNN with the same K=3, but without normalization, with minmax normalization, and with z score normalization.
5. In the first table, I used different learning rates for the linear regression with minmax normalization. The learning is a parameter that determines the step size for the iterations in the gradient decent. I compared the difference in R^2 which measures how the values are close to the line, MSE which is the mean squared error, it finds the error (the difference between the actual and predicted values) , RMSE which is the root of the MSE, and MAE which is the mean average error. We can realize from the numbers in the table that when alpha is bigger, R^2 value increases, and other measures such as the MSE decreases, this indicates that using bigger value for alpha improves the results of our models’ prediction. In this case, the best value for alpha between the three values was 0.1. Although using bigger value for alpha was better in this case, using a large value for alpha isn’t efficient in all cases.
6. In the second table, which represents the difference between the accuracy measures for the KNN such as accuracy, precision, recall, and F1 score when we change the K value, the K value specifies how many neighbours will be tested in order to classify a specific point. From the table above, when K=1, the accuracy measures were the lowest in the table, that is because the model will take the nearest point to the point we want to classify and put it in the same class with that point, but sometimes this one point may be noise in the data so it gives us wrong classification, but also this doesn’t mean we should take a large value for K such as 19, because it will also affect the accuracy for our model in a negative way. So to solve this problem, we choose neither too big nor small value for K, such as 3 or 7, in the table above, 7 showed better efficiency as the accuracy, precision, and other measures were the best values in the tables, but it’s not a rule or standard to follow, so choosing K value depends on the case.
7. The third table is about the normalization and how it affects our model’s predictions. Normalization is when we normalize all the numeric values in our dataset (before encoding) according to one scale to avoid the problem where the prediction depends on one feature only as it has big numbers and not giving the features with small number any importance. There are two ways for normalization, z score normalization where we normalize all values according to the z score (value- mean/ standard deviation), and minmax normalization where we normalize all values to the range (0,1) depending on the max and min values in every features. When I used the dataset without normalization in the linear regression algorithm, the results were very bad, R^2 was 1 but I believe it’s a fake value because all other values were nulls and this doesn’t make sense at all. But when using normalization, the results were good even though both normalization methods gave us results that are very close to each other.
8. The fourth table was like the third one but for the KNN model, so the accuracy measures changed also, but the comparison was the same, comparing results without normalization, with minmax normalization, and with z score normalizations. Here using the dataset without normalization gave us good results comparing with the linear regression. Also, the minmax and z score gave us the same results which was 1 for everything.

## The effectiveness of different models

**- Linear regression:**

We were required to use the linear regression to predict the sales of the product, which is a continuous value. While using the model, I changes the learning rate parameter, in addition to changing the data from minmax normalization to z score normalization and etc.. as shown in the tables above. All these changes affected in a clear way to the prediction of our model, some of them effected positively while others didn’t, this shows us the importance od understanding the main components and the things that affects our model so we know how we should use them. It’s also importance as trying these changes increases our model’s effectiveness as well as its accuracy. But if I want to talk about the effectiveness of the linear regression in general in our case, I would say it’s a good option but it’s not the best, so we may have other options that might suite our case better, because the best accuracy (R^2) I could get using this model was 0.56 which is a percent that can be improved if we use other models. Another reason why we may have to use another model is that the process where we specify alpha and the number of iteration takes times and sometimes may be hard.

**- KNN:**

KNN model was used to predict the store from which the product was purchased, which is considered as classification. KNN is knows as a lazy algorithm as it has no training time, but in fact, KNN performance in our case was great and the results were good. The main parameter that KNN relies on is K value, which means changing it will contribute a lot to our prediction as we can see in the table above, so changing the value of K affects the effectiveness of our model. Another thing to be taken into consideration is the normalization for the dataset, which is proved in the fourth table, normalization is very important stage specially in the KNN, as KNN uses distance to classify the point we want to the right class. Th effectiveness for the model in general is great( if we are talking about our case). Despite the accurate results, a disadvantage that we should look at is that KNN takes a lot of time, sand we may find difficulties trying to find the best K value. Although if we got accuracy measures =1 which can be sometimes considered as wrong value, this shows us that using KNN in our case can give us accurate results and prediction.

# References

Anon., 2019. *Pros and Cons of Supervised Machine Learning.* [Online]   
Available at: https://pythonistaplanet.com/pros-and-cons-of-supervised-machine-learning/  
[Accessed 26 1 2023].

Anon., 2021. *Pros and Cons of Unsupervised Learning.* [Online]   
Available at: https://pythonistaplanet.com/pros-and-cons-of-unsupervised-learning/  
[Accessed 26 1 2022].

Anon., 2022. *5 Types of Classification Algorithms in Machine Learning.* [Online]   
Available at: https://monkeylearn.com/blog/classification-algorithms/  
[Accessed 26 1 2023].

Anon., 2022. *Advantages and Disadvantages of Regression Model.* [Online]   
Available at: https://www.vtupulse.com/machine-learning/advantages-and-disadvantages-of-regression-model/  
[Accessed 26 1 2023].

Anon., 2022. *What are the common applications of supervised & unsupervised learning.* [Online]   
Available at: https://www.cognixia.com/blog/what-are-the-common-applications-of-supervised-and-unsupervised-learning/  
[Accessed 26 1 2023].