**A picture containing graphics, font, symbol, red

Description automatically generated**

**10204210 Data Analytics**

**Section 3**

**21110510**

**Nasrullah Rami Hamad**

**Dr. Bassam Kasasbeh**

**Eng. Yara Harahsheh**

**1: Introduction**

**1.1: Data Analytics Activities:**

Data analytics is the process of identifying patterns and trends inside data sets concluding by making judgments about the information they provide and offer. Data analytics is widely used in businesses and marketing because it helps them increase their revenues and improve their efficiency in the future. They give a chance for organizations and businesses to respond to trends as fast as possible, which creates a more productive and competitive business environment.

Organizations use data analytics by using old and historical data to study patterns and correlations regarding a certain feature. And then use the trends to increase their businesses by predicting how can the next pattern occur. The data used in predicting can either be numerical such as numbers and quantities or categorical like texts or images.

Data analytics provides us with more security by studying patterns and identifying fraud patterns. It can also be used in the medical field and healthcare by studying patterns of a disease and then getting to know if the patient is diagnosed by the disease or not by how patterns of the features play out.

**1.2: Data Analytics Techniques:**

Data analytics techniques are the methods used to achieve the best performance when studying patterns and correlations between data inside datasets, the techniques are necessary to make sure your data is shown in the intended way. Data analytics techniques differ from visualizations and EDA to statistics and feature measures.

**Time Series Analysis:** This technique studies the patterns and trends on a specific range of time and day, it can be used in businesses and organizations by helping them understand how and when the patterns can occur during a period, depending on some features. It can include some time-related visualizations about the patterns, and timestamps where the pattern occurred.

**EDA (Exploratory Data Analysis):** This technique involves studying the dataset and its features to provide a better examination of its structure and the correlations between the features, it includes visualizing charts and graphs same as the point above, and it can help the data analytic professional choose the best features depending on the dataset and the target they’re trying to predict.

**Statistics:** Statistics are very crucial in a data analytics professional’s work since it can give either a summary or a better understanding of the dataset and its features. For example, getting the mean for some of the features can help with getting a view at what the average of a single feature is, therefore it can help the data scientists with understanding the level of the feature. The mean can include some outliers, which sometimes give wrong feature information.

**Machine learning:** machine learning can help in gathering information about large sets of data to use visualizing methods on them after that. It can also help with predictions if the model was trained on accurate data since it can give high metric readings with high predicting accuracy. Machine learning models differ from supervised models such as classification and regression models to unsupervised models such as clustering.

**1.3: Data analytics tools:**

Data analytics tools differ depending on what each tool can do, what it is used for and when should it be used in data analytics. Some tools excel at visualizing the data and analyzing it, some other tools can provide the data in a more structured shape as rows and columns. And some tools can clean and transform data to be ready to be trained for models. Examples on these tools:

* **Microsoft Power BI:** This tool is normally used in business intelligence areas since it offers visualization tools and analyzing methods. It can be used similarly to Microsoft Excel, and it is used for data visualizing and charts to describe the dataset and its features, it can generate reports about some questions that may be frequent in organizations, and it helps with predicting too.
* **Python:** Python is one of the most popular programming languages that is known for its massive collection of libraries that can be used for analytical purposes, Numpy, pandas, seaborn, and matplotlib are some of the libraries that enhance the analytical environment by providing efficient graphs and charts, with the help of pandas and numpy’s functions and operations, the data used for visualizing wouldn’t be more accurate and understandable.
* **R:** R is also one of the most open-source programming tools that can provide data visualization methods and statistical information about data, R can also create and build a statistical model that uses data that updates constantly whenever new data is collected. R is widely used in Business Intelligence due to its high performance in mathematical functions and algorithms.
* **Microsoft Excel:** Microsoft is very popular when it comes to structuring data in a row-column type of structure, datasets are easily imported there, and the features are easily sorted. Excel provides data visualization with charts and includes statistical measures such as mode and mean. It can also be used to filter out outliers and prevent any missing values from getting inside the data set because it can give wrong analytical information and bad visualizations.
* **Tableau:** this tool allows analyzing and sharing the data with efficient visualizations that are understood by everyone, whether a person of the intellect for analytics or someone with no knowledge about it. It also excels in performing difficult and complex mathematical calculations, so it can be used widely in business intelligence too.

**1.4: Types of Data Analytics Methods:**

Types of data analytics are the key to answer questions about the analytical process of the data, first starting with, *“what happened?”* after answering this question, we must know *“what might happen in the future”*. After that, we need to know “*What should we do next”.* These questions can be answered by the three types of data analytics:

* Descriptive Analytics: this type of analytics is the foundation of the analytical process; all other types depend on it. This type lets us study the data set we have and gather information and see the correlations and patterns between the features, by answering the question *“What happened?”.*
* Predictive Analytics: this type of analytics answers the question “*what might happen in the future?”* by planning and predicting future patterns or events by offering old data that relate to the trend we’re trying to predict. The higher the data’s relation to the trend, the better the accuracy of the prediction is.
* Prescriptive Analytics: the last type of analytics answers the last question “what should we do next?”. This type focuses on making improvements for the analytical process, and providing meaningful suggestions that may be useful when making decisions in the future. It is also used to optimize and balance how the dataset works so it stays clean, so the process and pattern visualizations stay optimized.

**1.5: Uses of data analytic methods in real life:**

**Descriptive analytics:** This type is normally used in advertisements, since businesses advertise their products and services via internet, they need to know if their products are doing well and if they are getting good satisfaction levels from the customers, so they will be able to improve their performance and the product’s efficiency.

**Predictive analytics:** This type can also be used in the same field as advertisements and marketing, such as product recommendations that are based on previous purchases, this is already used in many companies such as Amazon and E-bay. The way it works is that customers purchase products from a specific category, then the system recommends them other similar products that work in the same field. This is based on machine learning and models that require previous data based on each category to be able to recommend accurate products for the customers.

**Prescriptive analytics:** prescriptive analytics can be used in fraud detection in the banking field since it can keep records of data and provides a recommendation if a change happens. For example, the algorithm keeps track of how much a single customer spends each month. Let’s say he spends 1,000 JDs per month, but on a specific month, he spent 500,000 JDs. What the algorithm does is it keeps saving the patterns of the money spent each month, then when it identifies a massive change in the patterns, it recommends the banking employees about the pattern, and they can take the appropriate action.

2.1: Descriptive Analytics

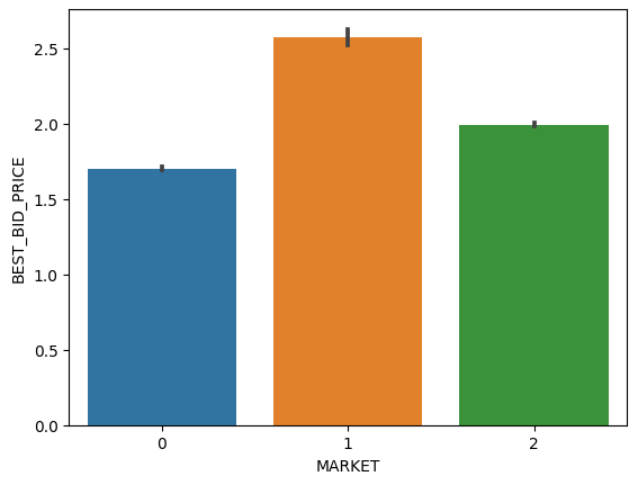
2.1.1: Features Analysis and explanation.

|  |  |  |  |
| --- | --- | --- | --- |
| **Feature no.** | **Feature Name** | **Descriptive Measure / Technique** | **Explanation** |
| **1** | Best Bidding Price | Mean | The mean was used to summarize the average value of the feature |
| Mode | The mode was used to identify the most common value between other values. |
| Median | Median can be compared to the mean to check if their values are close. So we can check if outliers exist. |
| **2** | HIGH | Deciles (1 -> 10) | Deciles were used to know how much the high price is getting increased after each 10th so we can deal with outliers if found |
| Variance | Variance was used to know how much each value is far from the mean, so it can show how different values are. |
| **3** | LOW | Z-score | By using Z-score, we can understand how many STDs values are far from the mean, which can aid in identifying outliers and solve them. |
| IQR | Measuring the range between Q1 and Q3 is efficient in identifying outliers and to get a better understanding of how data is spread. |

2.1.2: Feature Visualization and Explanation.

**Feature 1:**

The following graph shows the relationship between the Best Bid Price feature, and the market feature, it shows the market number 1 which is the second market having the highest bidding price in all the other markets, while market number 0 has the lowest bidding price of the three markets



**Feature 2:**

The graph below shows how scattered the HIGH feature values are with the number of trades from the NO OF TRADES feature, the graph shows that when the numbers of trades ranged between 0 -> 1000 , the highest value for the trades was between 0 -> 8, also 30 -> 40

A picture containing text, screenshot

Description automatically generated

**Feature 3:**

The graph below shows that the LOW feature values ranging between 0 to 3 numbered nearly 2000 counts. This means that the most valid LOW feature value was 1.5 if we took the average.

A picture containing text, screenshot, diagram, plot

Description automatically generated

**2.1.3: Contingency Table & Explanation:**

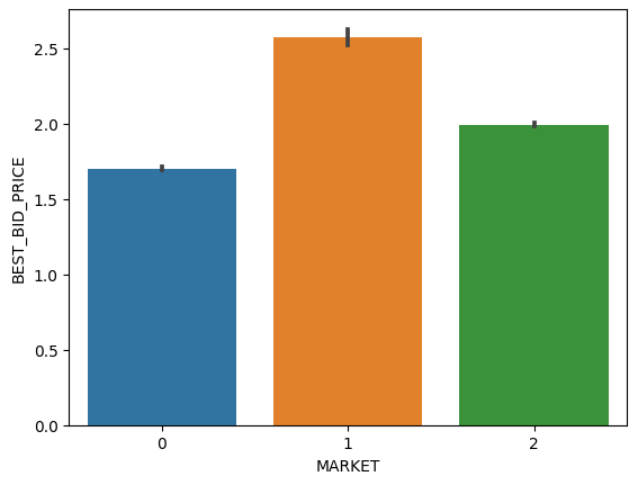
A screenshot of a number

Description automatically generated with low confidence

The contingency table above showed a whole month of visiting markets 0, 1, and 2 while counting how many visits each market had each day till the last day of the month. As shown in the table, Market 2 had the most visits each day nearly ranging between 400 -> 600. Doubling the visits from market 1 and tripling the visits from market 0. While market 0 had the lowest number of visits for the whole month nearly ranging between 100 -> 200.

**2.2: Decision-Making Techniques:**

- By getting to know and understand the measures used for both the HIGH and LOW features, I have decided to get rid of the outliers inside the dataset to get a more accurate understanding of the dataset. That way, it would give a more accurate explanation for the employees and stock buyers since they will have to buy stocks with its at its lowest price. And they should sell their stocks when the price is at its highest peak.



- The market graph above showed that MARKET 1 has the best bidding prices, which can be used as an example that bidding in MARKET 2 is more frequent in trading. This can encourage the MARKET owners to improve its infrastructure since it keeps attracting bidders and stock buyers.

**2.3: Evaluation:**

First, I selected three features from the dataset (HIGH, LOW, BEST BID PRICE) based on the need to study the target values, and to see if the values are close to each other or no. Then I used measuring techniques such as mean, median, deciles, Z-scores and IQR to know how the data is skewed and spread inside the dataset for these three features. And to know how far the values are from the mean.

After that, I used charts and graphs to have a visual understanding of feature relations and correlations. I used a Bar plot to plot how the Best Bid Price was inside each market. And a scatter plot to know what the Highest trading value was depending on the NO of trades. And I used a count plot to count the LOW feature values and to see what value was the most counted.

Then, I made a contingency table that includes a whole month day by day, and the markets and their number of trades each day. The table showed that market 2 had the most visits and market 0 had the lowest number of visits.

3: Predictive Analysis

3.1.1 Feature selection Techniques:

|  |  |  |  |
| --- | --- | --- | --- |
| **FS no.** | **Name** | **Description** | **Results (Selected Features)** |
|  | Low Variance | This feature selection eliminates features with low variance. Since these features have little to no use inside the modelling process. | [SEC CODE, SYMBOL1, MARKET, VOLUME, TRADEQTY, NOofTRADES, BEST ASK PRICE, BEST BID PRICE, BEST ASK QTY, BEST BID QTY] |
|  | Sequential Feature Selection | The sequential works by selecting the most appropriate features based on two ways, forward selection which starts with an empty set and keeps adding features, and it eliminates the feature with the least performance for the model, and backward selection which starts with a full set and removes elements, and eliminates the features that does the smallest decrease in model performance. | [SEC CODE, MARKET, VOLUME, NO OF TRADES, BEST ASK PRICE, BEST BID PRICE] |

3.1.2 Regression Techniques:

|  |  |  |
| --- | --- | --- |
| **Tech. no.** | **Name** | **Description** |
|  | Linear Regression | This model uses the linear formula to find the best fit line that collects most of the values. |
|  | Decision Tree Regression | This model works the same as the decision tree but instead, it uses a regression formula as the tree splits down into parts. |
|  | KNeighbors Regression | This model uses the Kneighbors algorithm with euclidian distance to find the best K value for the model. |

**3.2: Comparing the techniques:**

**“Low” Prediction**

**Comparison:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **FS no.** | **Tech no.** | **MAE** | **MSE** | **RMSE** | **R2** |
| **1** | **1** | 1.605 | 17.088 | 4.131 | 0.183 |
| **1** | **2** | 0.215 | 1.237 | 1.105 | 0.941 |
| **1** | **3** | 0.292 | 1.635 | 1.274 | 0.922 |
| **2** | **1** | 1.654 | 18.477 | 4.296 | 0.117 |
| **2** | **2** | 0.235 | 1.525 | 1.232 | 0.927 |
| **2** | **3** | 1.579 | 20.201 | 4.492 | 0.033 |

**Visualization of results:**

A screenshot of a graph

Description automatically generated with low confidenceA screenshot of a graph

Description automatically generated with low confidence

**“High” Prediction**

**Comparison:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **FS no.** | **Tech no.** | **MAE** | **MSE** | **RMSE** | **R2** |
| **1** | **1** | 1.63 | 17.633 | 4.196 | 0.188 |
| **1** | **2** | 0.224 | 1.372 | 1.162 | 0.937 |
| **1** | **3** | 0.295 | 1.691 | 1.295 | 0.922 |
| **2** | **1** | 1.63 | 17.633 | 4.196 | 0.188 |
| **2** | **2** | 0.224 | 1.378 | 1.165 | 0.937 |
| **2** | **3** | 0.295 | 1.691 | 1.295 | 0.922 |

**Visualization of results:**

A screenshot of a graph

Description automatically generated with low confidenceA picture containing text, screenshot, display, plot

Description automatically generated

**3.3: Evaluation:**

The feature selection methods used were Low variance and Sequential Feature Selection, they were used with the models Linear Regression, Decision Tree Regression, and KNeighbors Regression. The models were done with 30 iterations.

The target attributes were HIGH and LOW, they had many features that correlated with their prediction, and some features that were decreasing the performance of the models, the feature selection methods were used to find the best features that could improve the performance of the models and their predicting accuracy.

The four measures that were used to compare between the performance of the models were R2 score, MAE, MSE, and RMSE, the accuracy was measures by the R2 score, and the three other measures were the errors which if increased, the accuracy of the model decreases.

The results of the models were good for the most part, they produced good results for the Decision Tree regressor and the KNNR with an accuracy of 92%, but the Linear regression’s results were not good, since the accuracy measured was only 18%. The sequential feature selection and the low variance’s results were nearly similar in the HIGH column, but there was a difference inside the LOW column.

**4: Prescriptive Analysis:**

**4.1 Techniques with examples:**

|  |  |  |
| --- | --- | --- |
| **Tech. no.** | **Name** | **Description** |
|  | SSA | SSA is a method used to find the best path in a system with uncertainty. Its goal is to maximize the function and minimize the cost. |
|  | PSO | PSO is a method that is used to solve large search space optimization problems. This method is known for its efficiency, and its optimization performance in machine learning. |
|  | MFO | MFO is a method that’s based on how moths attract to light or flames. The model sees moths as the solution and and the light as the problem, it grabs the moths that are closest to the flame which are the optimal solution to the problem. |

**4.2:** The optimization techniques were used to find the optimal solution, and the methods I used were SSA (Shortest Path Analysis), PSO (Particle Swarm Optimization), and MFO (Moth Flame Optimization).

* SSA: This method is used to find the optimal path for a problem with uncertain outcomes, it uses the graph theory and probabilities combined with dynamic programming to model the problem as a directed graph. Nodes represent the current problem, and edges represent the possible solution. Probabilities are used in considering if the state should be changed, and the method’s goal is to find the solutions that maximize the optimization performance and minimize the cost. The method is known for its decision-making ability and its effectiveness and efficiency in optimizations for models.
* PSO: This method uses particles to find the optimal solution by searching in a large space. The particles swarm in the large space in their position and they often use velocities to get closer to the optimal solution, each particle is a solution. At each iteration, particles either get closer to each other by their velocities by communicating or they move far away. The process continues until the system turns down either by reaching the number of iterations needed or the optimal solution is met. The method is known for its flexibility and efficiency in optimization problems. It is also used for optimizing machine learning and clustering of data.
* MFO: This method is based on the idea of moths being attracted to the light of flames where each moth is a solution, and the light or flame is the optimal solution. Moths usually get attracted to brighter flames which are the best optimal solution and their movement is normally randomized, so they explore more in the search space looking for a more optimal solution. This method is known for solving a large range of problems. Such as feature selection, and image processing, and it’s also simple to use complex spaces to find the optimal solution.

**4.3:** Objective Function Code.

Def Calculator(q):

Minimum = 0

q = numpy.round(q)

x = numpy.array([1.33, 5.59, 1.6, 0.47, 0.33, 0.58, 0.5, 0.47, 0.83, 1.14, 1.23, 1.19, 0.1, 1.18, 1, 1.36, 0.5, 0.45])

mincost = numpy.multiply(q,x).sum()

if q.sum() < 10:

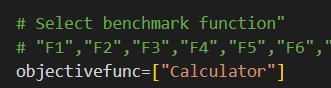
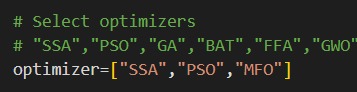
return 999999

else:

return mincost

4.4: Apply the techniques:

**4.4.1:** Code screenshots:



**4.4.2: Results & Explanation**

The SSA method was used to optimize and give the lowest prices for 10 stocks. The cheapest price to buy 10 stocks in the SSA method is 4.52 which is cheaper than the other two iterations since the other iterations gave a higher price for 10 stocks measuring by 6.01 and the most expensive price for 10 stocks inside the SSA method is 4.52.

The PSO method was used in three iterations, to get the cheapest price of buying 10 stocks. The cheapest price iterated for 10 stocks in the PSO method was 3.4 which is way cheaper than the SSA method’s cheapest stock prices. And the other two iterations resulted in stock prices between 8.25 and 5. Which are more costly than the SSA method.

The MFO method is also used in three iterations, to get the cheapest price for buying 10 stocks, the cheapest price generated for 10 stocks was the cheapest price in all the other methods resulting in 2.15 JDs for 10 stocks. The other values iterated for 10 stock prices ranged between 2.50 to nearly 2.75, which is cheaper than all the other stock prices generated in the other models.

Based on the stock prices from each method above, the best method used was the MFO method, resulting in the cheapest price of 10 stocks (2.15). and the most expensive price for 10 stocks was generated by the SSA method making it the worst method of the three.

**4.4.3: Visualization & Explanation:**

**A picture containing text, screenshot, diagram, rectangle

Description automatically generated**

The box plot above shows the range for 50 iterations of 10 stock prices generated by the three methods (SSA, PSO, MFO). Based on the graph, the SSA method generated the highest price for 10 stocks resulting in it being the worst method from the other three (4.52). The PSO method generated stock prices that were placed between the SSA and MFO range of stock prices (3.4). However, the MFO method generated the lowest and cheapest prices of 10 stocks, resulting in it being the best method there is (2.15).

**A picture containing line, plot, diagram, text

Description automatically generated**

This line plot includes the three methods generating the prices of 10 stocks 50 times. Looking at the graph, the lowest price generated was the green line (MFO method), that makes it the best method used to generate the cheapest price of 10 stocks. The yellow line was drawn between the other two lines making it the 2nd best method of the three (PSO method). The blue line however generated the highest price of 10 stocks, making it the worst method out of the three methods (SSA method).

**References:**

1. Stedman, C. (n.d.). *What is Data Analytics? - Definition from WhatIs.com*. [online] SearchDataManagement. Available at: <https://www.techtarget.com/searchdatamanagement/definition/data-analytics.>
2. Tableau. (n.d.). *What Is Data Visualization? Definition, Examples, And Learning Resources*. [online] Available at: [https://www.tableau.com/learn/articles/data-visualization#:~:text=The%20importance%20of%20data%20visualization.](https://www.tableau.com/learn/articles/data-visualization%23:~:text=The%20importance%20of%20data%20visualization.)
3. Scardina, J. (2022). *What is Microsoft Power BI? - Definition from WhatIs.com*. [online] SearchContentManagement. Available at: <https://www.techtarget.com/searchcontentmanagement/definition/Microsoft-Power-BI.>
4. Tableau (2022). *What is Tableau?* [online] Tableau. Available at: <https://www.tableau.com/why-tableau/what-is-tableau.>
5. Sisense. (n.d.). *What is R Analytics? Data Analytics With R Explained*. [online] Available at: [https://www.sisense.com/glossary/r-analytics/.](https://www.sisense.com/glossary/r-analytics/)

‌

1. Orbit Analytics. (n.d.). *Descriptive Analytics*. [online] Available at: [https://www.orbitanalytics.com/descriptive-analytics/#:~:text=Descriptive%20analytics%20come%20into%20picture](https://www.orbitanalytics.com/descriptive-analytics/%23:~:text=Descriptive%20analytics%20come%20into%20picture%20) [Accessed 10 Jun. 2023].

‌