# Introduction

## Data analytic activities

A number of tasks are involved in data analytic activities which are crucial for gaining insights by understanding the data. These can be divided into five main groups:

* Data collection: this includes collecting data from a variety of sources, such as surveys, databases, and online tools such as social media platforms. it is crucial to gather relevant and trustworthy data, to guarantee the accuracy and validity of the analysis.
* Data cleaning and preprocessing: after the data is collected, it needs to be cleaned from any errors and correcting them. These errors can be dealt with such as handling missing values, delete the duplicate entries, and standardizing the data in one format. These step ensures that the data is now in a high quality and ready to be explored and for modelling.
* Data exploration: in this step, we get to know the data, understand it, and gain insights from it. This includes applying statistical analyses, visualizing the data in charts and graphs, and spotting some patterns, trends, or correlations. This can be helpful in order to understand more the data and explore interesting relationships and potential areas for investigation.
* Data modelling: in this step we apply the models on the data after it is ready, these models will understand and interpret the underlying patterns and relationships in the data. Data analysts can then uncover valuable insights and make predictions which will affect in real problems decision making process based on the data. These predictions can play a crucial role in various domains such as business and healthcare.
* Data visualization: it is the process of representing the data and findings in a visual format such as charts, graphs, and many other visuals. In order to understand the results and findings of the models effectively, we need to visualize these findings and the data. Moreover, data visualization is important in order to help to deliver and communicate the results of the model to the stakeholders and non-technical people.

## Data analytic techniques

Data analytic techniques are a set of processes that can be performed in order to help in collect, preprocess, explore the data in order to extract meaningful insights and results, these techniques are such as:

* Statistical analysis: is the use of statistical methods to analyse the data, these can be helpful in finding patterns, relationships, and trends in data, moreover, we can identify outliers. Statistical analysis allow us to gain insights into the central tendency, dispersion, and the shape of distributions. Measures can be such as mean, median, standard deviation and correlation coefficient.
* Exploratory data analysis (EDA): it is a systematic approach to analyse the data sets in order to gain deeper understanding of the data and figure out the key features and patterns of the data. In EDA we spot errors such as anomalies, outliers, and missing values. Also, we understand the patterns such as trends, correlations, or seasonality. Moreover, we can answer questions such as about the presence of outliers, the relationship between variables, or know the distribution of the data.
* Data visualization: it is the process of representing the data in a visual format such as graphs and charts. Data visualization can help the data scientist to better understand the data and the main advantage is that data visualization help in the process of delivering the results of data analysis to stakeholders.
* Natural language processing (NLP): in NLP, we make the computer understand and process unstructured textual data which is the human language such as customer reviews. NLP works on analysing the textual data and extract a meaningful information from, can be used in many applications such as content categorization and recommendations systems.

## Data analytic tools

Data analytic tools have a crucial role in processing, analysing, and visualizing the data to come up with meaningful insights and results, these tools can be:

* Python: it is a versatile programming language that have many powerful libraries that can be used for data manipulation, analysis, using machine learning models, and visualizing the data. These libraries are such as NumPy, pandas, seaborn, and matplotlib.
* Microsoft excel: it is a spreadsheet software program that is widely used, can be used for data analysis such as in calculations, basic statistical analysis, and visualization tools.
* Power BI: in this tool users can connect to different data sources, transform the data, clean it up, and generate interactive visualizations and reports using this tool. Data modelling is one of the advanced features offered by Power BI. It enables businesses to collect insights from data and act based on that data.
* Tableau: is a powerful data visualization tool that helps users to analyse the data and present it in an interactive way, it helps in exploring the data and to gain insights of data.
* Apache spark: it a big data processing and analytics framework, it offers various libraries for data manipulation and machine learning.

## Types of data analytic methods

Data analytic methods can be categorized into three main methods: descriptive analytics, predictive analytics, and prescriptive analytics.

* Descriptive analytics: this method involves analysing the current and historical data in order to understand, describe and summarize to what happened in this data. It focuses on providing insights and visualizations about the trends, patterns, outliers, and relationships in the data. This is used in order to get a better understanding of the data which will help the data scientist in the predictive analytics, such as identifying relevant features and identifying patterns that will be used in building predictive models.
* predictive analytics: this method involves using the historical data for the predictive models in order to make predictions and forecast future outcomes. It uses techniques such as regression models, and machine learning algorithms to identify the patterns, trends, or construct equations to make informed predictions about future situations. This can be done by using many independent features that have an influence or relationship on a dependent feature or the target feature while minimizing the cost, error, or inaccuracies in the prediction. This can be helpful in predicting future outcomes based on some inputs.
* prescriptive analytics: this method takes the analysis a step further by providing recommendations and actions to optimize the outcomes. This use various optimization techniques that finds the optimum values for these decision variables that will lead to the best possible outcome, and it should meet quantity of interest such as reducing costs, maximizing profits, or minimizing errors. Also, it considers various possible scenarios, uses optimization algorithms, and decision-making techniques in order to reach the best possible outcome that meets the interests. Prescriptive analytics support decisions’ makers by providing them the best course of action to achieve the desired outcomes.

## Uses of data analytic methods in real life

Let’s assume that we have a telecommunication company that wants to reduce the customer churn and improve customer retention. They have a large dataset that contains historical customer information, demographics, usage patterns, and service subscriptions. This company wants to use the data analytic methods in order to improve customer retention and reduce customer churn.

Descriptive analytics: the company applies descriptive analytics to gain insights about customer behaviour and churn patterns. They begin with analysing historical data to identify common characteristics or features among churned customers, such as service dissatisfaction, pricing issues, or usage patterns. By understanding these factors, they can develop strategies and methods in order to mitigate customer churn and retain customers satisfaction.

Predictive analytics: the company aims to build a prediction model that predict the likelihood of customer churning in the future based on multiple inputs. They trained the model on features such as service usage, payment history, call durations, and frequency of complaints. By training this model, the company can predict the likelihood of customer churn, and this can benefit the company to proactive procedures to retain the customers through targeted campaigns or offers.

Prescriptive analytics: the company utilizes the prescriptive analytics, optimization algorithms, and decision-making techniques in order to identify the best strategies for reducing the customer churn. By considering multiple scenarios and constraints, they can determine the most appropriate and effective allocation of resources such as developing targeted campaigns and offers or enhancing customer support for those customers. This can benefit the company in taking the best course of actions and achieve their target by reducing customer churn and improve customer retention.

# Descriptive Analytics

## Techniques & Examples (Your work)

### Features Analysis and Explanation

|  |  |  |  |
| --- | --- | --- | --- |
| **Feature no.** | **Feature Name** | **Descriptive Measure / Technique** | **Explanation** |
| **1** | Market | Frequency and relative frequency | This will indicate to the markets that has the highest trades, which will affect the prediction of the model, but it will be beneficial for an investor to invest in the highest trades market |
| Number of trade in each market and classifying them by month | The purpose of this is to know the number of trades in each market from month to month, this will help to more understand the data set in order to know which months have higher number of trades and vice versa |
| **2** | VOLUME, TRADE\_QTY, NO\_OF\_TRADES | Measure of central tendency (mean) on every market value | The purpose of these descriptive statistics is to identify trends and patterns, which can be used to help in making investments decisions such as deciding which market is more active or have high number of trades |
| Measure of dispersion (slandered deviation, percentiles, min, and max values) on each market value |
| **3** | Correlation between the BEST\_BID\_PRICE and BEST\_ASK\_PRICE and TRADE\_QTY | Calculated the correlation between these three columns | Correlation can benefit us in many things such as identifying trends and patterns between the columns, also we can determine the strength of relationship between the columns. Moreover, it helps in feature engineering as if we want to remove some columns |

### Features Visualization and Explanation

**Feature 1:**

|  |  |
| --- | --- |
| Measures of Frequency on MARKET (I used the bar plot because it provides a clear representation of the categorical columns) | Figure 1.1 |
| Number of trade in each market and classifying them by month (I used the line plot because it clearly shows the changes in a column over time, which makes it easy to compare and identify the trend) | Figure 1.2 |

**Feature 2:**

|  |  |  |
| --- | --- | --- |
| Mean of VOLUME, TRADE\_QTY, NO\_OF\_TRADES columns in every market (bar plot is good to represent the distribution of a column values based on other factor such mean or standard deviation). | | |
| Figure 2.A picture containing text, screenshot, diagram, rectangle  Description automatically generated2.1 | Figure 2.**A picture containing text, screenshot, diagram, rectangle  Description automatically generated**2.2 | Figure 2.A picture containing text, screenshot, diagram, rectangle  Description automatically generated2.3 |
| Standard deviation of VOLUME, TRADE\_QTY, NO\_OF\_TRADES columns in every market | | |
| Figure 2.3.1 | Figure 2.3.2 | Figure 2.3.3 |

**Feature 3:**

|  |  |
| --- | --- |
| Correlation between the BEST\_BID\_PRICE and BEST\_ASK\_PRICE | The correlation = 0.9999988 |
| I used the heatmap because it provides a comprehensive and visual representation of the correlation between columns  Figure 3.1 | I used the scatter plot to visualize the correlation between the BEST\_BID\_PRICE and BEST\_ASK\_PRICE in order plot each value in them, in order to show the pattern and relationship between them.  Figure 3.2 |

### Contingency table and Explanation

I did a contingency table between MARKET and SYMBOL1 columns in order to understand the distribution of symbols across different markets. It gives insights on the distribution of symbols over every markets which can reveal any pattern.

This can be helpful for many reasons such as:

* Firstly, it can identify market specific symbols which means that it helps to determine the symbols that are exclusive or unique for one specific market. This can be helpful the most to investors, as they want to decide in which symbol, they would invest so they would search for symbols that have trades in their desired markets section.
* Allows us to see the diversification of symbols among markets, which will show if the distribution is balanced among different market section.

## Techniques for decision-making (Your work)

Feature 1:

Based on figure (1.1), we can determine two things:

* Firstly, is that the machine learning models can be biased to the markets that have more trades. Since the model has more data from market 2 compared to other markets, it may learn patterns and make predictions that are more biased towards market 2. The data scientist can try to do oversampling and under sampling techniques.
* Secondly, for an investor it would benefit him more to invest in market 2 as it has more trades, which means the market will be more active.

Based on figure (1.2), we can determine that the number of trades increase from the beginning of the year until March, then the number will decrease until April, then it will increase reaching the maximum among each market in June, after that it will decrease until the end of the year. This can be beneficial for an investor to buy just before those months where the number of trades increase, because when the number of trades increase the prices will increase also, so he would sell in those months.

Feature 2:

Based on the above figures for the mean and standard deviation for the columns (VOLUME, TRADE\_QTY, NO\_OF\_TRADES), we can figure out many things:

* Based on the figures (2.1.1, 2.1.2, 2.1.3). We can figure out that market 1 has the highest mean value in VOLUME column (152089) while market 2 mean value in VOLUME column (32769), and so on other markets and other column. But from these mean values we can determine that market 1 has the highest mean value of its trades upon other markets and this mean value is the highest whether in the average of sum total of all the shares of a particular company, or in the average of total number of shares that have been bought or sold in a specific period of time. Which can help the investor to know that market 1 has a high revenue if he decided to invest in it.
* Based on the figures (2.2.1, 2.2.2, 2.2.3). we can determine that the data is highly volatile, which will encourage the data scientists to focus on using more robust statistical models such as decision tree and random forest. Moreover, this high STD of market 1 in VOLUME can indicate that market 1 has lot of volatility unlike market 2 has less volatile, so if an investor would prefer high volatile, then market 1 would be a good option and vice versa.

Feature 3:

Based on figure (3.1), we can see that the correlation between TRADE\_QTY column with the BEST\_BID\_PRICE and BEST\_ASK\_PRICE is 1. But we can’t drop one of the columns as they are almost perfectly linearly related, because TRADE\_QTY and the other columns have different types of variables representing different aspects of the trading process, so it more reasonable to keep these columns as they distant information.

But on the other hand, based on figure (3.1, 3.2), we can see that the correlation between BEST\_BID\_PRICE and BEST\_ASK\_PRICE equal 0.9999988. It is recommended to drop one of those columns to avoid redundancy in data as they provide similar information and they are highly correlated, so keeping them both wouldn’t provide significant insights in most cases.

## Evaluation (Your work)

Based on the above techniques and the decision-making process depending on them and their results, I have dropped the BEST\_ASK\_PRICE, as it is almost linearly related with BEST\_BID\_PRICE, so one of these columns have to be removes in order to avoid redundancy in data and keeping them don’t provide additional significant information and to be more efficient in the data modelling.

Moreover, I should focus more on using robust statistical methods such as decision tree and random forest, because the data has many columns that are highly volatile. And these models is good to use because they are less sensitive to extreme values, ensuring that it has less influence on the models. Moreover, these models downweigh the influence of extreme values in order to reduce such model bias and provide more balanced representation and predicting of data. And by this we helped in solving the problem of having market section 2 has more data than other sections, which can influence the model to be biased towards it.

# Predictive Analysis

## Techniques & Examples (Your work)

Feature Selection Techniques

|  |  |  |  |
| --- | --- | --- | --- |
| FS no. | Name | Description | Results (Selected Features) |
|  | Select from model | It is a feature selection that train the machine learning model and choose the most important features depend on specific threshold. Firstly, it train on the whole dataset, then the importance of each feature is calculated, the feature of the lowest importance is removed, then the process repeats until the desired number of features is met. | Linear regression:  ’TRADE\_DATE', 'SYMBOL1', 'MARKET', 'NO\_OF\_TRADES', 'BEST\_BID\_PRICE' |
| Decision tree:  'SEC\_CODE', 'SYMBOL1', 'VOLUME', 'TRADE\_QTY', 'BEST\_ASK\_QTY' |
| Random forest:  'SEC\_CODE', 'SYMBOL1', 'VOLUME', 'TRADE\_QTY', 'BEST\_ASK\_QTY' |
|  | Recursive feature elimination | RFE runs by recursively removing features from the model that are of the least importance. A feature's importance is measured by how much it contribute to the model's accuracy. | Linear regression:  'TRADE\_DATE', 'SEC\_CODE', 'SYMBOL1', 'MARKET', 'VOLUME', 'TRADE\_QTY', 'NO\_OF\_TRADES', 'BEST\_BID\_PRICE' |
| Decision tree:  'TRADE\_DATE', 'SEC\_CODE', 'SYMBOL1', 'MARKET', 'VOLUME', 'TRADE\_QTY', 'BEST\_ASK\_QTY', 'BEST\_BID\_QTY |
| Random forest:  'TRADE\_DATE', 'SEC\_CODE', 'SYMBOL1', 'MARKET', 'VOLUME', 'TRADE\_QTY', 'BEST\_ASK\_QTY', 'BEST\_BID\_PRICE' |

### Regression Techniques

|  |  |  |
| --- | --- | --- |
| **Tech. no.** | **Name** | **Description** |
|  | Linear regression | It’s a statistical model that predicts a continuous outcome based on one or more feature. It aims to draw a linear relationship between the input features and the continues output variable. |
|  | Decision tree regression | is a hierarchical model that is built by recursively splitting the data into smaller and smaller subsets based on a series of binary decisions. Where each node represents a decision rule, and each leaf is a predicted value. |
|  | Random forest regression | it is a machine learning that combines multiple decision trees in order to make predictions, it trains multiple decisions tress on different subsets of the data. The predictions of these trees are combined to make the final prediction |

## Compare Techniques (Your work)

**“Low” Prediction**

**Comparison:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **FS no.** | **Tech no.** | **MAE** | **MSE** | **RMSE** | **R2** |
| **SFM** | **LR** | 1.670 | 19.092 | 4.367 | 9.864 |
| **SFM** | **DT** | 0.201 | 1.377 | 1.169 | 93.487 |
| **SFM** | **RF** | 0.139 | 0.524 | 1.169 | 97.529 |
| **RFE** | **LR** | 1.620 | 17.363 | 4.164 | 18.024 |
| **RFE** | **DT** | 0.198 | 1.274 | 1.118 | 94.010 |
| **RFE** | **RF** | 0.145 | 0.542 | 0.732 | 97.438 |

**Visualization of results:**

|  |  |  |
| --- | --- | --- |
| **DT & RF** | **MSE (bar plot)**  Graph 4.1.1 | **MAE (bar plot)**  Graph 4.1.2 |
| **RMSE (bar plot)**  Graph 4.1.3 | **R^2 (bar plot)**  Graph 4.1.4 |
| **LR** | **MSE (bar plot)**  Graph 4.2.1 | **MAE (bar plot)**  Graph 4.2.2 |
| **RMSE (bar plot)**  Graph 4.2.3 | **R^2 (bar plot)**  Graph 4.2.4 |
| **DT & RF** | **MSE (box plot)**  Graph 4.3.1 | **MAE (box plot)**  Graph 4.3.2 |
| **RMSE (box plot)**  Graph 4.3.3 | **R^2 (box plot)**  Graph 4.3.4 |
| **LR** | **MSE (box plot)**  Graph 4.4.1 | **MAE (box plot)**  Graph 4.4.2 |
| **RMSE (box plot)**  Graph 4.4.3 | **R^2 (box plot)**  Graph 4.4.4 |

**“High” Prediction**

**Comparison:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **FS no.** | **Tech no.** | **MAE** | **MSE** | **RMSE** | **R2** |
| **SFM** | **LR** | **1.697** | **19.748** | **4.441** | **10.148** |
| **SFM** | **DT** | **0.185** | **1.055** | **1.019** | **95.215** |
| **SFM** | **RF** | **0.140** | **0.551** | **0.738** | **97.4976** |
| **RFE** | **LR** | **1.646** | **17.909** | **4.229** | **18.509** |
| **RFE** | **DT** | **0.203** | **1.347** | **1.149** | **93.894** |
| **RFE** | **RF** | **0.147** | **0.572** | **0.751** | **97.397** |

**Visualization of results:**

|  |  |  |
| --- | --- | --- |
| **DT & RF** | **MSE (bar plot)**  Graph 5.1.1 | **MAE (bar plot)**  Graph 5.1.2 |
| **RMSE (bar plot)**  Graph 5.1.3 | **R^2 (bar plot)**  Graph 5.1.4 |
| **LR** | **MSE (bar plot)**  Graph 5.2.1 | **MAE (bar plot)**  Graph 5.2.2 |
| **RMSE (bar plot)**  Graph 5.2.3 | **R^2 (bar plot)**  Graph 5.2.4 |
| **DT & RF** | **MSE (box plot)**  Graph 5.3.1 | **MAE (box plot)**  Graph 5.3.2 |
| **RMSE (box plot)**  Graph 5.3.3 | **R^2 (box plot)**  Graph 5.3.4 |
| **LR** | **MSE (box plot)**  Graph 5.4.1 | **MAE (box plot)**  Graph 5.4.2 |
| **RMSE (box plot)**  Graph 5.4.3 | **R^2 (box plot)**  Graph 5.4.4 |

## Evaluation (Your work)

depending on all the figures and graph above we can observe that:

Comparing different feature selection techniques:

We can see that LR (Linear Regression) performed better in the Recursive Feature Elimination (RFE) method, which achieved an R2 (coefficient of determination) of 18 but had a lower R2 value of 9.8 in the SelectFromModel (SFM) method. For both techniques, the other error measures were approximately the same. This suggests that the feature selection methods used affected the performance of LR.

On the other hand, the decision tree (DT) performed exceptionally in the RFE method, achieving a remarkable R2 score of 94. The R2 value for DT in the SFM method was 93.4, which was a little lower. The other error measures for both techniques were also approximately the same. As a result, DT was able to predict the target variable with greater accuracy and is generally more robust in terms of feature selection.

The R2 values for random forest (RF) were almost identical for the two feature selection techniques. Additionally, the results of the error measures for RF were similar, demonstrating consistent performance across the various techniques.

Comparing different models:

We can figure out that LR had higher error measures than DT and RF. The R2 value for LR was noticeably low, indicating that the data were not well fit, or the model is not appropriate for the data. On the other hand, DT and RF both showed better results. In terms of error measures, RF outperformed DT, displaying lower errors across a variety of metrics. Additionally, RF outperformed DT in terms of R2 value. This demonstrates that RF is a better model than DT for predicting the target variable.

In conclusion, the feature selection techniques comparison showed that LR, DT, and RF performed differently. Although LR had a lower R2 value in SFM, it performed better in RFE. While RF demonstrated consistent outcomes across feature selection techniques, DT demonstrated superior performance in both RFE and SFM. Compared to DT and RF, LR showed lower R2 values and higher error measures when the models were considered. moreover, RF demonstrated lower error measures than DT and also achieved higher R2 values. These results offer useful guidance for choosing the best feature selection method and model for the particular dataset and prediction task. The Recursive Feature Elimination (RFE) method and the Random Forest (RF) model emerged as the most effective approaches for achieving higher accuracy in predicting the target variable based on the comparison of feature selection techniques and models.

When observing the model's error measures and R2, a large difference between the minimum and maximum values within the same feature selection technique suggests that the performance of the Decision Tree (DT) model is more inconsistent. The Random Forest (RF) model, on the other hand, displays a lower difference in error measures and R2, indicating a more steady and consistent performance across all error measures. This result demonstrates that, using the same feature selection technique, the RF model is more stable and reliable than the DT model.

# Prescriptive Analysis

## Techniques with Examples (General)

|  |  |  |
| --- | --- | --- |
| **Tech. no.** | **Name** | **Description** |
|  | Gradient decent | it is an iterative method works on minimize the function cost by continuously moving towards the negative gradient. |
|  | Genetic algorithm | Is a nature inspired meta-heuristic method that is inspired by the process of natural selection and genetics. It mimics the principle of survival of the fittest. |
|  | Sine cosine algorithm | Is a stochastic optimization algorithm that finds function global minima by using simulated annealing. |
|  | JAYA | It’s a hybrid metaheuristic algorithm that combines between genetic algorithms and simulates annealing techniques. |

## Techniques for finding the best course of action (General)

* PSO: is a technique that was inspired by the social behaviour of bird flocks. Each particle in the population of potential solutions adjusts its position based on its own knowledge and the influence of the best-performing particle in closest proximity. PSO uses velocity and position adjustments to iteratively update particle positions in order to find the best outcome. The particles can explore the search space and converge on the best answer using this iterative process based on their individual and collective experiences.
* SCA: it’s a simulated stochastic optimization algorithm. Repeatedly, the algorithm updates the solutions based on the sine and cosine functions in order to find the global minima. This algorithm has a robust search for optimal solution algorithm, because it provides a balance between exploration and exploitation.
* JAYA: it’s a hybrid metaheuristic algorithm has features from genetic algorithms and simulated annealing algorithms. It works on enhancing and updating the search process by repeatedly updating the best solution and investigating various other options. It focuses on improving the current solution constantly.

## Objective Function Code (Your Work)

prices=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])

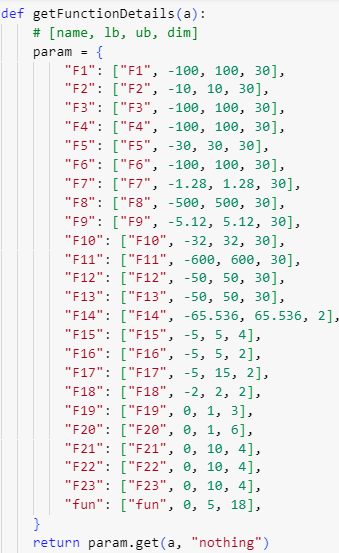
if (numpy.sum(q) < 10):

return 999999999

else:

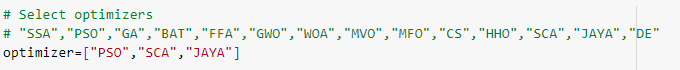
cost = numpy.sum(q \* prices)

return cost



## Apply the techniques (Your Work)

### Code screenshots





### Results and Explanation

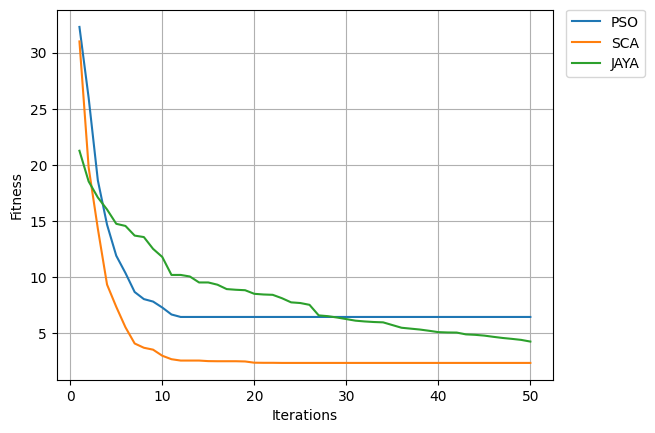
The above techniques have reached different solutions with different function cost value:

Best results

* PSO tried to decrease the cost of purchasing stocks, which came up with a good result of a cost of 6.47. according to this technique optimized solution, the investor should only think about purchasing 10 stocks from two banks: 5 from the ARBK bank stock and 5 from the AHLI bank stock on Thursday. This technique has tried to identify the stocks that have a potential for hight returns. PSO reached the lowest cost at iteration 12.
* Moreover, SCA has utilized to minimize the cost of purchasing stocks, resulting into a cost of 2.38, this cost is an excellent solution for purchasing stocks as it have a very low cost. This solution recommends the investors to buy 5 stocks from the ARBK bank stock on Wednesday and another 5 stocks from the AHLI bank on Saturday. SCA reached the lowest cos tat iteration 23.
* JAYA also has been implanted in order to minimize the cost and achieved a very good result of a cost of 4.28, this cost can be accepted for an investor to depend on it. This technique has suggested a list of multiple stocks with different quantity. These stocks includes the ARBK bank stock on Saturday, Tuesday, and Wednesday. Also purchasing HBTF bank stock on Saturday, Monday, and Wednesday. Moreover, AHLI bank stocks on Sunday and Wednesday. This solution can return a profit for in the investor. JAYA continued to reach the lowest cost at iteration 50.

As a result, the sine cosine algorithm proved to be the most effective algorithm by achieving the best solution as it have the lowest cost among other algorithms of 2.38. Moreover, this algorithm has indicated that the investor should only purchase from two stocks on different day, this can be helpful for investors so they can focus only on two stocks rather than to follow and keeping up with all stocks, and this can prove how precise the algorithm is. Although, the algorithms have reached to the lowest cost at different iterations, we can conclude that PSO has reached to the lowest cost first, then SCA, and lastly JAYA.

### A picture containing text, screenshot, diagram, display Description automatically generatedVisualization and Explanation



In the boxplot we can obtain that SCA has the lowest cost among other algorithms, JAYA in the middle, PSO is the highest. Moreover, we can see that the difference between the max and min cost in PSO is high because it tries a wide range of solutions.

Moreover, through the line plot we can examine that JAYA has started with lower cost compared to other algorithms but the decrease in the cost was slight. While the other algorithms started with higher cost, but the decrease was strong as they reached a cost that is lower than JAYA. Although, we can figure out that the algorithms have reached to the lowest cost at different iterations, where PSO reached first, then SCA, and the last was JAYA.

# References

* Top 10 data analytics tools in 2023: Data Analytics Tools (2023) Edureka. Available at: https://www.edureka.co/blog/top-10-data-analytics-tools/ (Accessed: 18 June 2023).
* Frankenfield, J. (2023) Data analytics: What it is, how it’s used, and 4 basic techniques, Investopedia. Available at: https://www.investopedia.com/terms/d/data-analytics.asp (Accessed: 18 June 2023).
* Gad, A.G. (2022) Particle swarm optimization algorithm and its applications: A systematic review - archives of computational methods in engineering, SpringerLink. Available at: https://link.springer.com/article/10.1007/s11831-021-09694-4 (Accessed: 18 June 2023).
* Gabis, A.B. et al. (2021) A comprehensive survey of sine cosine algorithm: Variants and applications, Artificial intelligence review. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8171367/ (Accessed: 18 June 2023).