MACHINE LEARNING

REGRESSION

CONTENTS

- Regression
- Stock Market Prediction
- Decision Trees
- Linear Regression
- Support Vector Regression
- Conclusion
- References

STOCK MARKET PREDICTION USING REGRESSION

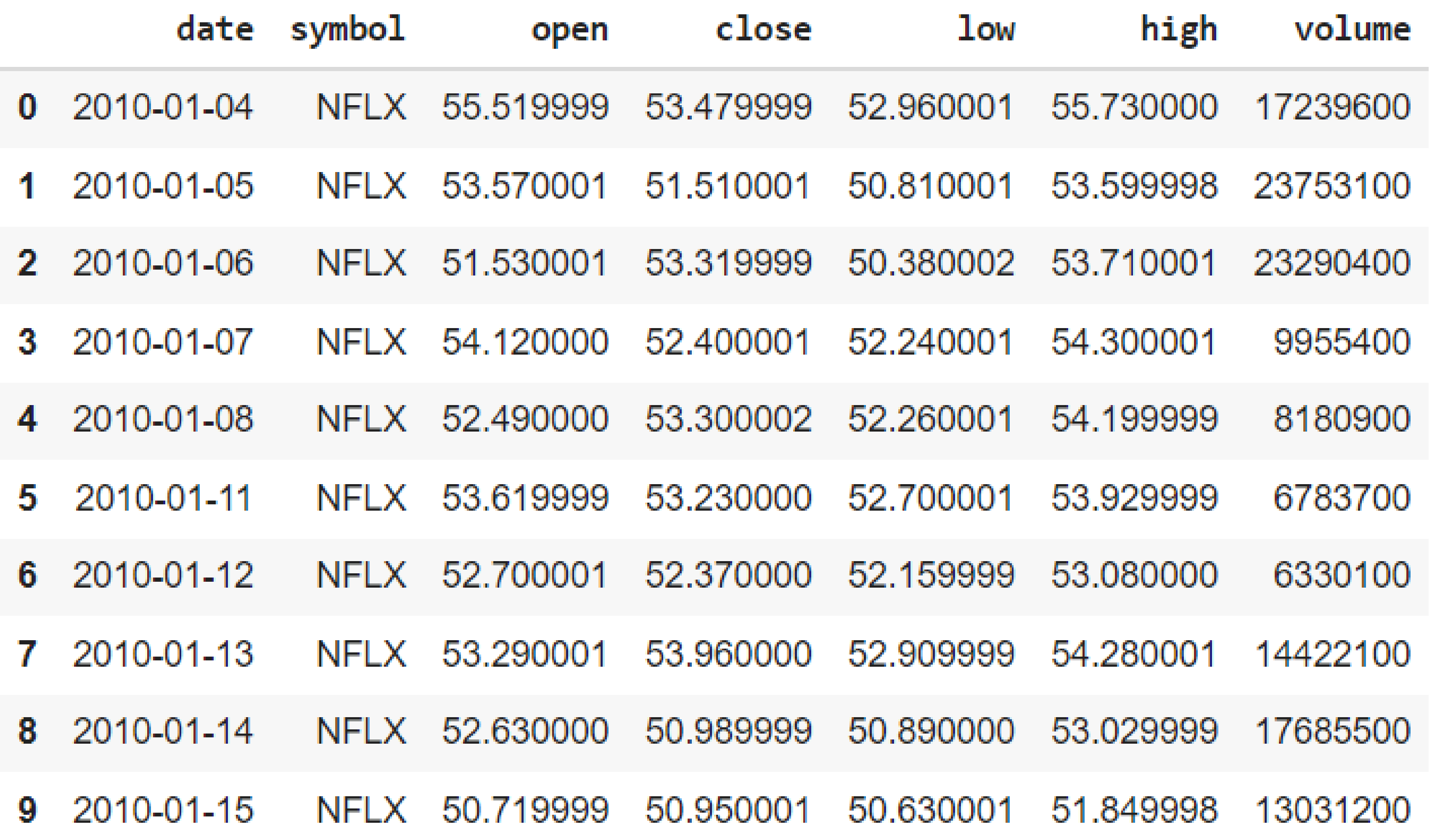
- **Regression is a supervised machine learning technique which is used to predict continuous values.**
- **The ultimate goal of the regression algorithm is to plot a best-fit line or a curve between the data**

- Predicting how the stock market will perform is one of the most difficult things to do.
- There are so many factors involved in the prediction – physical factors vs. psychological, rational and irrational behavior, etc. All these aspects combine to make share prices volatile and very difficult to predict with a high degree of accuracy.

STOCK MARKET PREDICTION

In this project, we try to implement a stock prediction model that helps to predict the price of a stock on nasdaq, for the upcoming days.
The dataset used contains 1762 closing prices and 7 columns
The various columns present in the dataset are:

- **date**- date on which the stock price is mentioned
- **symbol**-letters used to represent the stock on exchange
- **open**-price of the stock when the market opens
- **close**-price of the stock when the market closes
- **low**- lowest price of the stock throughout the trading time
- **high**- highest price of the stock throughout the trading time
- **volume** - how much the stock is being traded



	date	symbol	open	close	low	high	volume
0	2010-01-04	NFLX	55.519999	53.479999	52.960001	55.730000	17239600
1	2010-01-05	NFLX	53.570001	51.510001	50.810001	53.599998	23753100
2	2010-01-06	NFLX	51.530001	53.319999	50.380002	53.710001	23290400
3	2010-01-07	NFLX	54.120000	52.400001	52.240001	54.300001	9955400
4	2010-01-08	NFLX	52.490000	53.300002	52.260001	54.199999	8180900
5	2010-01-11	NFLX	53.619999	53.230000	52.700001	53.929999	6783700
6	2010-01-12	NFLX	52.700001	52.370000	52.159999	53.080000	6330100
7	2010-01-13	NFLX	53.290001	53.960000	52.909999	54.280001	14422100
8	2010-01-14	NFLX	52.630000	50.989999	50.890000	53.029999	17685500
9	2010-01-15	NFLX	50.719999	50.950001	50.630001	51.849998	13031200

STOCK MARKET PREDICTION USE CASE

Discount broker applications

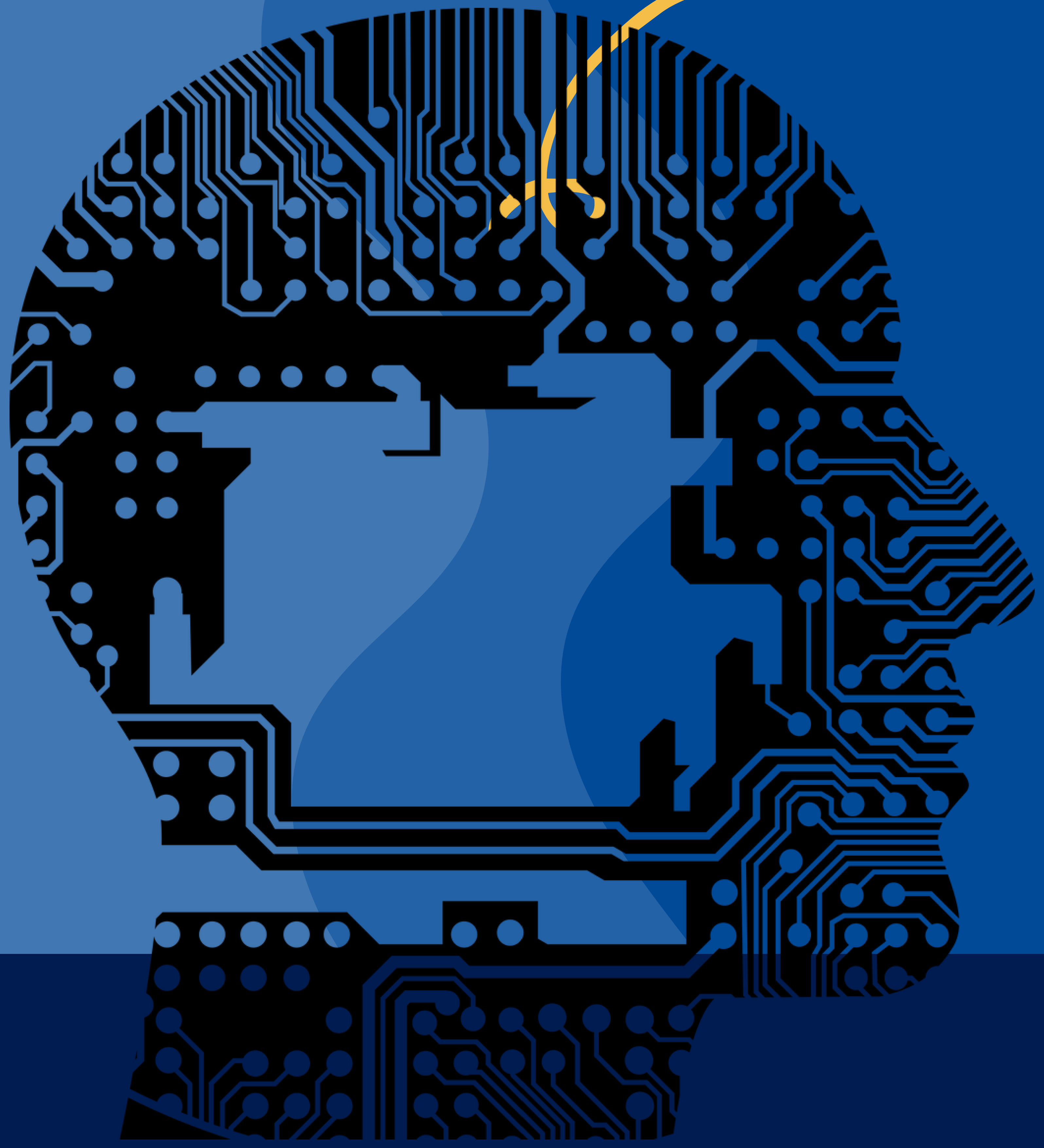
Many online broker applications use stock market prediction

News Agency

Finacial news channels use it to give suggestions

Day Traders

Professional traders depend on hints from ML powered stock predictions

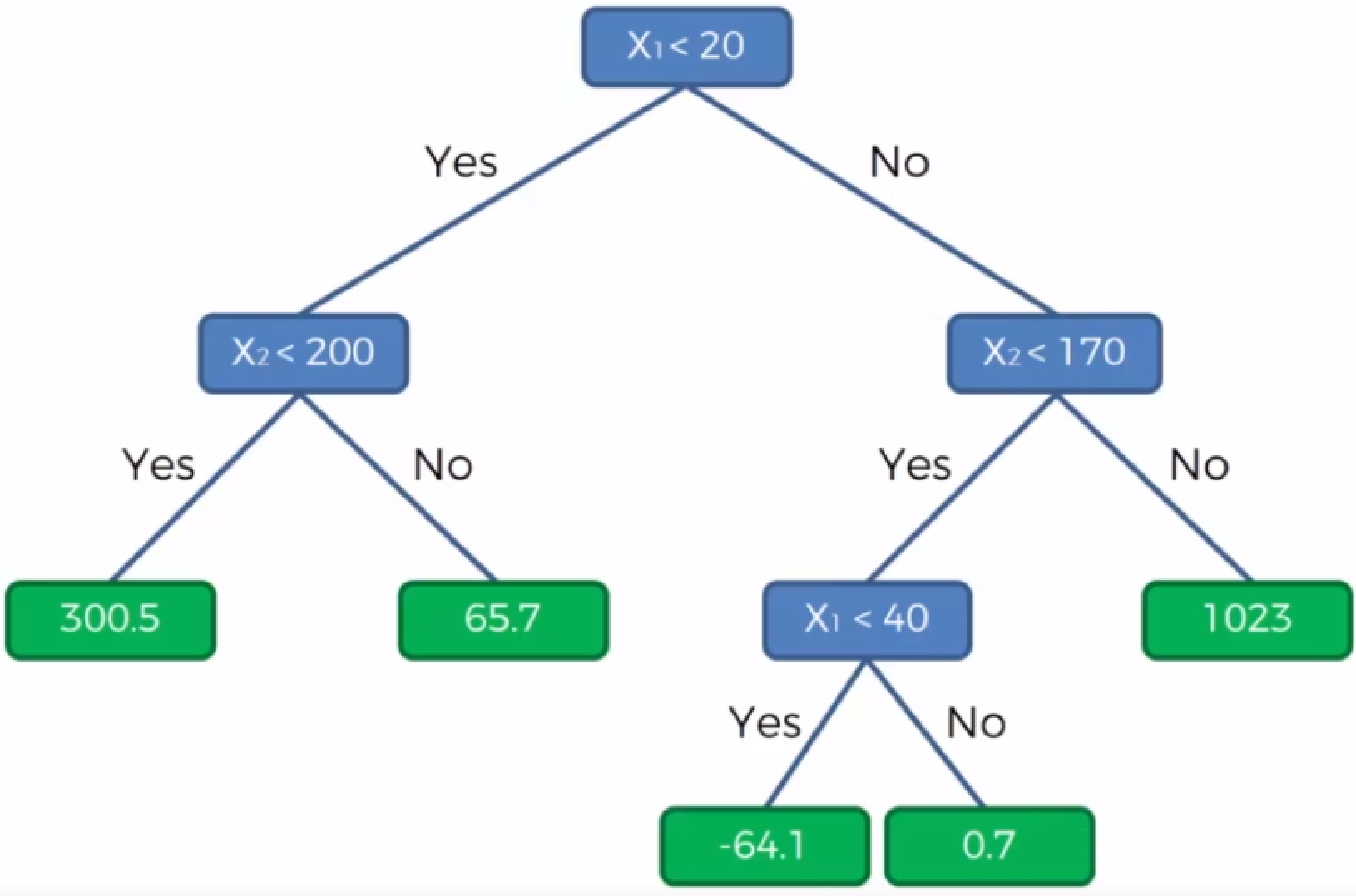


Decison Tree

- Decision tree builds regression or classification models in the form of a tree structure.
- It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed.
- The final result is a tree with decision nodes and leaf nodes.

How does it work ?

- Decision tree builds regression or classification models in the form of a tree structure.
- It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed.
- The final result is a tree with decision nodes and leaf nodes.
- A decision node (e.g., Outlook) has two or more branches (e.g., Sunny, Overcast and Rainy), each representing values for the attribute tested.
- Leaf node (e.g., Hours Played) represents a decision on the numerical target. The topmost decision node in a tree which corresponds to the best predictor called root node. Decision trees can handle both categorical and numerical data.



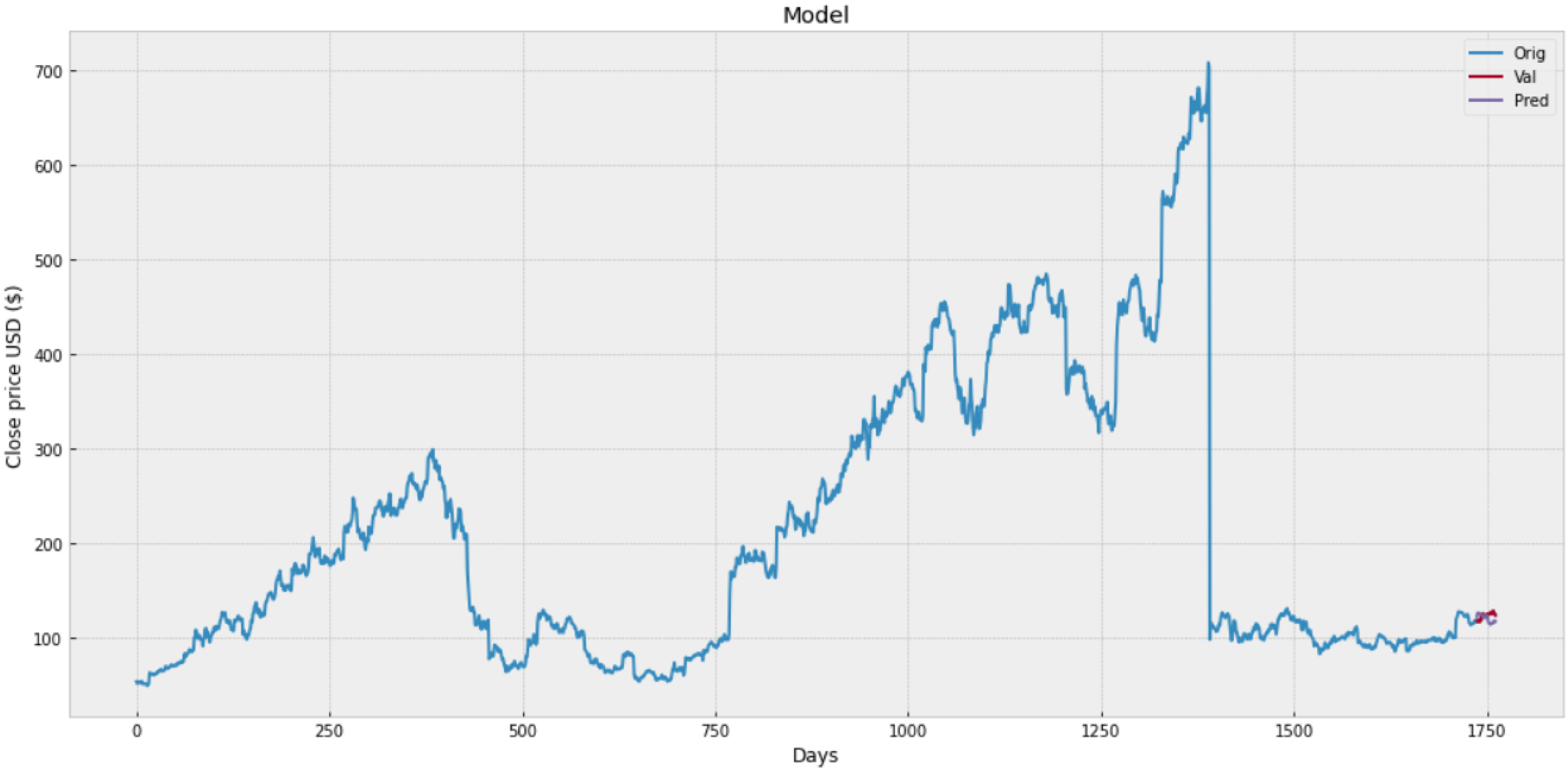
Advantages :

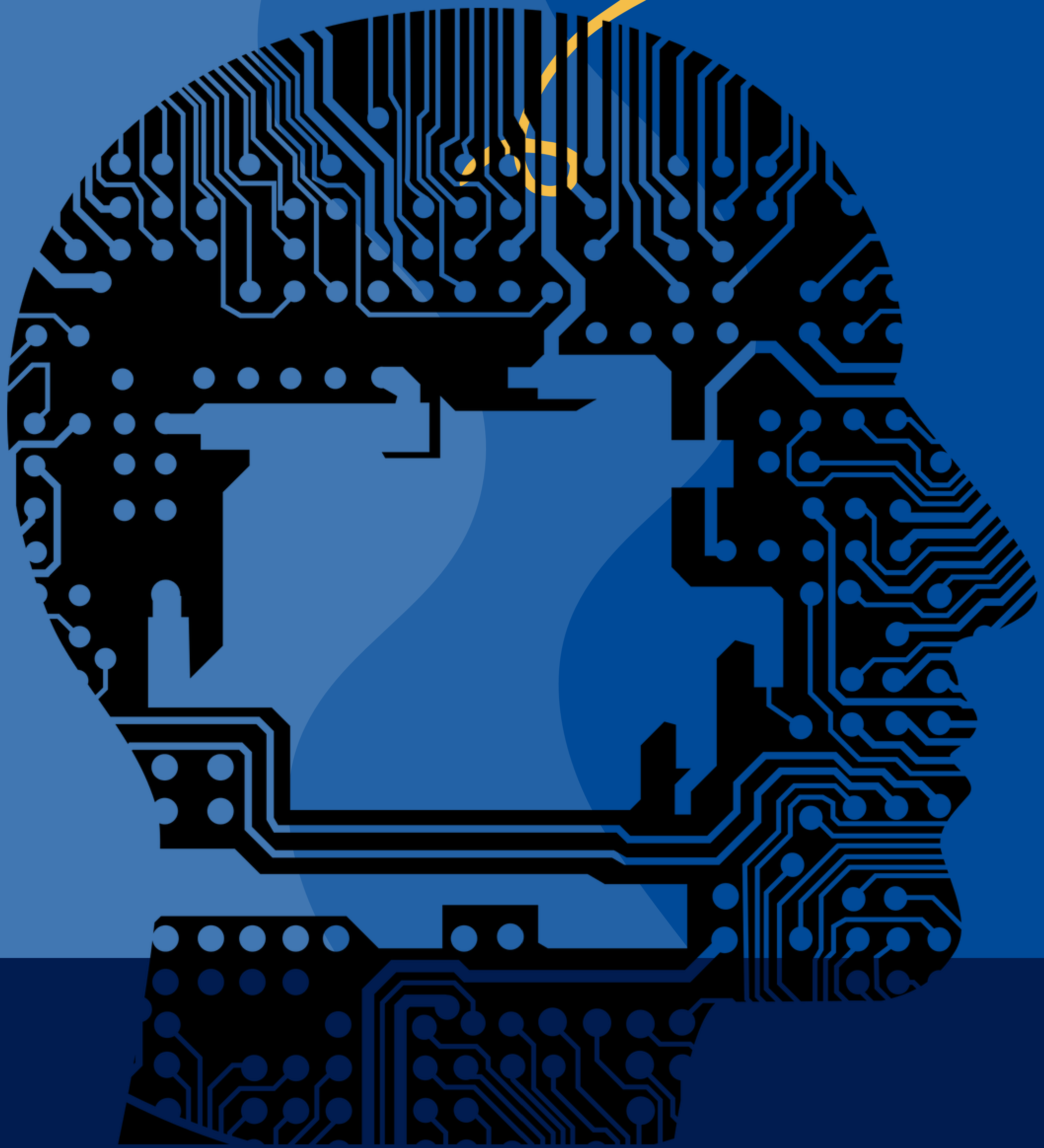
1. Clear Visualization: The algorithm is simple to understand, interpret and visualize as the idea is mostly used in our daily lives. Output of a Decision Tree can be easily interpreted by humans.
2. Simple and easy to understand: Decision Tree looks like simple if-else statements which are very easy to understand.
3. Decision Tree can be used for both classification and regression problems.
4. Decision Tree can handle both continuous and categorical variables

Disadvantages :

1. Overfitting: This is the main problem of the Decision Tree. It generally leads to overfitting of the data which ultimately leads to wrong predictions. In order to fit the data (even noisy data), it keeps generating new nodes and ultimately the tree becomes too complex to interpret. In this way, it loses its generalization capabilities. It performs very well on the trained data but starts making a lot of mistakes on the unseen data.
2. High variance: As mentioned in point 1, Decision Tree generally leads to the overfitting of data. Due to the overfitting, there are very high chances of high variance in the output which leads to many errors in the final estimation and shows high inaccuracy in the results. In order to achieve zero bias (overfitting), it leads to high variance.

PREDICTION USING DECISION TREE





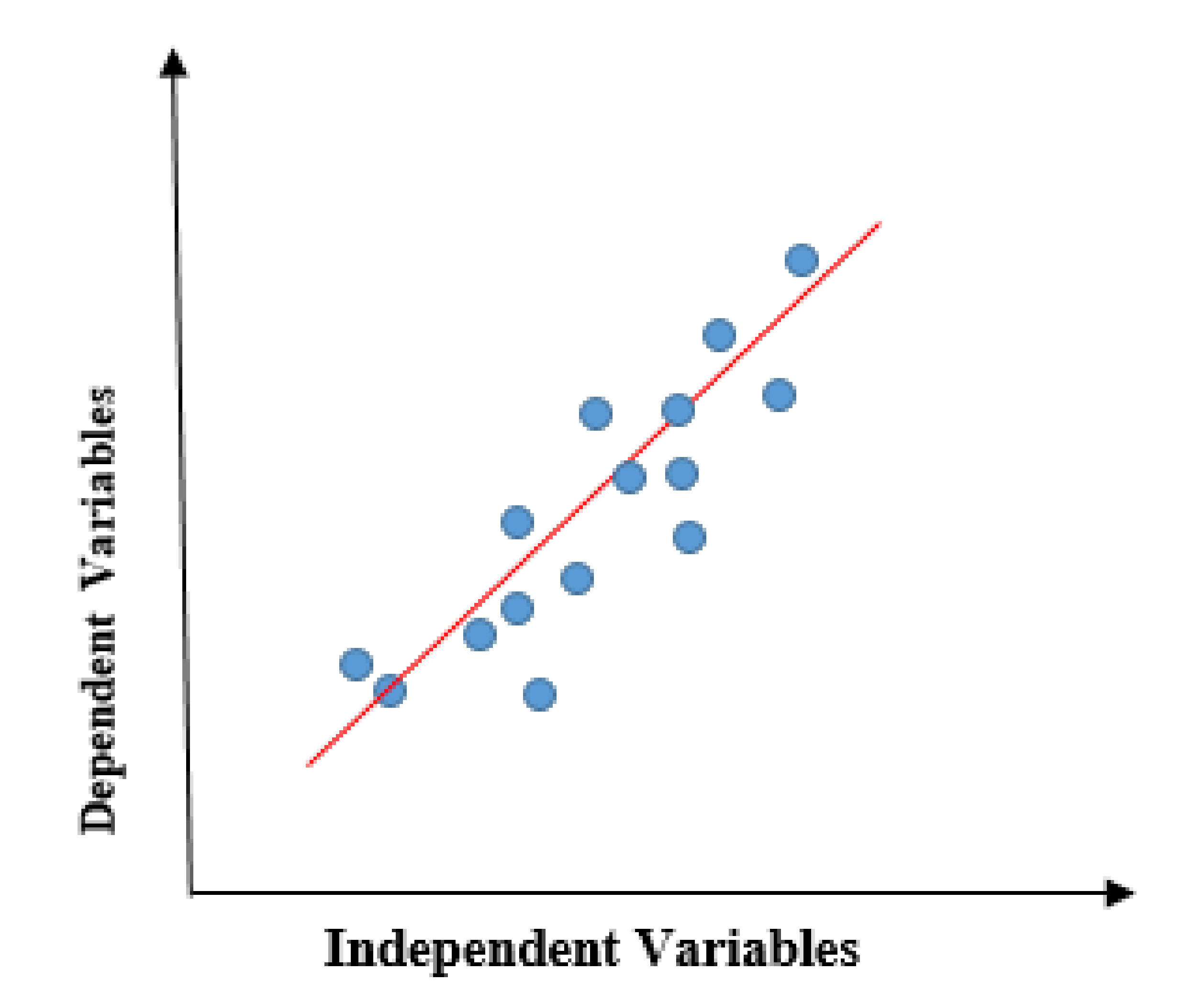
LINEAR REGRESSION

- Linear Regression is a machine learning algorithm based on supervised learning
- Linear regression is one of the easiest and most popular Machine Learning algorithms.
- Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (x) variables, hence called as linear regression.
- It is mostly used for finding out the relationship between variables and forecasting. .

HOW THEY WORK?

- The linear regression model gives a sloped straight line describing the relationship within the variables.
- When working with linear regression, our main goal is to find the best fit line that means the error between predicted values and actual values should be minimized.
- To calculate best-fit line linear regression uses a traditional slope-intercept form given below
- In the below formula

1. y= Dependent Variable.
2. x= Independent Variable.
3. a1= intercept of the line.
4. a2 = slope of the line

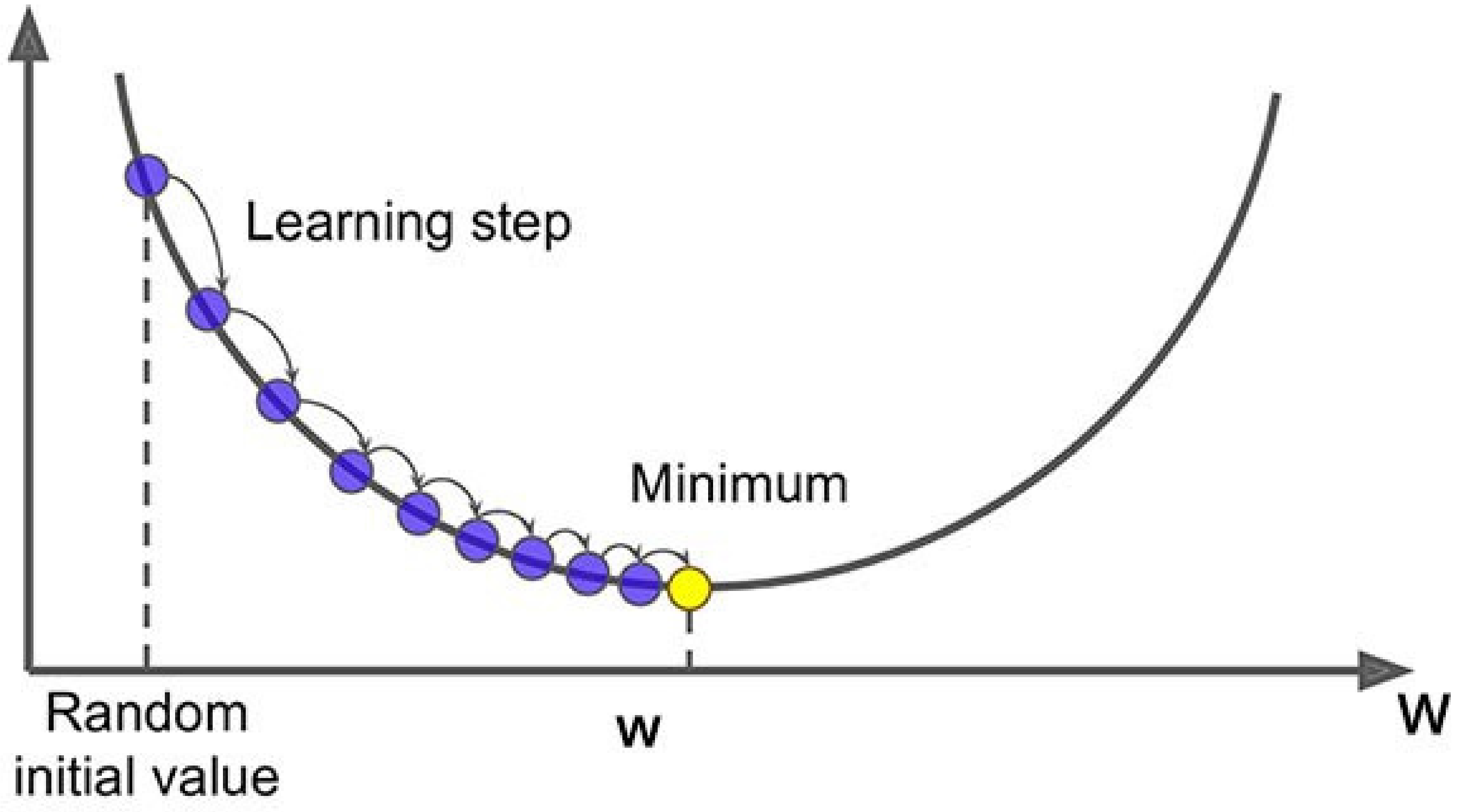


$$y = \theta_1 \quad + \quad \theta_2 . x$$

SO HOW DO WE FIND A1 AND A2

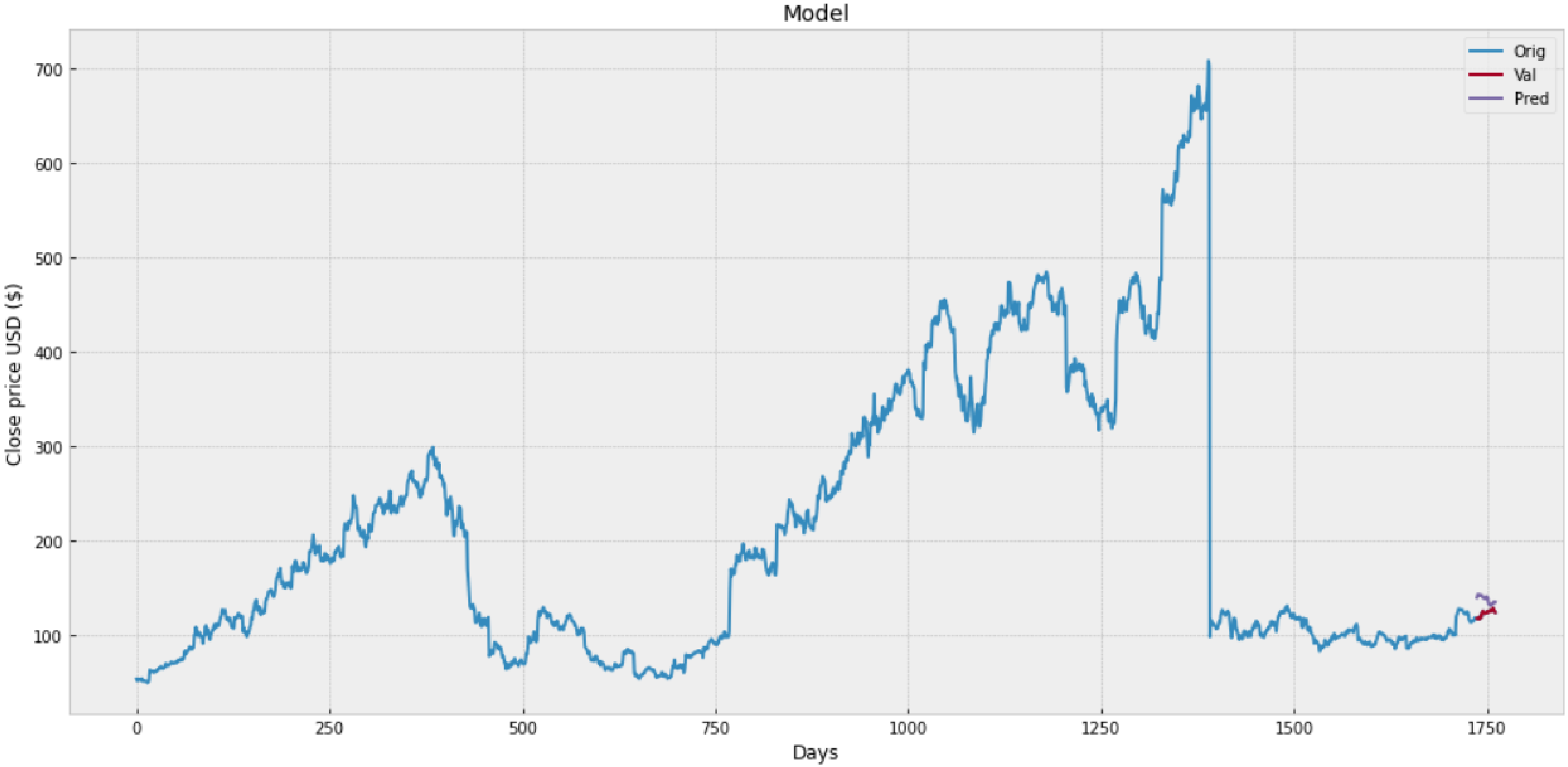
$$MSE = \frac{1}{N}\sum_{i=1}^{n}(y_i - (mx_i + b))^2$$

- The cost function helps to figure out the best possible best fit line for the data points.
- In Linear Regression, Mean Squared Error (MSE) cost function is used, which is the average of squared error that occurred between the predicted values and actual values.
- The cost function should be as minimum as possible.Thats where gradient descent comes in.
- Gradient descent is used to minimize the MSE by calculating the gradient of the cost function.



- A learning rate is used as a scale factor and the coefficients are updated in the direction towards minimizing the error.
- it does only a few calculations far from optimal sol and increases the no of calculations close to the optimal value
- It is done by a random selection of values of coefficient and then iteratively update the values to reach the minimum cost function.

PREDICTION USING LINEAR REGRESSION



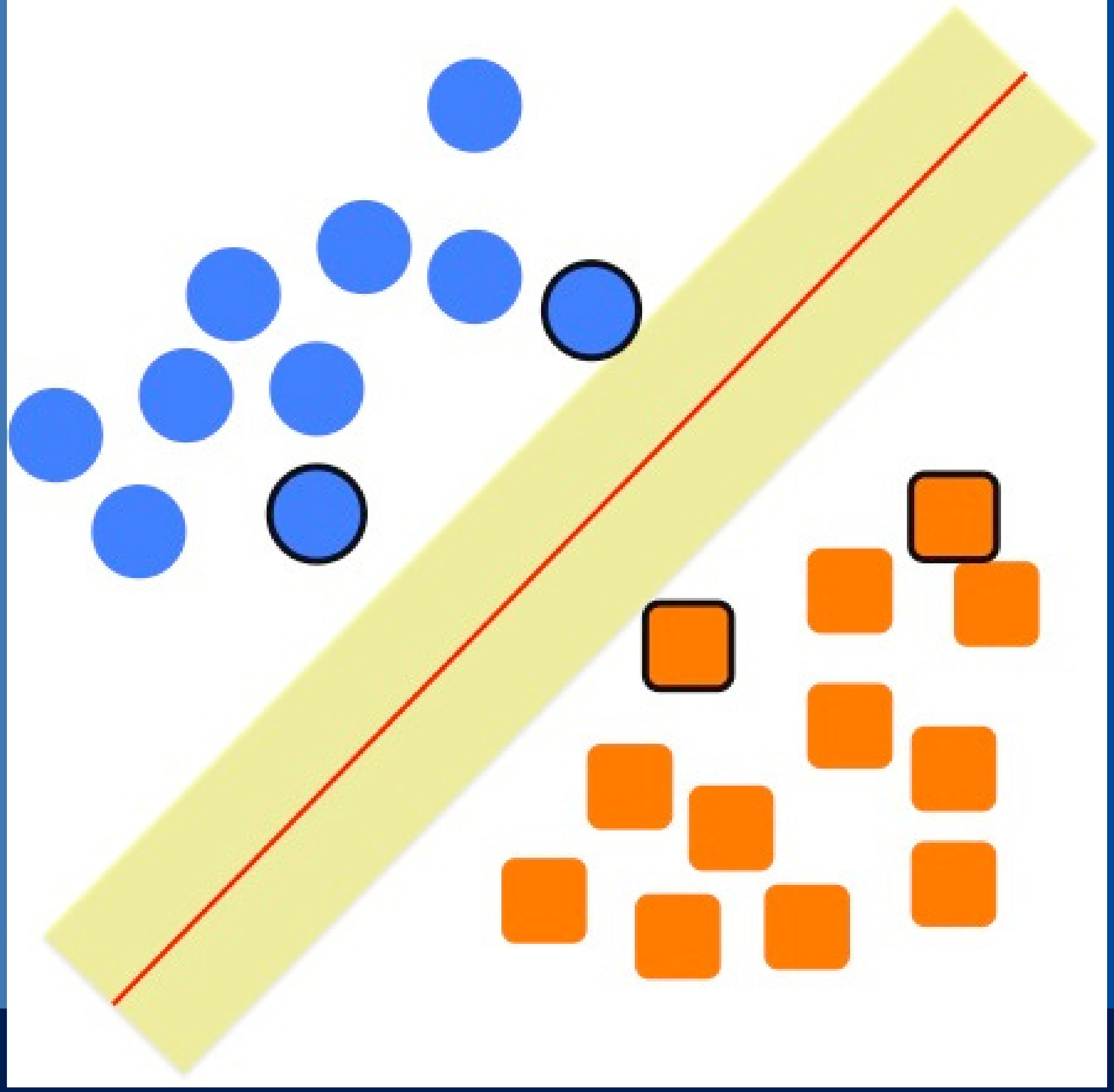
Advantages :

1. Linear Regression is simple to implement and easier to interpret the output coefficients.When you know the relationship between the independent and dependent variable have a linear relationship,
2. Linear Regression is susceptible to over-fitting but it can be avoided using some dimensionality reduction techniques, regularization (L1 and L2) techniques and cross-validation.

Disadvantages :

1. On the other hand in linear regression technique outliers can have huge effects on the regression and boundaries are linear in this technique.
2. Diversely, linear regression assumes a linear relationship between dependent and independent variables. That means it assumes that there is a straight-line relationship between them. It assumes independence between attributes.
3. But then linear regression also looks at a relationship between the mean of the dependent variables and the independent variables. Just as the mean is not a complete description of a single variable, linear regression is not a complete description of relationships among variables.

SUPPORT VECTOR MACHINE



- Support Vector Machine (SVM) is a very popular Machine Learning algorithm that is used in both Regression and Classification.

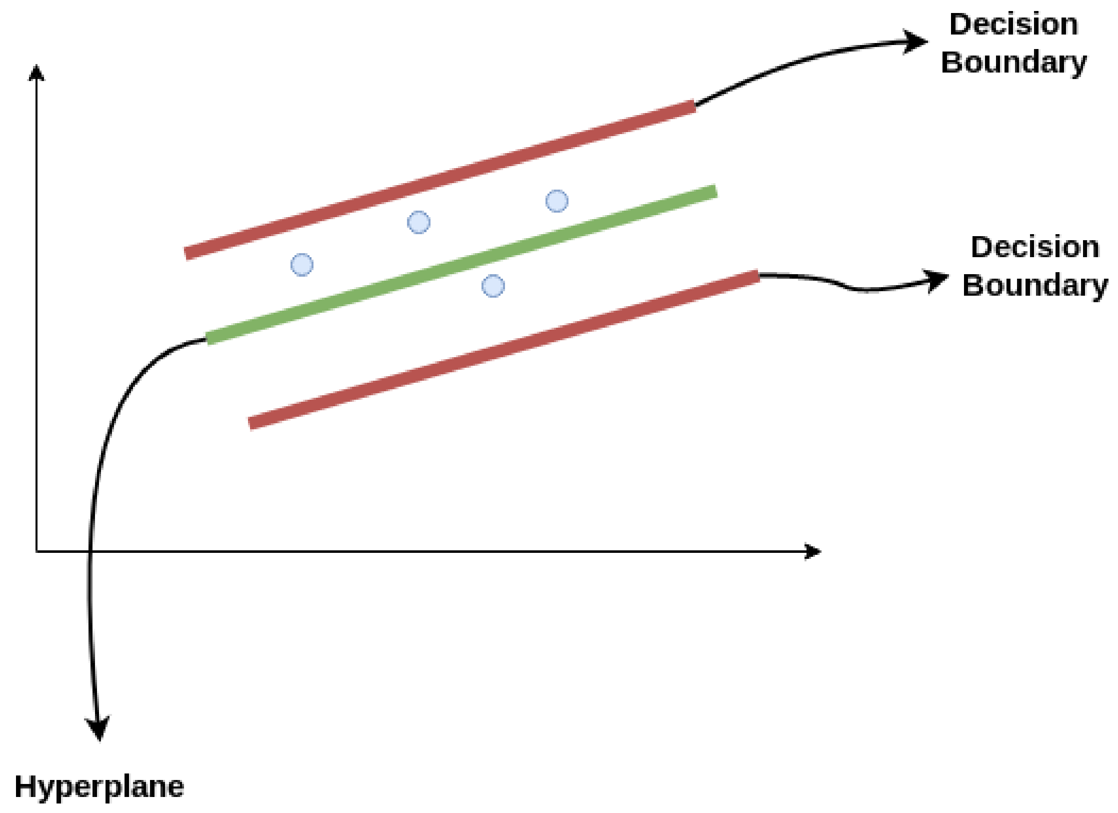
A few important parameters of SVM are:

- **Kernel**: A kernel helps us find a hyperplane in the higher dimensional space without increasing the computational cost.
- **Hyperplane**: This is basically a separating line between two data classes in SVM. But in Support Vector Regression, this is the line that will be used to predict the continuous output
- **Decision Boundary:** A decision boundary can be thought of as a demarcation line (for simplification) on one side of which lie positive examples and on the other side lie the negative examples.

SUPPORT VECTOR REGRESSION

- The problem of regression is to find a function that approximates mapping from an input domain to real numbers on the basis of a training sample.

- Consider the two red lines in the diagram as the decision boundary and the green line as the hyperplane.
- Our objective is to consider the points that are within the decision boundary line.
- Our best fit line is the hyperplane that has a maximum number of points.
- Consider these lines as being at any distance, say ‘a’, from the hyperplane.
- So, these are the lines that we draw at distance ‘+a’ and ‘-a’ from the hyperplane.
- This ‘a’ in the text is basically referred to as **epsilon**.
- Unlike other Regression models that try to minimize the error between the real and predicted value, the SVR tries to fit the best line within a threshold value (Distance between hyperplane and boundary line), a.



- Assuming that the equation of the hyperplane is as follows:
- Y = wx+b (equation of hyperplane)
- Then the equations of decision boundary become:
  i. wx+b= +a
  ii. wx+b= -a
- Thus, any hyperplane that satisfies our SVR should satisfy:
  iii. -a < Y- wx+b < +a

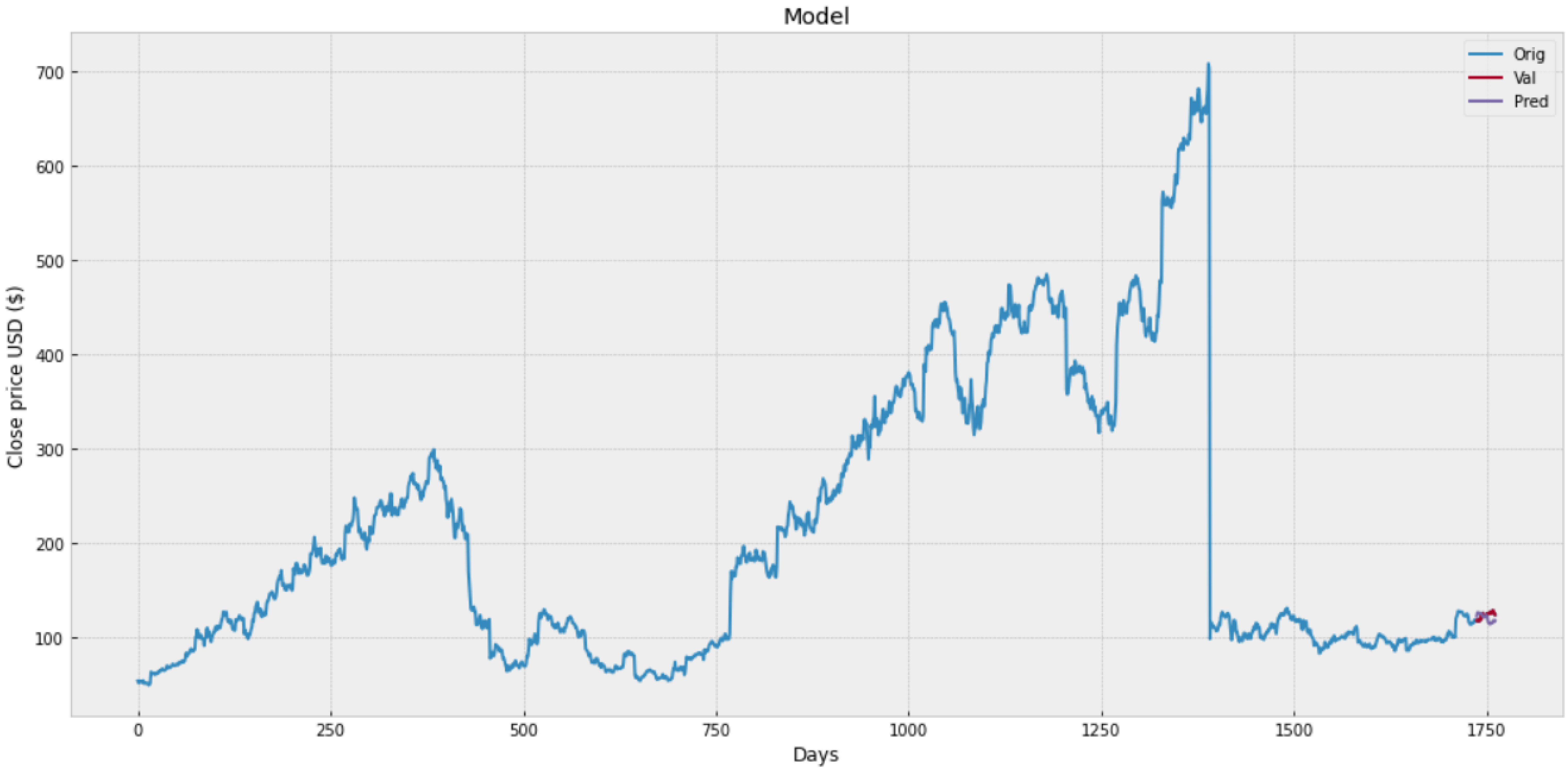
ADVANTAGES :

1. SVM works relatively well when there is a clear margin of separation between classes.
2. SVM is more effective in high dimensional spaces.
3. SVM is effective in cases where the number of dimensions is greater than the number of samples.
4. SVM is relatively memory efficient

DISADVANTAGES

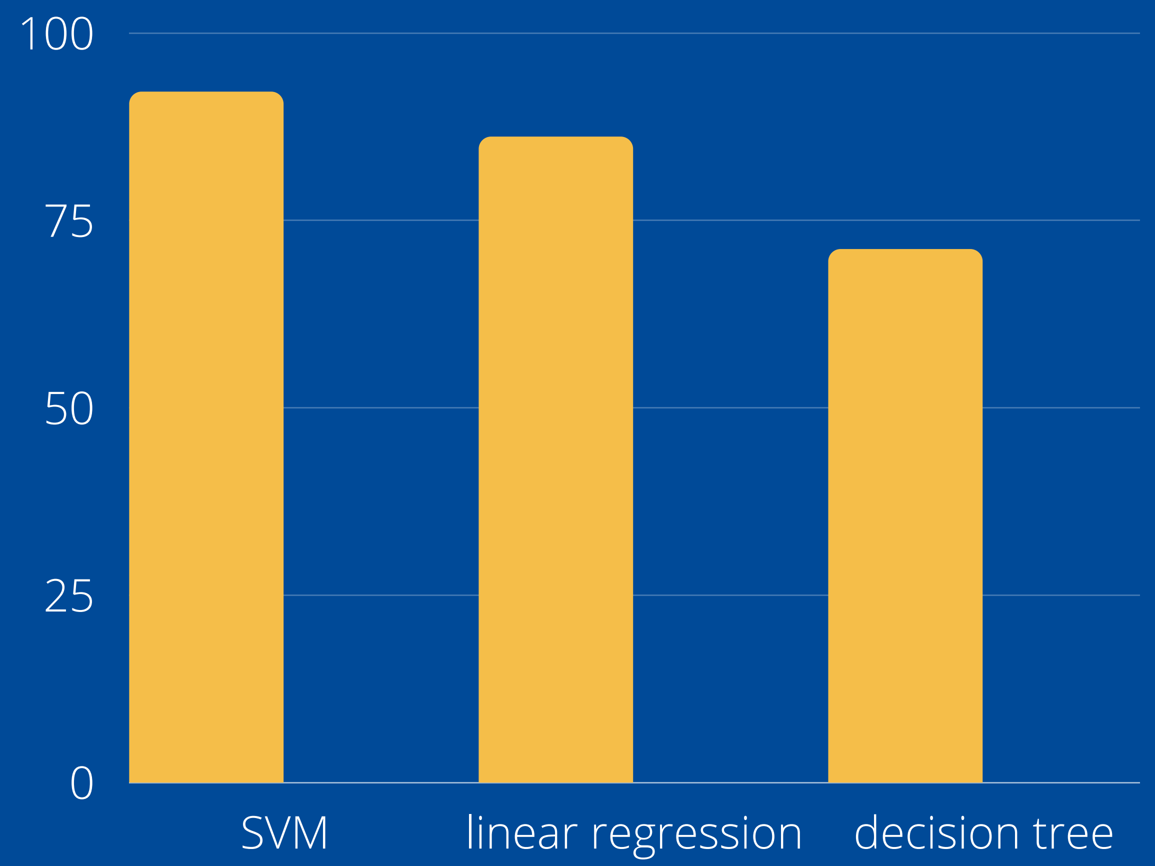
1. SVM algorithm is not suitable for large data sets.
2. SVM does not perform very well when the data set has more noise i.e. target classes are overlapping.
3. In cases where the number of features for each data point exceeds the number of training data samples, the SVM will underperform.
4. As the support vector classifier works by putting data points, above and below the classifying hyper plane there is no probabilistic explanation for the classification.

PREDICTION USING SVM



CONCLUSION

ALGORITHM	ACCURACY
Linear Regression	86%
Decision Tree	71%
Support Vector Machine	92%



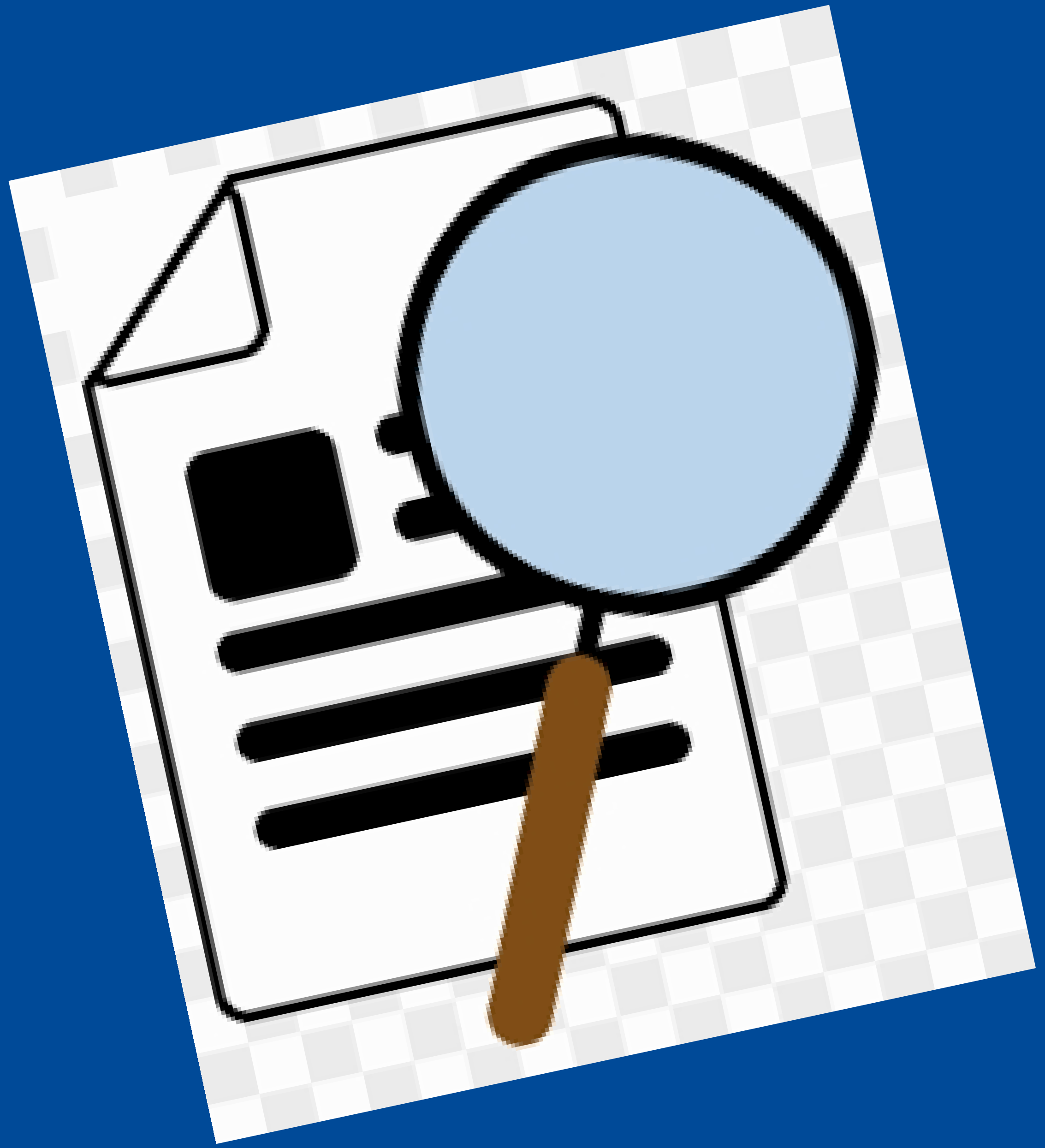
The **SVR Algorithm** gives the highest accuracy, which is **92%**.
We conclude that **the SVR Algorithm** is the best fit, out of the 3 algorithms, for this particular dataset because:

- The **SVR model** performs lower computation compared to other regression techniques and its implementation is comparitively easy.
- **Support vector regression** algorithm is a huge improvement over linear regression, as it allows us to build non-linear models and gives us the control over the flexibility vs. robustness of the model.
- Altough it is difficult to understand and interpret the SVM model compared to Decision tree as SVM is more complex but it is memory systematic and more productive in high dimensional spaces.

***Note**: SVM model is not suitable for large data sets, i.e. when the target classes are overlaping in a data set.*

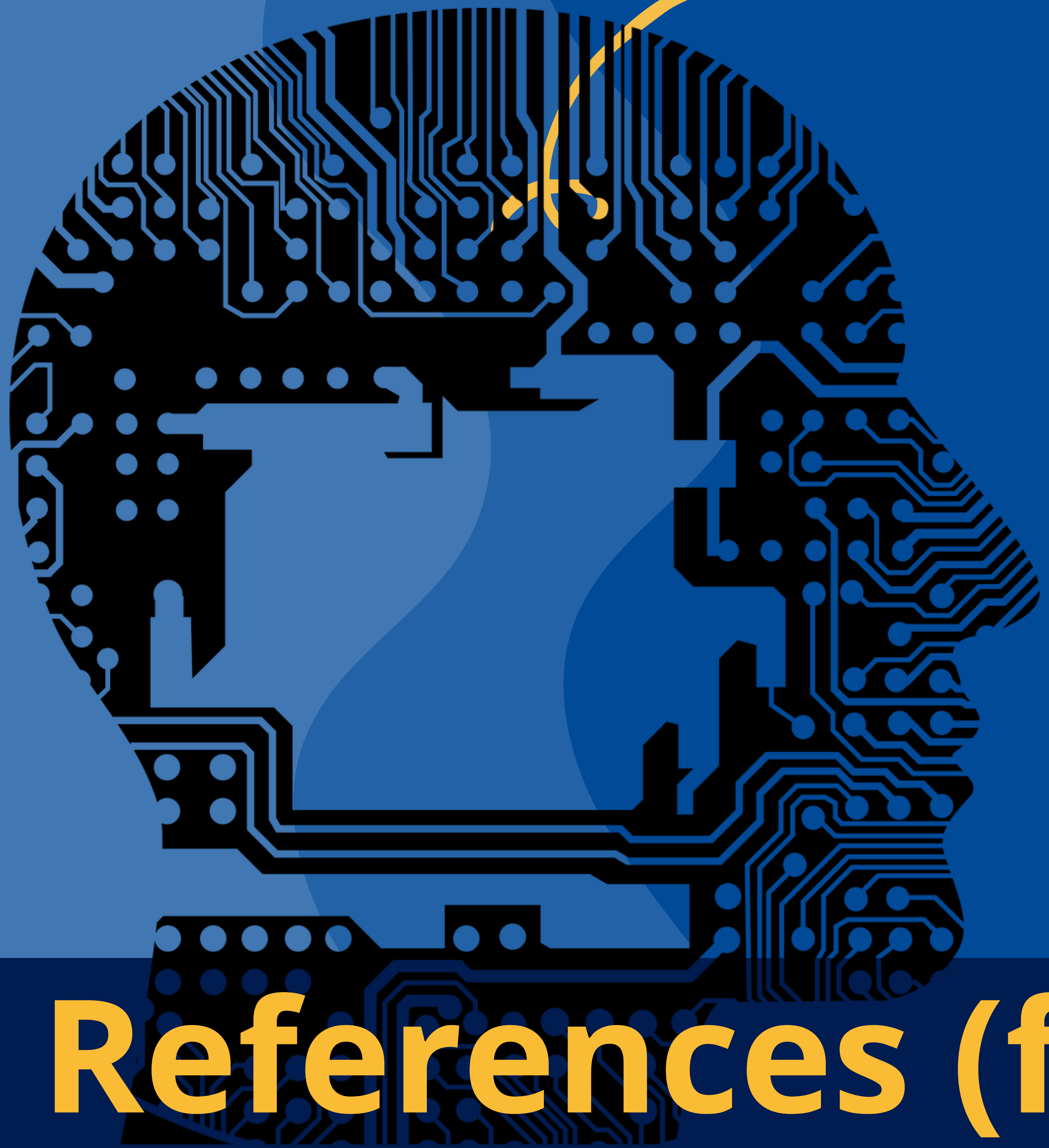
REFERENCES

1. https://monkeylearn.com/
2. https://www.analyticsvidhya.com/
3. Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems (Book)
4. https://arxiv.org/ftp/arxiv/papers/1601/1601.06971.pdf
5. https://machinelearningmastery.com/linear-regression-for-machine-learning/



References for Viva:

- Raw data set from Kaggle: https://www.kaggle.com/dgawlik/nyse
- Cleansed data set: https://drive.google.com/drive/folders/1wfsUjrn11a1OWC4itHx-3es9DKVRvE2_
- The algorithm (via google colab): https://colab.research.google.com/drive/1k0whRXEqysWtiOh0iW7IwCphc4pkie46?usp=sharing



References (for algorithm):

- Raw data set from Kaggle: https://www.kaggle.com/dgawlik/nyse
- Cleansed data set: https://drive.google.com/drive/folders/1wfsUjrn11a1OWC4itHx-3es9DKVRvE2_
- The algorithm (via google colab): https://colab.research.google.com/drive/1k0whRXEqysWtiOh0iW7IwCphc4pkie46?usp=sharing