Supervised Learning Algorithm of Classification on Basis of Ranges

Ahmer Hasan[1] Usman Khan[2]

Abstract

A supervised learning algorithm of classification which is implemented on linear data. Ranges/ Boundaries of each group are calculated. As the data is linear the number of ranges will be equal to a number of classes. New input data will be tested by in which range/boundary it lies. And the range in which it lies, the class of that specific range will be assigned to the new data. The experimental results show that the accuracy depends on the linearity of the data. Therefore our model can show accuracy up to 99.99% also if the data is completely linear in nature.

Keyword: Classification,Linear Data,Supervised Learning,Unsupervised Learning, Ranges, Algorithms, Clustering, Decision Tree

1 Introduction

Machine learning [15] enriches us with many of its learning techniques, some of which includes the technique of classification [1]. Supervised learning [2] [17] includes data which is classified or you can say that each data is labeled with a class instance. Unsupervised learning [2] [17] includes data which is not classified or unlabeled data. Data which is classified is trained for predicting the class of the new incoming inputs. And the data which is not classified is grouped/ clustered [5] [17] in unsupervised learning.

There are many supervised and unsupervised learning classification algorithms which include K-Nearest Neighbor [3] for supervised learning and K-Means [4] [17] for unsupervised learning.

K-Nearest Neighbor is an example of a supervised learning algorithm for classification. The new features checks 'k' number of data points (neighbors) closest to them by calculating their distances (distances are usually calculated by Manhattan [6] or Euclidean [6] methods). The most repeated classes among the closest 'k' number of neighbors are the predicted class for the new input.

K-Means is an example of an unsupervised learning algorithm for clustering. The number of 'k' decides the number of centroids/means/clusters of the data. The 'k' no of means are selected from the data randomly. Each mean point or centroid has a distance with each other data point. The data point which is most near among all the centroids is thenconsidered to be the part of the cluster or group of the respective centroid/mean point. New means are calculated again but this time it is among the groups, i.e each group will a have mean (a new centroid). Now the distance is calculated again of each data point with the new centroids, and each data point is assigned to the nearest centroid. This process continues until any data point does not change the group.

[1] *Karachi Institute of Economics & Technology, Karachi |ahmerhasan123@yahoo.com*
[2] *Karachi Institute of Economics & Technology, Karachi |usman@pafkiet.edu.pk*

The structure of the paper is as follows. In section 2, literature review; in section 3, explanation of the data and feature selection; in section 4, we propose our classification algorithm; section 5, clustering of an unsupervised data using K-Means; section 6, practical implementation and the results; Finally, the conclusion of the paper is shown in section 7.

2 Literature Review

A Machine Learning

It is one of the types of Artificial Intelligence [8], which works on the development and designing of algorithms which uses past data that provide computers ability to predict or make decisions by the use of statistical/mathematical methods [9].

The historical data is used for the training part of the algorithms, which is actually making the computer able to learn from data. The new inputs are tested on behalf of the empirical data, these new inputs are called testing data.

The measures of accuracy [10] of any learning algorithm depend on many factors which include features extraction [11], features normalization [12], selecting a suitable algorithm according to the nature of the data [11], etc.

Following are four types of learning algorithms which depend on the nature of the data:

1) Supervised Learning: It consists of labeled data. Predictions are made by regression models [17] and classifications are made by classifiers.

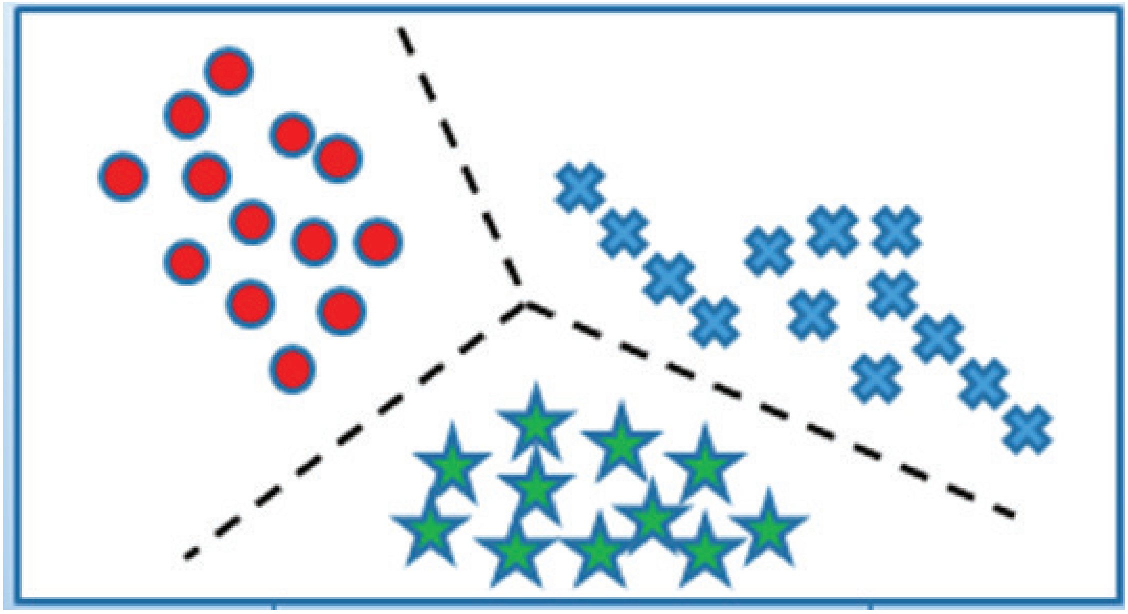


Figure 1: Graphical View of Supervised Learning Data

2) ***Unsupervised Learning:*** It consists of unlabeled data. Clustering, probability distribution estimation, finding an association (in features) and dimension reductions are used in unsupervised learning.

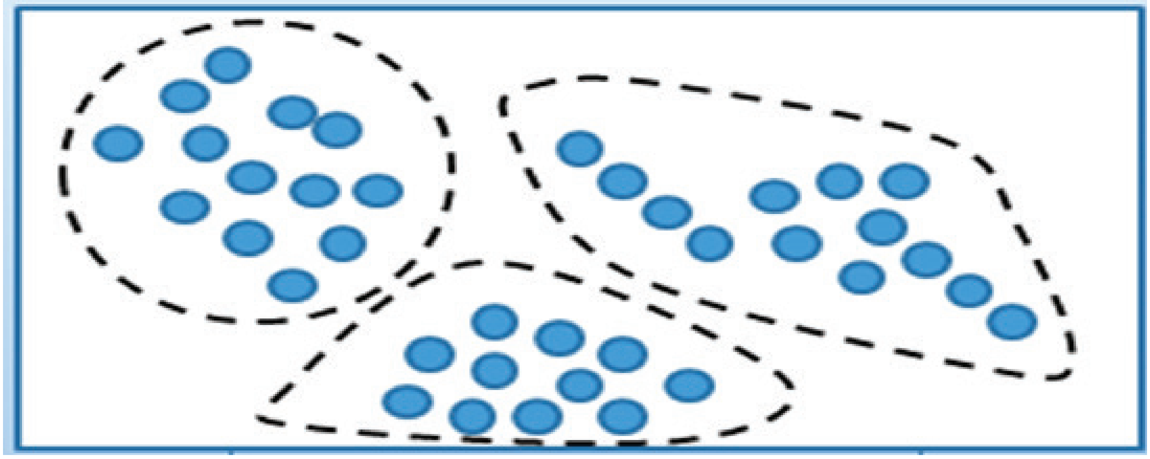


Figure 2: Graphical View of Unsupervised Learning Data

3) ***Semi-Supervised Learning:*** The data is partially labeled in nature. Semi-supervised learning [13] is a mixture of supervised and unsupervised learning.

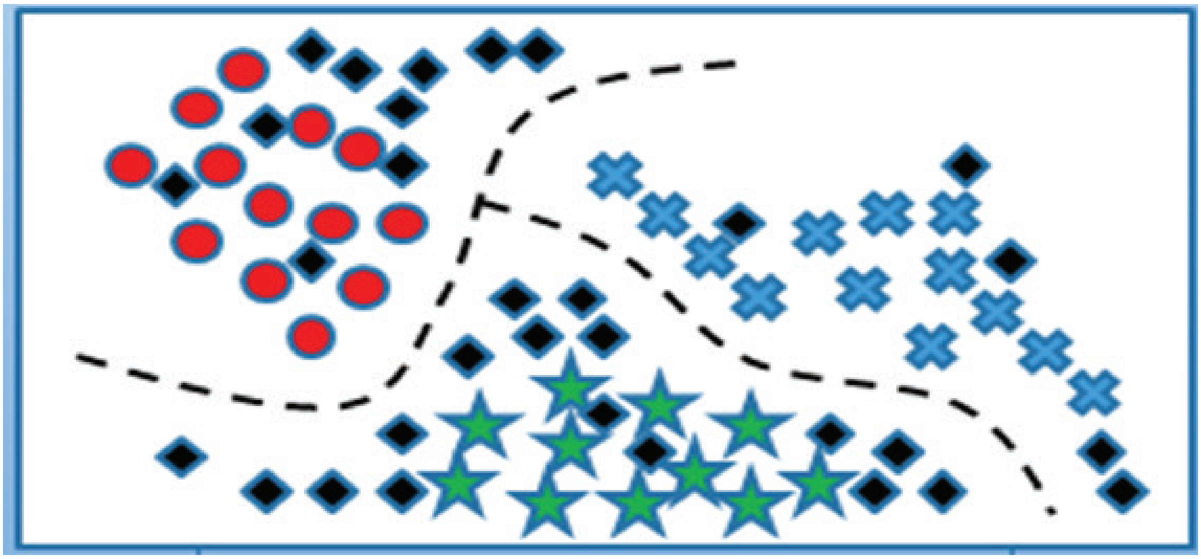


Figure 3: Graphical View of Semi-Supervised Learning Data

4) ***Reinforcement Learning [14][17]:*** It is used in decision making. For example chess game or the ability of a robot for making efficient actions/decisions.

B Supervised Learning

This is one of the four types of machine learning. In supervised learning, the nature of the data defines output on behalf of the feature/s. These outputs vary into two types 1)Classes(discrete labels) 2)Regression(Real values). The data in which there are classes assigned for every input is known as classified data i.e the data is in classes/groups, therefore classification algorithms [16][17] are implemented on this kind of data. The data which has output as decimal values are handled by regression algorithms [17] for predictions.

Table 1: Classification Data

Return July	Return Aug	Return Sep	Return Oct	Return Nov	Positive Dec
-0.020517029	0.024675868	-0.020408163	-0.173317684	-0.025385313	0
-0.025321312	0.211290002	-0.580003262	-0.267141251	-0.151234568	0
-0.135384615	0.033391916	0	0.091695502	-0.059561129	0
-0.094	0.095290252	0.056680162	-0.096339114	-0.40511727	1
0.35530086	0.056842105	0.033602151	0.03626943	-0.085305106	1
0.274446938	0.538343949	0.127068167	-0.171428571	-0.195374525	1

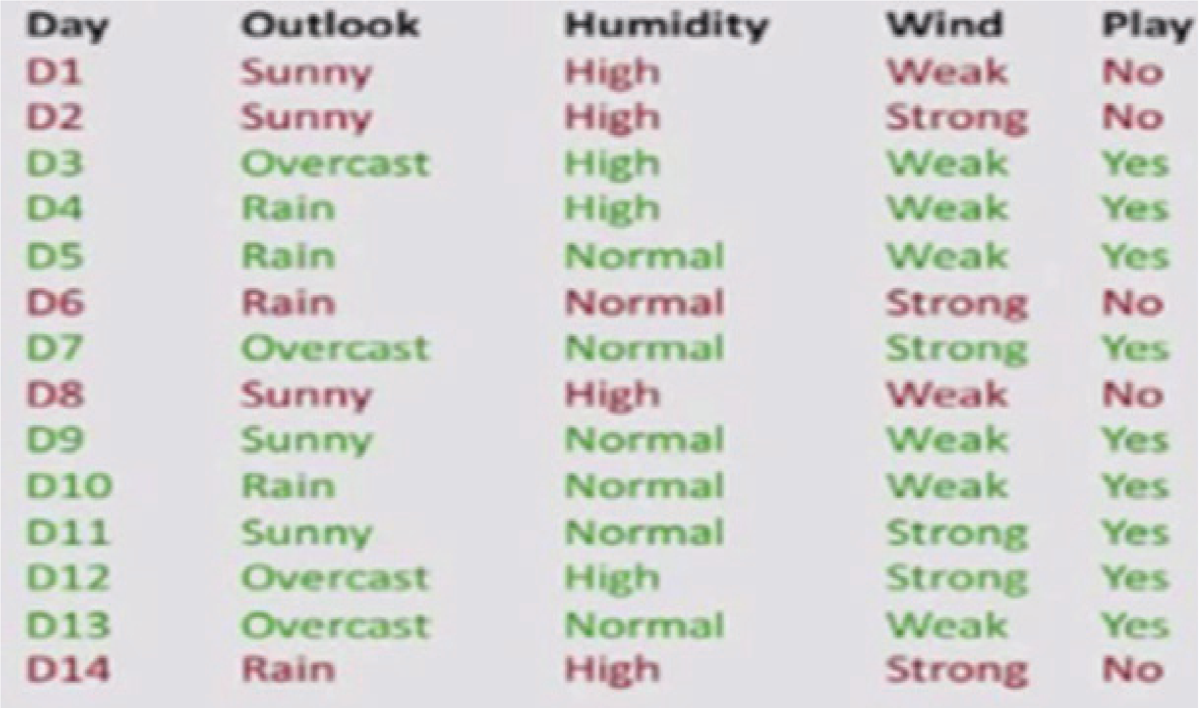
The “PositiveDec” column denotes the binary classes [17]of the features.

Table 2: Regression Data

PT08.S2(NMHC)	NOX(GT)	PT03.S3(NOX)	NO2(GT)	PT08.S4(NO2)	PT08.S5(03)	T	RH	AH
1046	166	1056	113	1692	1268	13.6	48.9	0.7578
955	103	1174	92	1559	972	13.3	47.7	0.7255
939	131	1140	114	1555	1074	11.9	54.0	0.7502
948	172	1032	122	1584	1203	11.0	60.0	0.7867
836	131	1205	116	1490	1110	11.2	59.6	0.7888
750	89	1337	96	1393	494	11.2	59.2	0.7848
690	62	1462	77	1333	733	11.3	56.8	0.7603

The “AH” column is the output on behalf of the values of the rest of the columns in that single row. There are two types of regression models namely simple and multiple which comprise of two types of nature of data, linear and non-linear.

The data which is in textual format, algorithms like decision tree [17] is implemented on them. This nature of data lies in the context of classification but the classes, as well as the features, are in a textual format as shown on next page:



Day	Outlook	Humidity	Wind	Play
D1	Sunny	High	Weak	No
D2	Sunny	High	Strong	No
D3	Overcast	High	Weak	Yes
D4	Rain	High	Weak	Yes
D5	Rain	Normal	Weak	Yes
D6	Rain	Normal	Strong	No
D7	Overcast	Normal	Strong	Yes
D8	Sunny	High	Weak	No
D9	Sunny	Normal	Weak	Yes
D10	Rain	Normal	Weak	Yes
D11	Sunny	Normal	Strong	Yes
D12	Overcast	High	Strong	Yes
D13	Overcast	Normal	Weak	Yes
D14	Rain	High	Strong	No

Figure 4: Textual Data

The "Play" column shows the decisions on behalf of the features. A decision tree for the above data is shown below:

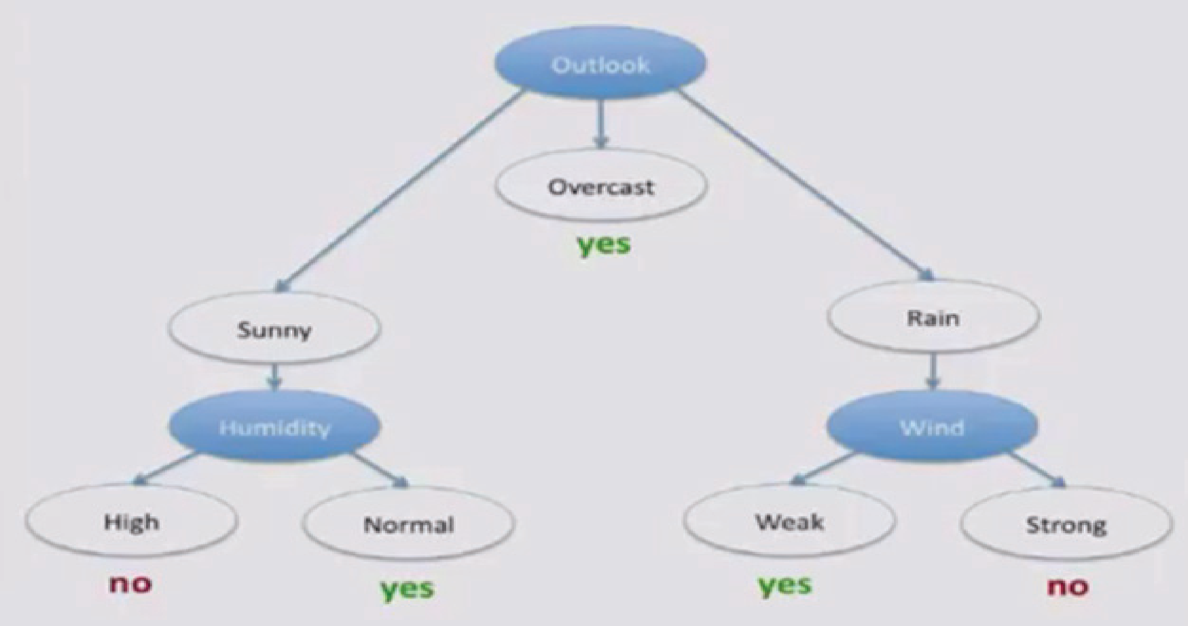


Figure 5: The Decision Tree of theTextual Data

3 Explanation & Filtering of Data

A Explanation of Data

This is the 5 years past data of Newyork (NYSE: TheNewyork Stock Exchange) stock market from date 2/2013 to 1/2018. There are 505 different stocks. Features include Date (current date of the stock price), Open (price of the stock at market open), High (the highest price reached in that day), Low (lowest price of that day), Close (price of the last trade of the stock), Volume (number of stocks traded) & Name (stock's ticker name). We are going to use the first 5 stocks from the data.

B Feature Selection

There are features which need to filter out as they contribute no benefit to our training, some of them will disturb the linear nature of the data if they are not filtered. We are going to remove the features of 'Volume', 'Date', 'Name' from the data.

C Nature of Data

To check the linearity of the data plotting is the optimal way of representing. Hence we check all 5 stocks data nature by plotting them in Figure 6.

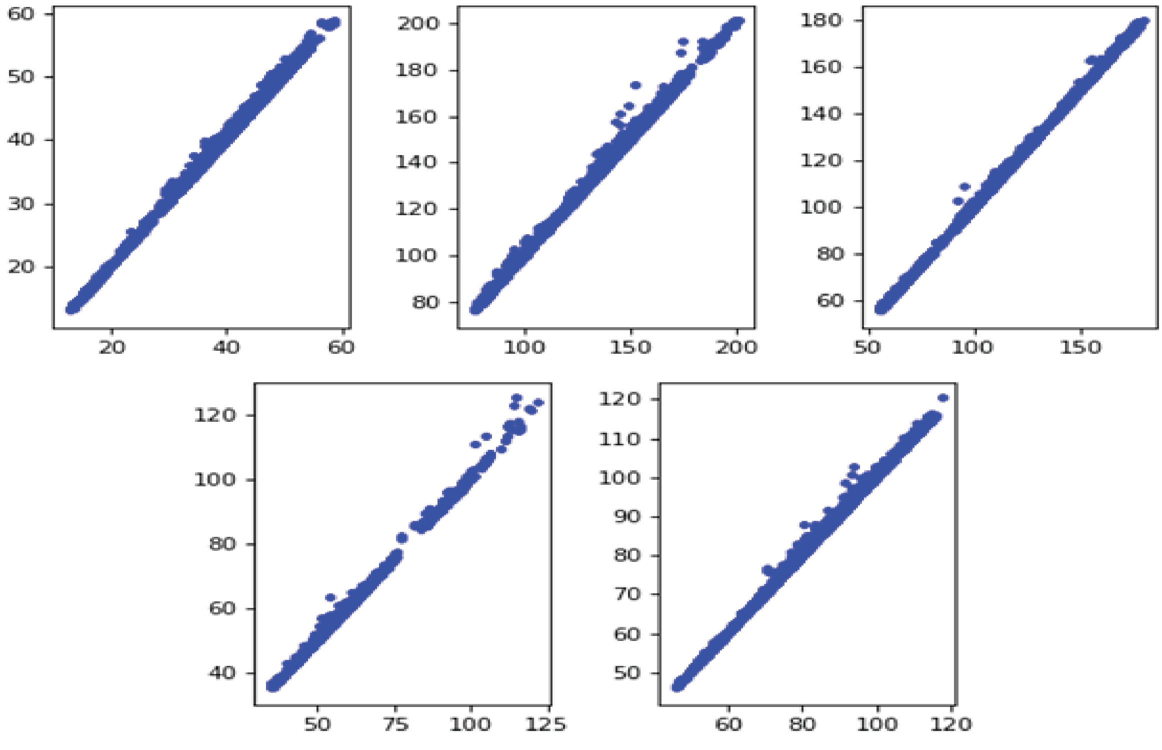


Figure 6: Graphs of 5 different stocks data

The above plots illustrate the linear nature of the data.

4 Classification Algorithm on Basis of Ranges

The classification algorithm is implemented in a supervised learning environment on linear data. Calculate the minimum and the maximum value of each group from any one of the features (as the data is linear therefore it can be calculated from any of the features). These minimum and maximum values will define the range of each group. Now that we have our ranges we can predict the class of an incoming new input/s. We will pick the same feature's value, which we used for calculating our ranges, to check the predicted class. The value lies between one of the ranges and the class of that specific range is assigned as the class of the new input. If the input value exceeds the maximum value of the highest range, then the class of the highest maximum value is assigned to the new input and the maximum value of the highest range is updated with the new input value. If the input value falls short of the lowest minimum value of the smallest range, then the class ofthe lowest minimum value among all the ranges is assigned as the new class and the minimum value of the smallest range is replaced by the new input value.

5 Clustering

The data of all the stocks are unlabeled, that means it is not classified into groups/clusters. There are many clustering algorithms for unsupervised learning. We have used the K-Means algorithm to cluster the data of each stock into 3 groups. After clustering, the plotting of all 5 stocks data is shown in Figure 7.

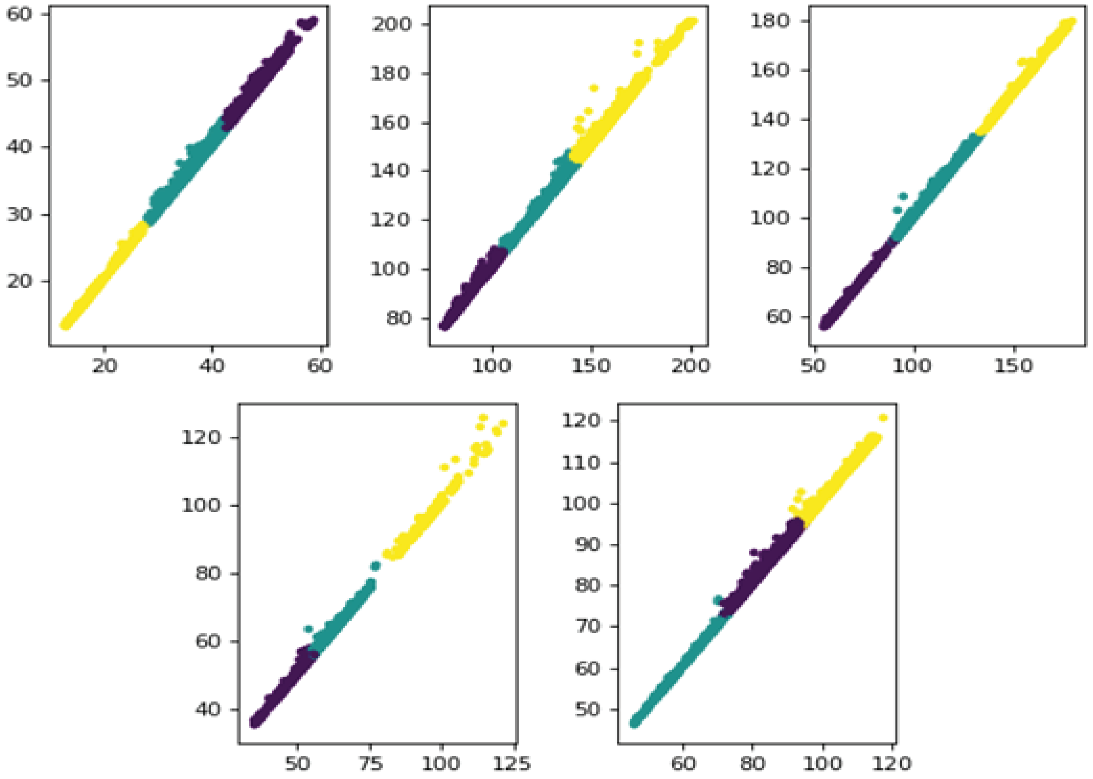


Figure 7: Linear data after K-Means Clustering

Now each stock is classified into 3 groups. K-Means classifies groups into numeric values. For instance, the lower part of the graph shown in each subplot is group '0', the middle part of the graph in each subplot is grouped as '1' and the last part as '2'. So there are 3 groups namely (0,1,2).

6 Experiments and Results

A Nature of Data

After features extraction [7] and classification, we have 4 features and their class, namely (Open, High, Low, Close, Groups). A sample image of the data is shown below in Figure 8 to explain the structure of the features:

Open	High	Low	Close	Group
153.88	154.770	153.31	154.08	0

Figure 8: Stock Features

We have split the data into 2 parts, 90% for training and 10% for the testing. The data is randomly selected for training and testing.

Now we select one feature from where the minimum and maximum values will be calculated. In our case, we selected the "close" feature for the calculation. For each group minimum and maximum values will be calculated from the "close" feature. The Minimum and Maximum value for each group of stock data are shown below in Figure 9:

For group 0 Min Max Values are : (152.05, 220.37)
For group 1 Min Max Values are : (216.05, 266.2)
For group 2 Min Max Values are : (258.07,309.91)

Figure 9: MinMax ranges of each cluster/group

We have our ranges/boundaries which will predict the class of the new inputs. The value of the "close" feature in the new input will be tested, that in which range or boundary it fits and then the class of the specific range is the new class for the input.

B Results

The best accuracy achieved up till now is 99.76%.The accuracy of the test data of the first stock is shown below in Figure 10.

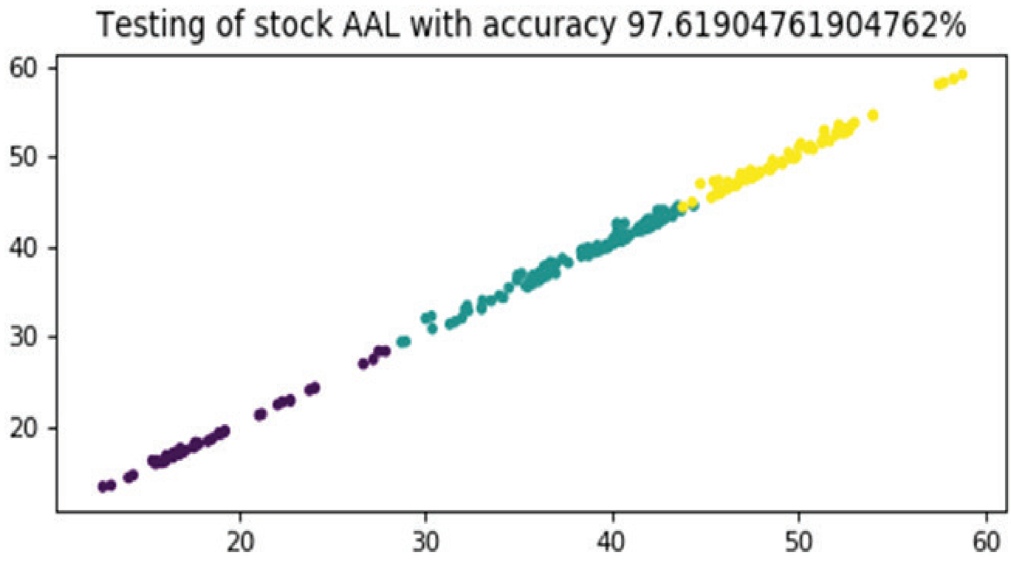


Figure 10: Stock 'AAL' accuracy 97.62%

For the second, third, fourth and fifth datasets of stocks, graphs are listed below:

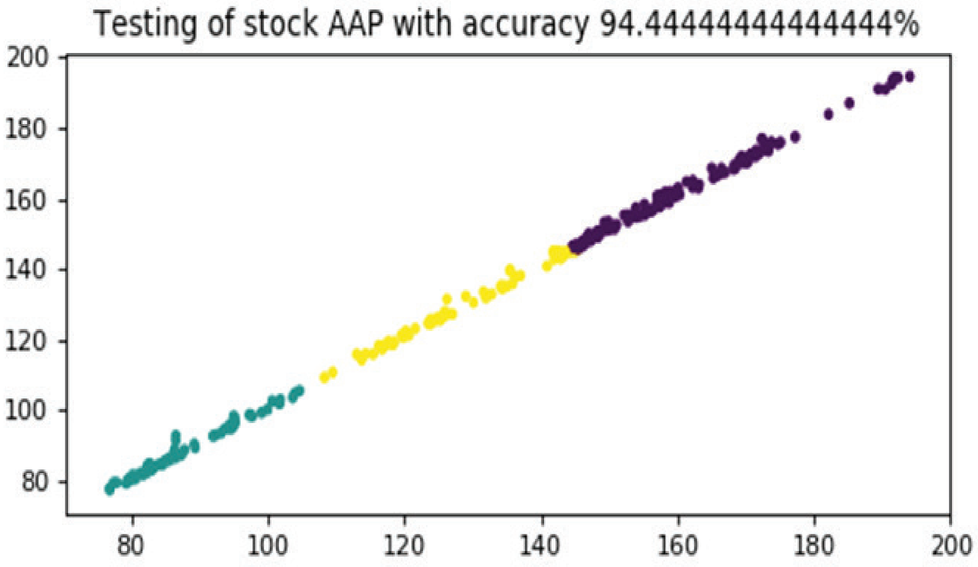


Figure 11: Stock 'AAP' accuracy 94.44%

The images of the above plots show that the accuracy is directly proportional to the linear nature of the data:

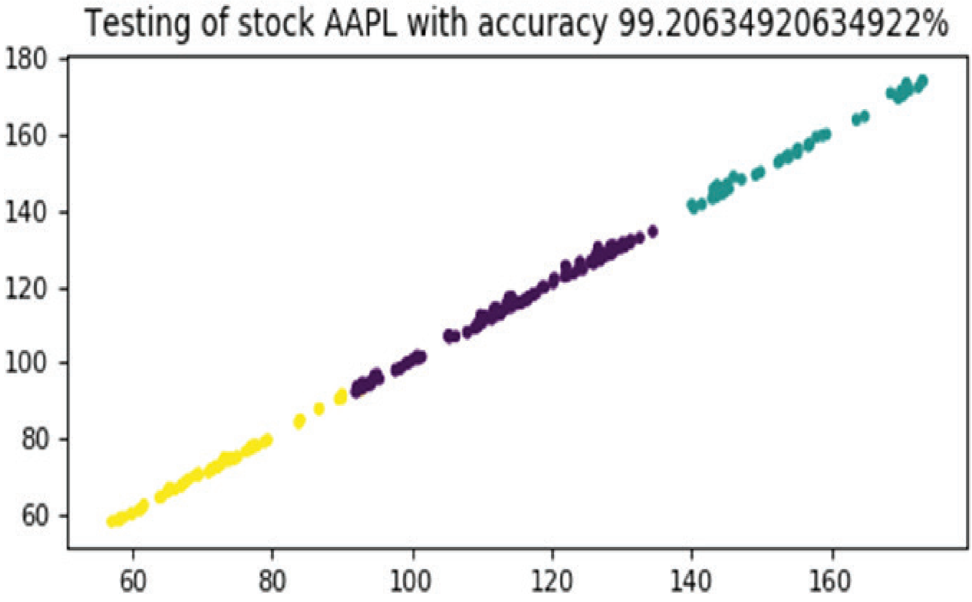


Figure 12: Stock 'AAPL' accuracy 99.21% (Highest Accuracy)

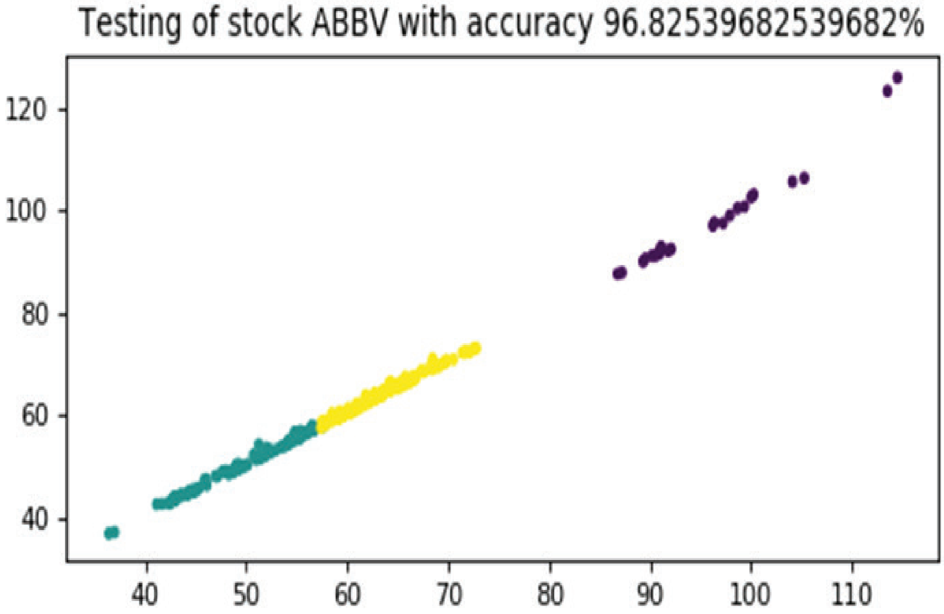


Figure 13: Stock 'ABBV' accuracy 96.82%

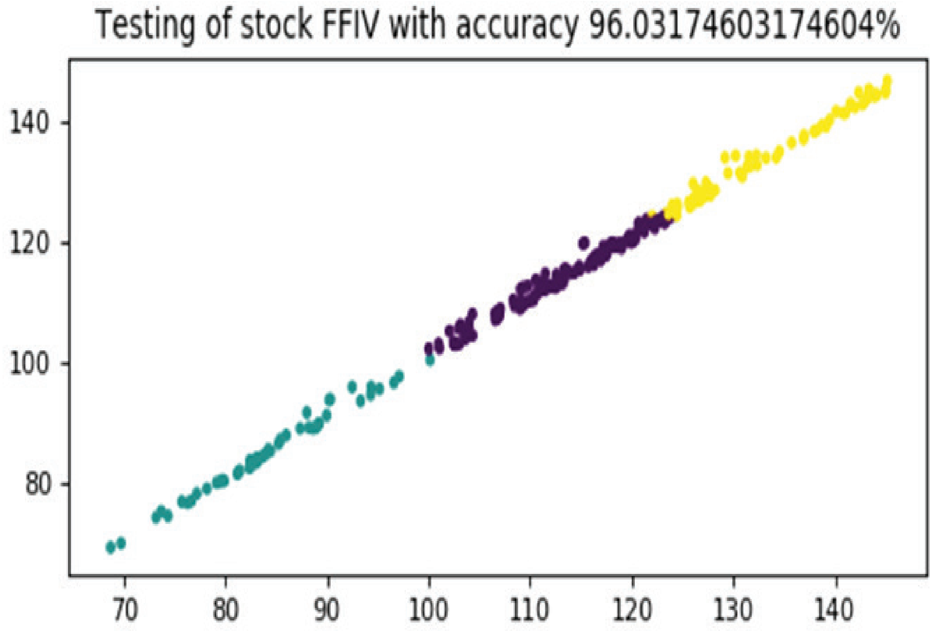


Figure 14: Stock 'FFIV' accuracy 96.03%

The images of the above plots show that the accuracy is directly proportional to the linear nature of the data:

Accuracy Nature of Data

C Experimental Environments

Table 3: PC Combination

CPU	Intel(R) Core(TM) i5-3340M CPU @ 2.70GHz 2.70 GHz
RAM	DDR3 8GB Ram
OS	Windows 7 Ultimate
TOOLS	Pycharm (python, scikit-learn, pandas, matplotlib, scipy)

7 Experiments and Results

This paper introduces a classification algorithm for supervised learning on linear data. On 5 stocks data which consists of Open, High, Low and Close as their features. The data is split into two parts, 90% for the training part and 10% for testing. As for the results, the accuracy percentages are all above 90%, giving a maximum accuracy of 99.20%. The advantage of implementing this model is efficiency, accuracy, and simplicity of the algorithm as other classification algorithms maybe costly on linear data. However, the restriction of the data being linear is a drawback of this model. So, the improvement of this algorithm efficiency on non-linear data is needed soon. For these improvements, we can develop a way of extracting optimal ranges from non-linear classification data as well.

References

[1] Thomas G. Dietterich, “Ensemble Methods in Machine Learning” in 2000. International Workshop on Multiple Classifier Systems, On 2000. LNCS 1857, pp 1–15.

[2] TaiwoOladipupoAyodele (2010).” Types of Machine Learning Algorithms”, New Advances in Machine Learning, Yagang Zhang (Ed.), ISBN: 978-953-307-034-6, InTech

[3] Sahibsingh A. Dudani, “The Distance-Weighted K-Nearest-Neighbor Rule” in1976, IEEE Transactions on Systems, Man, and Cybernetics (Volume: SMC-6, Issue: 4), pp 325-327.

[4] J. A. Hartigan and M. A. Wong, “A K-Means Clustering Algorithm”, in 1979, Journal of the Royal Statistical Society, Series C (Applied Science), Wiley for the Royal Statistical Society, pp 100-108

[5] Gleen W. Milligan and Martha C.Cooper, ”Methodology Review: Clustering Methods ”, in 1987, Ohio State University, Volume: 11 Issue: 4, pp 329-354.

[6] T. SoniMadhulatha, ”An Overview On Clustering Methods”, in 2012, IOSR Journal of Engineering, Vol. 2(4), pp: 719-725.

[7] Shigeo Abe, “Feature Selection and Extraction”, in 2010, Support Vector Machine for Pattern Classification, pp 331-341.

[8] Eugene Charniak, “Introduction to Artificial Intelligence”, in 1985, Brown University.

[9] Makrufa S. Hajirahimova and Aybeniz S. Aliyeva, “Review of Statistical Analysis Methods of Large-Scale Data”, in 2015, 9th IEEE International Conference on Application of Information and Communication (AICT), pp 67-71.

[10] Jin Huang and Charles X. Ling, “Using AUC and Accuracy in Evaluating Learning Algorithms”, in 2005, IEEE Transaction on Knowledge and Data Engineering (Volume: 17, Issue: 3), pp 299-310.

[11] Foram P. Shah and Vibha Patel, ”A Review on Feature Selection and Feature Extraction for Text Classification”, in 2016, IEEE International Conference on Wireless Communications, Signal Processing and Networking(WiSPNET), pp 2264-2268.

[12] C. Barras and J. –L. Gauvain, ”Feature and Score Normalization for Speaker Verification of Cellular Data”, in 2003, IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP ‘03), Volume: 2, pp II-49.

[13] O. Chapelle, B. Scholkopf and A. Zien, Eds, ”Semi-Supervised Learning (Chapelle, O. et al., Eds.;2006)[Book Revies]”, in 2009, IEEE Transactions on Neural Networks (Volume: 20, Issue: 3), pp 542-542.

[14] Leslie Pack Kaelbling, Michael L. Littman and Andrew.W. Moore, ”Reinforcement Learning: A Survey”, in 1996, Journal of Artificial Intelligence Research 4, Volume: 4, pp 237-285.

[15] Jaime G. Carbonell, Ryszard S. Michalski, and Tom M. Mitchell, ”Machine Learning: An Artificial Intelligence Approach”, in 2013.

[16] S. B. Kotsiantis, I. D. Zaharakis, and P. E. Pintelas, ”Machine Learning: A Review of Classification and Combining Techniques”, in 2006, Artificial Intelligence Review, Volume: 26, Issue: 3, pp 159-190.

[17] Giuseppe Bonaccorso, ”Machine Learning Algorithms”, in 2017.