Comparative Analysis of Id3 and Naïve Bayes Algorithm on Stock Market Prediction

I. S. L. Sarwani
Assistant Professor
Dept. of I.T., ANITS
Visakhapatnam

N. Siva Nandhalahari
Dept. of I.T., ANITS
Visakhapatnam

S. Sobha Sri
Dept. of I.T.,ANITS
Visakhapatnam

ABSTRACT
Stock market is a high risk investment influenced by many factors. Stock market prices prediction is not an easy task. With the aid of classification, a data mining technique predicting stock prices considering some factors of influence had been done. This paper put a light on the performance of ID3 and Naïve Bayes algorithms on a given Stock market data. ID3 and Naïve Bayes were classification algorithms which classifies the given data to be classified (test data) basing on the historical data (training data) provided. The historical and test datasets contain attributes which are the factors influencing the stock prices. ID3 algorithm is a Decision Tree technique which constructs a decision tree using the historical data. After decision tree construction, prediction is done for the test dataset values and forecast accuracy is calculated using original value dataset values. Bayesian networks are also used for prediction. The Naïve Bayes algorithm is a Bayesian Network technique used for the Bayesian Network construction using the historical data. The constructed Bayesian Network aids in prediction of the test dataset stock prices and forecast accuracy is calculated using original value dataset values. For computing forecast accuracy root mean square deviation is used. Along with forecast accuracy, under and upper forecasting of the algorithms are also presented. These two algorithms namely ID3 and Naïve Bayes are evaluated on various stock market datasets and the comparison of their performance is provided.

Keywords
Stock Market, Decision Trees, ID3 algorithm, Bayesian networks, Naïve Bayes algorithm, comparative analysis.

1. INTRODUCTION
Data Mining in simple terms is a process of finding information, trends and patterns using large datasets usually collected from various sources which helps in making decisions regarding future activities. The information aimed to be found is usually non-trivial, implicit, unknown and potentially useful. Large amounts of data usually stored data warehouses or other information repositories which were collected from various sources like company's customer feedback, customer purchasing habits etc. are used in the mining activity. As the size of the data collected increases it is more useful for the data mining because knowledge gained from larger amounts of data tends to be more accurate compared with smaller ones. Usually the collected data is known as historical data, learning data or training data. Usage of Data Mining has been increased in recent times as large amount of data useful for the mining activity can be collected in a very short time with the help of internet, email, electronic form. The applications using the techniques of data mining are found in Artificial Intelligence, Machine Learning, Market Analysis, Decision Support and the list continues adding new fields. Though the concept of Data Mining is a part of Knowledge Discovery from Data (KDD) [1] in recent times of the industry, media, and database research the term Data Mining is becoming more popular than the process of KDD.

Data mining consists of six tasks which can be grouped into Supervised and Unsupervised learning. In Supervised learning using the available data a model is constructed. Target attributes are described which are the attributes of interest in terms of remaining attributes. In Unsupervised learning no target attribute is taken. Relationships among the attributes of the available data are established. Classification and prediction [2] comes under the Supervised Learning. In Classification the whole data given is classified into classes so when a new value is given it is fitted into any of the previously classified classes. For the purpose of classification, a classifier is constructed by appropriate techniques. Prediction is used to predict future values based on the given data. This work is a comparative study of very well-known classification algorithms ID3 Decision Tree generation algorithm and Naïve Bayes algorithm.

2. DATASETS USED
The datasets used in the project was provided by Yahoo finance [3] available for all and can be retrieved on daily, monthly, annually intervals. Datasets consists of the Date, Open, High, Low, Close, Volume attributes. Stock price records of APPL company were taken ranging between the dates 12th December 1980 and 12th December 2015 on daily intervals. Open, High, Low, Close attributes were taken into consideration. Below provides the description of the chosen attributes as per the NASDAQ [4] was given below.

- Open: It is used in the context of general equities. It is either the buy or sell interest at the indicated price level and side of a preceding trade.
- High: The highest closing price of a stock over the past 52 weeks, adjusted for any stock splits, or the highest intraday price of a stock in the most recent (or current) trading session.
- Low: In the context of general equities, this is a specific minimum limit required by a seller in execution an order.
- Close: The close is the period at the end of the trading session. Sometimes used to refer to closing price.

There are three types of datasets used in the work. They are named as Training Dataset, Test Dataset and Original Value Dataset. The Training Dataset aids in the construction of Decision trees and Bayesian networks. It was divided into five cases as follows.

Table 1. Tabular representation of date ranges taken in the training dataset

Case number	Date range
1	1980 – 1987
2	1980 – 1994
3	1980 – 2001
4	1980 – 2008
5	1980 – 2015

In the above table,5 cases are in increasing in number of years representing the increase in size of training dataset given. Test Dataset consists of 19 entries. Close attribute value needed to be predicted for each Open, High and Low attribute values of the entry using both the algorithms. Original Dataset consists of 19 entries of Test dataset along with the original Close attribute value in order to calculate the forecast accuracy of the algorithm and for plotting the graph.

3. METHODOLOGIES

3.1 Decision Trees & ID3 Algorithm

3.1.1 Decision Trees

A Decision Tree in simple terms is a classification scheme. It represents a model of different classes, for a given learning dataset. It generates a tree and a set of rules which mainly consists of nodes. The nodes are classified [5] into 3 based on their positions namely root, internal and leaf node. The top most node is called as the root node from where the classification of the dataset starts. The nodes positioned next level are intermediate nodes which denotes a test on attribute. The branches of the nodes represent the outcome of the test. The lowest level from which no branches were present are the leaf nodes which represent class distributions.

3.1.2 Entropy & Information Gain

Entropy H(S)

In Information Theory, the Entropy characterizes the given dataset S. It is a measure of the amount of uncertainty in the dataset. It is also known as Shannon Entropy.

$$H(S) = -\sum_{i=1}^{n} \quad p(xi)log2p(xi)$$

Where,

- S - The current dataset for which entropy is being calculated.
- X - Set of classes in S
- p(xi) - The proportion of the number of elements in class X to the number of elements in set S

When H(S)=0, the given dataset S is perfectly classified in other words all the elements in the dataset are of the same class and no further classification can be done. In ID3, Entropy is calculated for each remaining attribute in the dataset S. The attribute with the smallest entropy is used to split the dataset S on this iteration.

Information Gain

Information gain IG(A) for a given attribute A is the measure of the difference in Entropy from before making decision on the value of an attribute to after the decision is made. In simpler terms, the difference between the Entropy values before and after the split of the dataset S is on an attribute A. It is the measure of decrease of in uncertainty in S due to the split. The attribute with the largest information gain is used to split the set S on this iteration.

$$IG(S,A) = H(S) - \sum_{teT} \quad p(t) * H(t)$$

Where,

- H(S) - Entropy of the dataset S calculated before the split.
- T - The subsets created from splitting dataset S by attribute A.
- p(t) - The ratio of number of elements in subset t to the number of elements in dataset S
- H(t) - Entropy of subset t

3.1.3 ID3 Algorithm

Iterative Dichotomiser3 is widely known as ID3 algorithm. J. Ross Quinlan developed the ID3 algorithm. ID3 is a decision tree generation algorithm. For a given dataset containing attributes and entries ID3 algorithm generates the Decision tree using Shannon Entropy. Following were the steps of the algorithm.

1. Consider a target attribute.
2. Calculate target attribute's Entropy (described in 3.1.3).
3. Calculate the Information Gain (described in 3.1.3) of all remaining attributes in the dataset.
4. Determine the attribute with highest Information Gain as the node.
5. Make the residual dataset by eliminating the selected node.
6. If all the attributes in the dataset are completed go to Step 7, else go to Step 8.
7. Generate the Decision Tree.
8. Continue from Step 2.

3.2 Bayesian Networks & Naïve Bayes Algorithm

3.2.1 Bayesian Networks

A Bayesian network also known as Bayes network, belief network, probabilistic directed acyclic graphical model is a probabilistic graphical model which is a type of statistical model that represents a set of random variables X = {X_1, X_2, X_3, …, X_n} and their conditional dependencies. The conditional dependencies among the set of random variables is represented using a directed acyclic graph (DAG). The DAG consists of nodes and edges. The nodes denote random variables which are either unknown parameters, hypothesis, observable quantities etc. Edges represent the conditional dependencies among the nodes. If any two nodes are disconnected, then the random variables are independent of each other.

3.2.2 Naïve Bayes Algorithm

Bayes Theorem

Bayes theorem describes the probability of occurrence of an event (say A), based on conditions that might be related or

dependent on the occurrence of another event (say B). Bayes theorem is stated mathematically as the following equation:

$$p(A|B) = \frac{p(B|A)\, p(A)}{p(B)}$$

Where,

- P(A) # 0 and P(B) # 0 i.e. the probability of occurrence of two events A and B is non-zero.
- P(A) and P(B) are the probabilities of A and B without dependence on each other.
- P (A | B), a conditional probability, the probability of occurring of event A given that the event B has already occurred.
- P (B | A), also a conditional probability, the probability of occurring event B given that A has already occurred.

Naïve Bayes Algorithm

Naïve Bayes algorithm is a classification technique which generates Bayesian Networks for a given dataset based on Bayes theorem. The classifier based on Naïve Bayes algorithm is the Naïve Bayes classifier [6]. It assumes that the given dataset contains a particular feature in a class which is unrelated to any other feature. For example, an object is considered to be A because of some features. These features presence may depend on each other or on other features but all of the features presence independently contribute to the probability that this object is A and that is the reason it is known as 'Naïve'. Advantages of Naïve Bayes algorithm are it is easy to build and useful for very large datasets and even known to outperform highly sophisticated classification techniques. Following were the important steps to be performed in this algorithm.

1. The given dataset is to be converted into a frequency table.
2. Calculate probabilities of the events and using the probabilities create Likelihood table.
3. Using the Naive Bayesian equation, calculate the posterior probability for all classes.
4. The class with the highest posterior probability is the outcome of prediction.

3.3 Calculating Forecast Accuracy and Model tends

The forecast accuracy in this work is the degree of closeness of the predicted value to that actual value [7]. Actual values which are usually not available at the time of prediction are derived from the original value datasets in this work. Following were the steps for calculating forecast accuracy and determine whether model tends to under-forecast or over-forecast.

1. Calculate error = actual value – forecast (predicted) value.
2. Calculate mean of errors of all cases i.e. the Mean Forecast Error (MFE).
   2.1. If MFE = 0, then it is ideal i.e. actual and predicted values are same.
   2.2. If MFE > 0, model tends to under-forecast.
   2.3. If MFE < 0, model tends to over-forecast.
3. Calculate Root Mean Square deviation
4. Calculate Forecast Error = (|error|/actual value) * 100
5. Calculate Forecast Accuracy = Maximum (0, 100 – Forecast Error)

Forecast accuracy and model tends were to be calculated for every case for both the algorithms.

4. OBSERVATIONS

A total of 5 cases were applied with increasing in learning dataset as represented in Table 1. Both the algorithm implementations are done in Python programming language [8]. Case observations were described below.

Case observations:

Learning Dataset: From the year 1980 – 2015. Total number of entries in the dataset were 8829

Test and Original Value Datasets are described in Datasets Uses (Section 2) section.

The original values from the Original value dataset, ID3 and Naïve Bayes algorithms predicted values were shown in Table 2 and their graphical plot (see Figure 1) was presented.

Table 2. Case Observation

ORIGINAL VALUES	ID3 PREDICTED VALUES	NAÏVE BAYES PREDICTED VALUES
94.09	98.66	96.99
94.09	94.45	94.5
93.42	95.97	95.97
99.99	104.02	100.75
99.44	103.48	100.75
101.42	101.13	100.75
96.3	103.12	94.0
96.79	95.91	96.99
96.66	97.33	94.99
97.13	98.71	95.97
99.52	99.63	94.99
97.39	103.76	100.75
99.96	101.86	100.75
98.53	101.87	94.99
96.96	101.16	94.99
96.45	101.13	94.99
100.7	101.86	102.50
102.71	110.23	102.50
105.35	107.94	102.50

Observations of the case were the forecast accuracies and model tends of both the algorithms.

Naïve Bayes Algorithm

- Forecast Accuracy: 70.328
- Algorithm tends to Under-forecast

ID3 Algorithm

- Forecast Accuracy: 82.739
- Algorithm tends to Under-forecast

Observations were plotted and shown in Figure 1. The 'Open" attribute values are in X-axis original "Close" attribute values from original value dataset and the predicted "Close" values of both algorithms are in Y-axis. Figure 1 is generated using Tkinter package [9] of Python programming language.

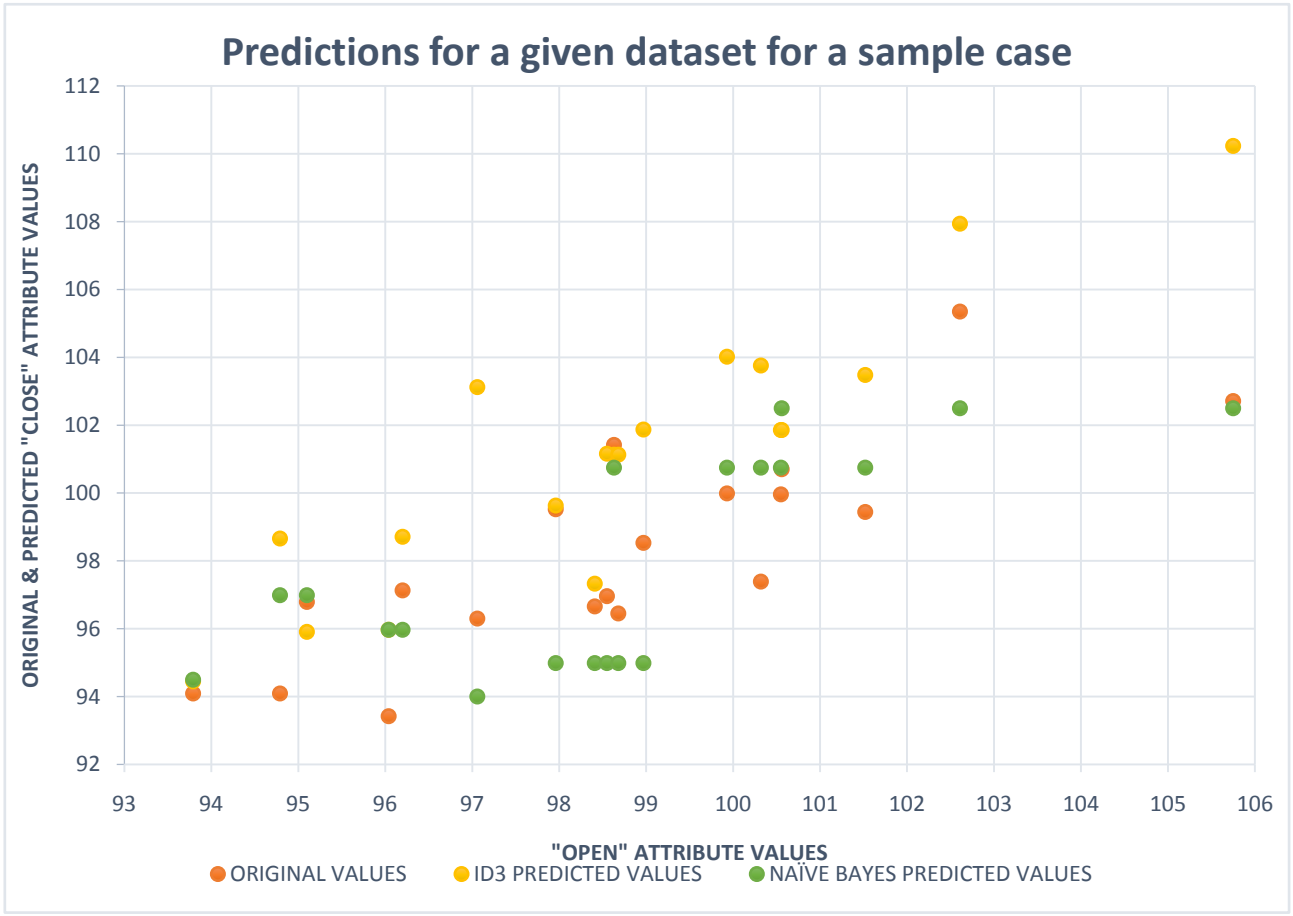

Figure 1: Graphical Representation of sample case observation

ID3 and Naïve Bayes algorithms were applied for all the five cases of Training Datasets (see Table 1) and their prediction accuracies are calculated. Below Table (Table 3) provides the accuracy values of the algorithms case wise and graphical representation was provided in Figure (see Figure 2) below.

Case numbers were taken for X-axis and their corresponding accuracy values of both the algorithms were taken for the Y-axis.

Table 3: Case wise forecast accuracies of algorithms

CASE NUMBER	ID3 ALGORITHM	NAÏVE BAYES ALGORITHM
1	82.73	70.32
2	82.55	72.27
3	94.49	95.44
4	95.27	97.75
5	95.63	98.11

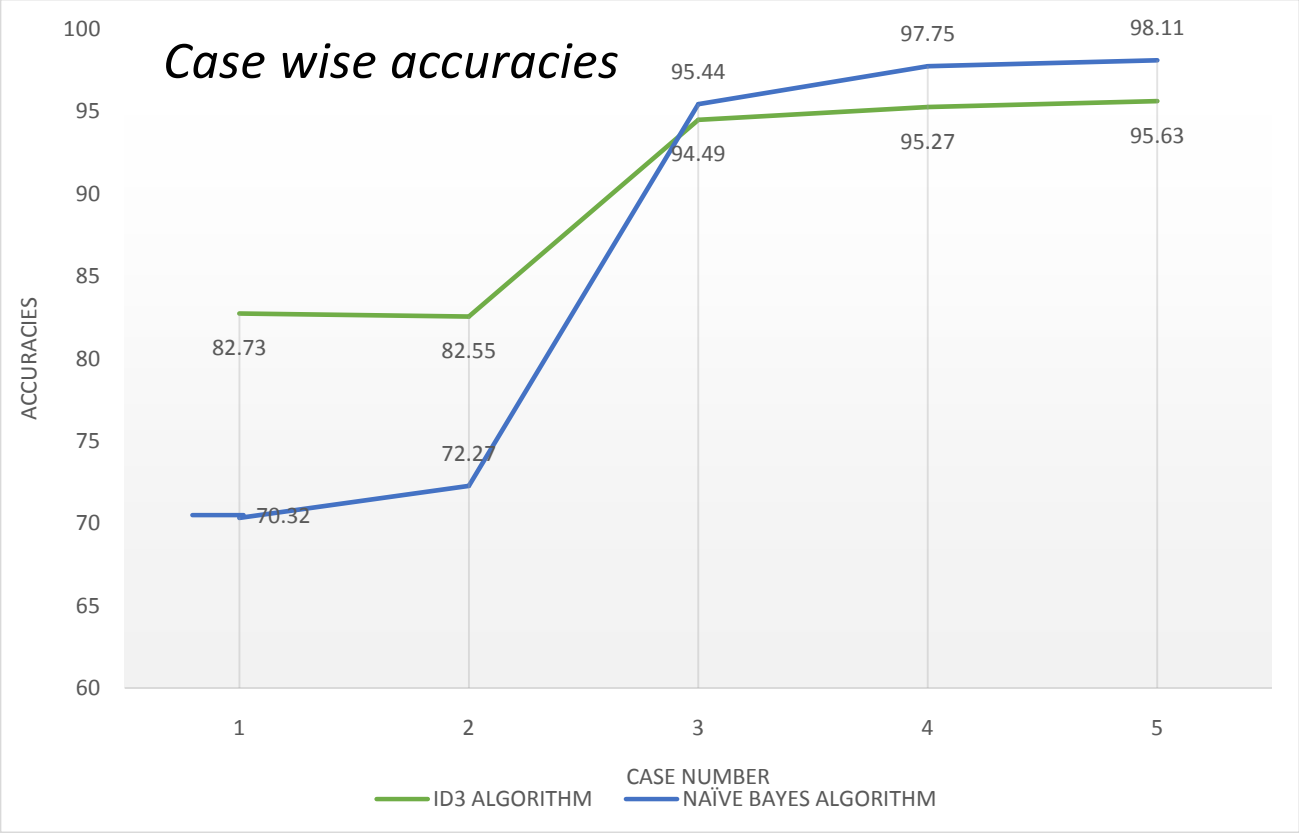

Figure 2: Graphical representation of the case wise accuracies algorithms of the algorithms

5. CONCLUSION AND FUTURE WORK

Data mining consists of many streams like classification, clustering etc. Classification was chosen for the project. After applying different cases of datasets based on the results shown in the graph (see Figure 2) and the Table 3 it being concluded that Naïve Bayes algorithm has better forecast accuracy and tends to learn faster compared to the ID3 algorithm on stock market datasets. Consideration of the missing values in the training datasets which are very much essential for the construction of the classifiers (Decision Trees and Bayesian Networks) could be a serious problem in predicting values of stock market test dataset. Missing values in attributes should be considered so that performance of the two algorithms on missing attributes can be analyzed. Other factors which influence the performance of the prices should be considered as newer attributes in datasets so that ID3 and Naïve Bayes algorithms performance on inclusion of the newer attributes can be analyzed.

The results of this paper would provide the guideline for the research which helps in improving existing algorithms, developing algorithms suitable for stock market prediction.

6. REFERENCES

[1] Data Mining: Concepts and Techniques, Second Edition by Jiawei Han University of Illinois at Urbana-Champaign Micheline Kamber

[2] Data Mining Tasks(http://www.tutorialspoint.com/data_mining/dm_tasks.htm)

[3] Yahoo finance (https://in.finance.yahoo.com/)

[4] NASDAQ Glossary terms(http://www.nasdaq.com/investing/glossary/)

[5] Decision for Buying and Selling Stock with Decision Tree Algorithm by Nur Aini Rakhmawati, Erma Suryani (http://old.its.ac.id/personal/files/pub/4546-erma-is-ICTSbaru.pdf)

[6] Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification by Tina R. Patil, Mrs. S. S. Sherekar (http://www.researchpublications.org/IJCSA/NCAICN-13/189.pdf)

[7] Measuring Forecast Accuracy: Approaches to Forecasting: A Tutorial(https://scm.ncsu.edu/scm-articles/article/measuring-forecast-accuracy-approaches-to-forecasting-a-tutorial)

[8] Mastering Python for Data Science by Samir Madhavan

[9] Tkinter GUI Application Development HOTSHOT by Bhaskar Chaudhary