Stock Market Prediction with Gaussian Naïve Bayes Machine Learning Algorithm

Ernest Kwame Ampomah
School of Information & Software Engineering,
University of Electronic Science and Technology of China, China
E-mail: ampomahke@gmail.com

Gabriel Nyame
Department of Information Technology Education
Akenten Appiah-Menka University of Skills Training and Entrepreneurial Development, Kumasi-Ghana
E-mail: kwakuasane1972@gmail.com

Zhiguang Qin
School of Information & Software Engineering,
University of Electronic Science and Technology of China, China
E-mail: qinzg@uestc.edu.cn

Prince Clement Addo
School of Management and Economics, University of Electronic Science and Technology of China, China
E-mail: prince@std.uestc.edu.cn

Enoch Opanin Gyamfi
School of Information & Software Engineering,
University of Electronic Science and Technology of China, China
E-mail: enochopaningyamfi@outlook.com

Michael Gyan
Department of Physics Education, University of Education, Winneba-Ghana
E-mail: mgyan173@gmail.com

Keywords: machine learning, gaussian naïve Bayes, stock price, feature extraction, scaling

Received: January 8, 2021

The stock market is one of the key sectors of a country's economy. It provides investors with an opportunity to invest and gain returns on their investment. Predicting the stock market is a very challenging task and has attracted serious interest from researchers from many fields such as statistics, artificial intelligence, economics, and finance. An accurate prediction of the stock market reduces investment risk in the market. Different approaches have been used to predict the stock market. The performances of Machine learning (ML) models are typically superior to those of statistical and econometric models. The ability of Gaussian Naïve Bayes ML algorithm to predict stock price movement has not been addressed properly in the existing literature, hence this attempt to fill that gap in the literature by evaluating the performance of GNB algorithm when combined with different feature scaling and feature extraction techniques in stock price movement prediction. The performance of the GNB models set up were ranked using the Kendall's test of concordance for the various evaluation metrics used. The results indicated that, the predictive model based on integration of GNB algorithm and Linear Discriminant Analysis (GNB_LDA) outperformed all the other models of GNB considered in three of the four evaluation metrics (i.e., accuracy, F1-score, and AUC). Similarly, the predictive model based on GNB algorithm, Min-Max scaling, and PCA produced the best rank using the specificity results. In addition, GNB produced better performance with Min-Max scaling technique than it does with standardization scaling techniques

Povzetek: Predstavljena je metoda Gausovega naivnega Bayesa za borzne napovedi.

1 Introduction

The stock market is one of the key sectors of a country's economy. It provides investors with an opportunity to invest and gain returns on their investment. Predicting the stock market has attracted serious interest from researchers from many fields such as statistics, artificial intelligence, economics, and finance. An accurate

prediction of the stock market reduces investment risk in the market. Different opinions exist as regards to the predictability of the stock market. The efficient market hypothesis (EMH) states that all available information is fully incorporated by current market price immediately, therefore, changes in price of the stocks are as a result of new information [1]. The EMH implies that stock prices would trail a random walk pattern, hence, the stock market cannot be forecasted from past data to make any meaningful returns [2]. However, numerous researches have been conducted since the beginning of the 21st century which contradicts the EMH and show that the stock market can be predicted to some extent [3-5]. Exploration of many prediction algorithms in stock market forecasting has taken place and showed that the behavior of stock prices can be forecast [6]. The prediction of the stock market behavior is a very difficult task since this market is very complex, non-linear, and evolutionary. The market is influenced by situations such as investors' sentiments, political events, and overall economic conditions [7]. Three main approaches: fundamental analysis, technical indicators, and machine learning (ML) are used to forecast the stock market. In fundamental analysis, the value of a stock is derived from the general economic and financial factors such inflation, return on equity (ROE), price to earnings (PE) ratios, and debt levels. In technical analysis approach, technicians use charts and market statistics from historical price data to identify market trends and patterns so that they can make fairly accurate forecast of the trajectories of the stock market behavior [8]. The machine learning approach offers system the ability to learn and improve automatically from massive amount of historical data without them being explicitly programmed. Machine learning models have been shown to perform better than both fundamental, and technical analyses in the literature [9-11]. Distributional assumptions are not required by ML models. Also, ML models are able find hidden patterns in time series data [12-13]. Several machine learning algorithms exist, but the focus of this study is on Gaussian Naïve Bayes (GNB) algorithm. GNB is a probabilistic classifier based on Bayes' theorem with assumption of strong (naïve) independence between the features [14]. GNB algorithm is very simple and easy to implement and does not require too many training data. It is highly scalable (it scales linearly with the number of features and data points), not sensitive to irrelevant features and able to deal with missing data very effectively. A major weakness with GNB algorithm is the assumption of independence between predictors. GNB assumes that all the predictors are mutually independent. This assumption is hardly true in real life especially with financial data. However, this assumption can be met by applying feature extraction techniques to extract independent predictors from the given data. Many feature extraction techniques are available in the literature which can be used to achieve this goal. Hence, this work assesses the performance of GNB with different feature scaling and feature extraction techniques in predicting the direction of movement of stock prices.

2 Related studies

Many ML algorithms have been used in the literature of forecasting the direction of stock price. A review of some of those works is provided. Ampomah et al, (2020) [14] studied the effectiveness of tree-based AdaBoost ensemble ML models (namely, AdaBoost-DecisionTree (Ada-DT), AdaBoost-RandomForest (Ada-RF), AdaBoost-Bagging (Ada-BAG), and Bagging-ExtraTrees (Bag-ET)) in predicting stock prices. The experimental results showed that AdaBoost- ExtraTree (Ada-ET) model generated the highest performance among the tree-based AdaBoost ensemble models studied. Kumar and Thenmozhi (2006) [15] carried out a study to forecast the direction of S&P CNX NIFTY Market Index of the National Stock Exchange (NSE). Random forest, linear discriminant analysis, artificial neural network, logit, and SVM machine learning algorithms were used by the researchers. The experimental results indicated that SVM is the best performer among the classification algorithms used. Ou and Wang (2009) [16], studied and applied ten different data mining techniques to forecast stock price movement of Hang Seng index of Hong Kong stock market. The techniques included neural network, Linear discriminant analysis (LDA), Logit model, Quadratic discriminant analysis (QDA), K-nearest neighbor classification, Naïve Bayes based on kernel estimation, Bayesian classification with Gaussian process, Tree based classification, SVM and Least squares support vector machine (LS-SVM). The empirical results presented indicate that the performance of SVM and LS-SVM models are superior to those of the other models. Subha and Nambi (2012) [17] examined the predictability of the movement of BSE-SENSEX and NSE-NIFTY stock indices of the Indian Stock Market by using k-Nearest Neighbours algorithm (k-NN) and Logistic Regression model to predict the daily movement of the indices. Data for the period between January 2006 to May 2011were used. The research outcome shows that the k-NN classifier performed better than the logistic regression model in all the model evaluation metrics used. Saifan et al, (2020) [18] applied the Quantopian algorithmic stock market trading simulator to evaluate ensemble models performance in daily prediction and trading. The ensemble models used are Extremely Randomized Trees, Random Forest, and Gradient Boosting. The models were trained using multiple technical indicators and automatic stock selection. The results showed a significant returns relative to the benchmark and large values of alpha were generated from all models. A study to verify whether modified SVM classifier can be applied successfully in prediction of short-term trends in the stock market was undertaken by Zikowski, (2015) [19]. The author computed and used several technical indicators and statistical measures as input features. Fisher's method was applied to perform feature selection. The study outcome shows that using the modified SVM in conjunction with feature selection enhance significantly the trading strategy results in terms of the total rate of return, as well as the maximum drawdown during a trading period. Patel, et al (2015) [20], compared the performance of Artificial Neural Network

(ANN), support vector machine (SVM), random forest and Naive-Bayes with two different approaches for input data to the models in forecasting the direction of movement of stock and stock price index. The first approach to input data computed ten technical indicators from the stock trading data (open, high, low & close prices) and the second approach represent the technical indicators as trend deterministic data. They evaluated the models with 10 years of historical stock data from 2003 to 2012 of Reliance Industries, Infosys Ltd, CNX Nifty and S&P Bombay Stock Exchange (BSE) Sensex. The outcome of the study shows that for the first approach random forest outperforms other three prediction models on overall performance. Also, that the performance of all the prediction models improved when these technical indicators are represented as trend deterministic data. Sun et al, (2018) [21], proposed a hybrid ensemble learning model combining AdaBoost and LSTM network to predict financial time series. Daily datasets of two major exchange rate and two stock market indices were used for evaluating the model. The experimental outcome shows that the AdaBoost-LSTM ensemble model outperformed the other single forecasting models and ensemble models that were compared with it. Khan et al, (2020) [22] assessed the impact of social media and financial news data on stock market prediction accuracy. The authors performed feature selection and spam tweets reduction on the data sets. Experiments were performed to find stock markets that are difficult to predict and those that were more influenced by social media and financial news. They compared the results of different algorithms to find a consistent classifier. Deep learning and some ensemble classifiers were used. The experimental results indicated that highest prediction accuracies of 80.53% and 75.16% were achieved using social media and financial news, respectively. Also, New York and Red Hat stock markets were difficult to predict, New York and IBM stocks are more influenced by social media, while London and Microsoft stocks by financial news. Random forest classifier was found to be consistent and highest accuracy of 83.22% was achieved by its ensemble. Bhandare et al, (2020) [23] used the Naive Bayes classifier to provide analyse and quantify the performance of stock market analysts by providing ratings. The recommendations given by the analysts was analysed and factors relevant to the success or failure of the recommendation extracted. The Naive Bayes classifier was used provide a rating on the factors thus extracted. The results indicated that the system efficiently analyse the performance of an analyst given their passed records by matching it with the actual stock prices and provide a rating for the analyst using the Naive Bayes classifier. The performance of the system is optimal when Gaussian Naive Bayes Classifier was used.
From the above discussion, and to the best of our knowledge, the ability of Gaussian Naïve Bayes to predict stock price movement has not been addressed properly in the existing literature. Hence, a gap study aims to fill in that gap by evaluating the impact of feature scaling and feature extraction techniques on GNB algorithm in prediction of stock price movement.

3 Method

3.1 Experimental design

Stock Data set used for the study were gathered randomly from three different stock market (NYSE, NASDAQ and NSE) through yahoo financial application programming interface (API). Daily data of seven stocks were gathered. Details of the stock data used are given in Table A1 in the Appendix. Forty (40) technical indicators were computed from the raw stock data which comprise of open price, low price, high price, close price and volume. The computed technical indicators were used as input features for the GNB models. Details of these technical indicators are presented in Table A2-A4 in the Appendix. Each data set was split into training and test set. Initial seventy percent (70%) of the data was used as the training set, and the final thirty percent (30%) of the data was used as the test set. In this work, the ability of GNB algorithm in combination with different feature scaling techniques (i.e., Standardization scale and Min-Max scale) and different feature extraction techniques (i.e., PCA, LDA, and FA) to forecast stock price movement were evaluated.

The following GNB models were evaluated and compared: (i) GNB model, (ii) Integrated model based on GNB algorithm and standardization scaling (GNB_Z-Score) (iii) Integrated model based on GNB algorithm and Min-Max normalization (GNB_Min-Max) (iv) Integrated model based on GNB algorithm and principal component analysis (GNB_PCA) (v) Integrated model based on GNB algorithm and factor analysis (GNB_FA) (vi) Integrated model based on GNB algorithm and linear discriminant analysis (GNB_LDA) (vii) Integrated model based on GNB algorithm, standardization scaling, and principal component analysis (GNB_Z-Score_PCA) (viii) Integrated model based on GNB algorithm, standardization scaling, and factor analysis (GNB_Z-Score_FA) (ix) Integrated model based on GNB algorithm, Min-Max normalization, and principal component analysis (GNB_Min-Max_PCA) (x) Integrated model based on GNB algorithm, Min-Max normalization, and factor analysis (GNB_Min-Max_FA).
GNB model applies the GNB algorithm to the raw stock data without any feature scaling or feature extraction to make prediction. GNB_Z-Score model first used the standardization scaling technique to scale the data, and then the GNB algorithm was applied to forecast the movement of stock price. GNB_Min-Max model applied Min-Max scaling technique to scale the data before the GNB algorithm was applied to the scaled data to make predictions. With GNB_PCA model, the PCA was first applied to the unscaled stock data to extract important features from the data, and then the GNB algorithm was applied to the extracted data to make prediction. GNB_FA model initially applied FA technique to the unscaled stock data to extract relevant features from the original data, and then applied the GNB algorithm to the extracted data to make predictions. GNB_LDA model first used LDA technique to extract relevant features from the initial input data, and then applied the GNB algorithm to the extracted data. GNB_Z-Score_PCA model first applied

standardization scaling technique to the initial data to scale it, after that it then applied PCA to extract important features from the scaled data and finally applied the GNB to the extracted scaled stock data to make predictions. GNB_Z-Score_FA model initially used standardization scaling technique to scale the data, then applies the FA technique to extract relevant features from the scaled data and the GNB algorithm was applied to the extracted scaled data to make predictions. GNB_Min-Max_PCA model initially scaled the input data with Min-Max scaling technique, then applied PCA to extract important feature from the original stock data, and then applied the GNB algorithm to the extracted scaled data to make predictions. GNB_Min-Max_FA uses Min-Max to first scaled the data, then applied the FA technique to extract relevant features from the scaled data, and finally applied the GNB algorithm to make predictions.

3.2 Feature scaling techniques

Feature Scaling is a way of standardizing the independent features that are present in the data within a fixed range. The two most widely used feature scaling techniques are standardization scaling and Min-Max Normalization.

3.2.1 Standardization scaling

Standardization scaling (Z-score) is a scaling method that centers the values around the mean with a unit standard deviation. The data is scaled to a specific area to enable a thorough analysis. The variables are rescaled to have a mean of zero and the resulting distributions have a unit standard deviation. The standardized scaling is expressed by the formula below.

$$X' = \frac{X-\mu}{\sigma} \quad (1)$$

μ = mean of the feature values

σ = standard deviation of the feature values

3.2.2 Min-max normalization

Min-Max normalization (Min-Max) is a scaling approach in which features are re-scaled so that the data will fall in the range of zero and one. It undertakes a linear alteration on the initial data [24]. In the Min-Max scaling, the minimum value of every feature is converted to zero, and the maximum value of each feature is converted to one. The formula below expresses how the normalized form of each feature is computed.

$$X' = \frac{X-X_{min}}{Xmin_{max}} \quad (2)$$

X_{min}= minimum value of the feature, $X_{\max}$ = maximum value of the feature

3.3 Feature extraction techniques

Feature extraction is a dimensionality reduction process that extracts important features or attributes of the data in order to reduce the initial set of data to generate a more concise description of the data for processing. There are many feature extraction techniques in existing literature, however, in this study the principal component analysis (PCA), Linear discriminant analysis (LDA), and factor analysis (FA) were applied.

3.3.1 Principal component analysis

Principal Component Analysis (PCA) is a dimensionality-reduction technique that transform higher data sets to a lower dimensional set. It transforms a data set of interrelated features, into a new set of uncorrelated features called principal components (PCs) and the initial few of these PCs hold most of the variation present in the entire data set [25]. The PCs are linear combination of the actual features in such a way that the first PC has the largest amount of variation and the second PC is orthogonal to the first PC and has the most variance among the remaining PCs. The subsequent PCs follow in that order. The underlying assumption in PCA is that the coordinates with the large variants demonstrate the divergence between sample points, while the coordinates with lesser variants may be a source of noise, which must be ignored or suppressed. The correlation between two dimensions denotes irrelevant information, which will not be presented. This is why PCA requires the subsequent coordinates to be orthogonal to previous coordinates [26]. PCA is sensitive to scaling.

3.3.2 Linear discriminant analysis

Linear Discriminant Analysis (LDA) is a supervised linear transformation technique that computes the linear discriminants (directions) that will represent the axes which maximize the differences between multiple classes. The objective of the technique is to maximize the ratio of the between-group variance and the within-group variance. When the ratio is maximum, then the instances within each group have the least possible scatter and the groups are separated from each other the most. LDA is used to map features in higher dimension space into a lower dimension space while keeping the class-discriminatory information [27]. LDA is not sensitive to scaling, hence, the performance of LDA remains the same with or without scaling. The LDA uses two criteria to generate a new axis: (i) maximize the distance between means of the two classes, (ii) minimize the variation within each class.

3.3.3 Factor analysis

Factor analysis (FA) is a data reduction technique that describes variability among observed, correlated features in terms of a potentially smaller number of unobserved (latent) features called factors. The observed features are modeled as linear combinations of the factors plus error. FA extracts maximum common variance from all features and place them under a common score. This score as an index of all features can be used to do further analysis. FA evaluates how much of the variability in the data is as a result of common factors. The main goals of FA are to display multidimensional data in a lower dimensional space with minimum loss of information and to extract the independent latent of the data [28]. The FA technique makes the following assumptions: linear relationship

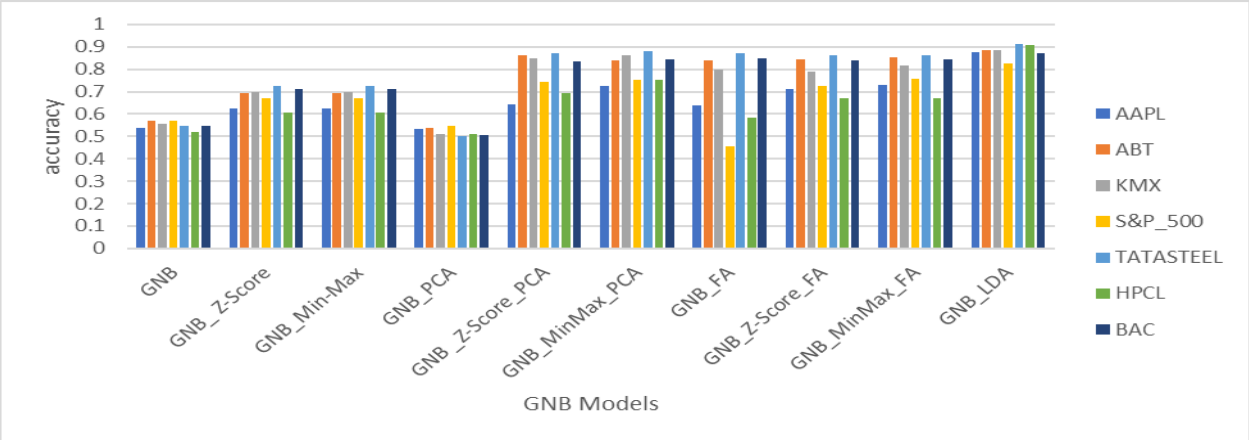

Figure 1: Accuracy results of the GNB models on the different stock data sets.

DataSets	GNB	GNB_ Z-Score	GNB_ MinMax	GNB_ PCA	GNB_ Z-Score_PCA
AAPL	0.5361	0.6241	0.6241	0.5342	0.6444
ABT	0.5713	0.6954	0.6954	0.5361	0.8639
KMX	0.5583	0.6982	0.6982	0.5111	0.8509
S&P_500	0.5722	0.6704	0.6704	0.5472	0.7444
TATASTEEL	0.5461	0.7232	0.7232	0.5012	0.8713
HPCL	0.5197	0.6085	0.6085	0.5126	0.6953
BAC	0.5472	0.7111	0.7111	0.5056	0.8333
Mean	**0.5501**	**0.6758**	**0.6758**	**0.5211**	**0.7862**

DataSets	GNB_ MinMax_ PCA	GNB_ FA	GNB_ Z-Score_FA	GNB_ MinMax_FA	GNB_ LDA
AAPL	0.7241	0.6370	0.7111	0.7314	0.8769
ABT	0.8398	0.8407	0.8462	0.8528	0.8861
KMX	0.8648	0.7963	0.7907	0.8176	0.8870
S&P_500	0.7546	0.4537	0.7269	0.7583	0.8259
TATASTEEL	0.8809	0.8701	0.8616	0.8637	0.9142
HPCL	0.7548	0.5832	0.6700	0.6700	0.9092
BAC	0.8435	0.8509	0.8407	0.8454	0.8713
Mean	**0.8089**	**0.7188**	**0.7782**	**0.7913**	**0.8815**

Table 1: Accuracy results recorded by the GNB models.

exists between the observed features and the common factors, no multi-collinearity is present, it includes relevant features into analysis, and there is true correlation between features and factors.

3.4 Evaluation metrics

The performances of the models were evaluated using the following evaluation metrics:

Accuracy: The percentage of entire instances rightly predicted by the model.

$$accuracy = \frac{tp+tn}{tp+tn+fp+fn} \quad (3)$$

F1-score: This is a harmonic mean of precision and recall

$$F1-score = \frac{2\times precision\times recall}{precision+recall} \quad (4)$$

Specificity: The proportion of negative instances rightly predicted by the classifier out of the total instances that are actually negative. This shows a model's ability to classify true negative instances as negative.

$$specificity = \frac{tn}{tn+fn} \quad (5)$$

Area Under Receiver Operating Characteristics Curve (AUC): Measures the ability of the classifier to distinguish between the positive and negative classes. A perfect classifier will have AUC of one. AUC measures tradeoff between specificity and recall.

Kendall's coefficient of concordance (W): is a metric that uses ranks to establish an agreement among raters. It measures the agreement among different raters who are evaluating a given set of objects [29]. Depending on the area where it is being used, the raters can be variables, characters, and so on. The raters are the different data sets in this work.

4 Experimental results

Table 1 provides the accuracy results generated by the various GNB models on the different stock data sets used. GNB_LDA model produced accuracy results which were better than all the other GNB models on each of the stock data used. The highest accuracy value recorded by any of

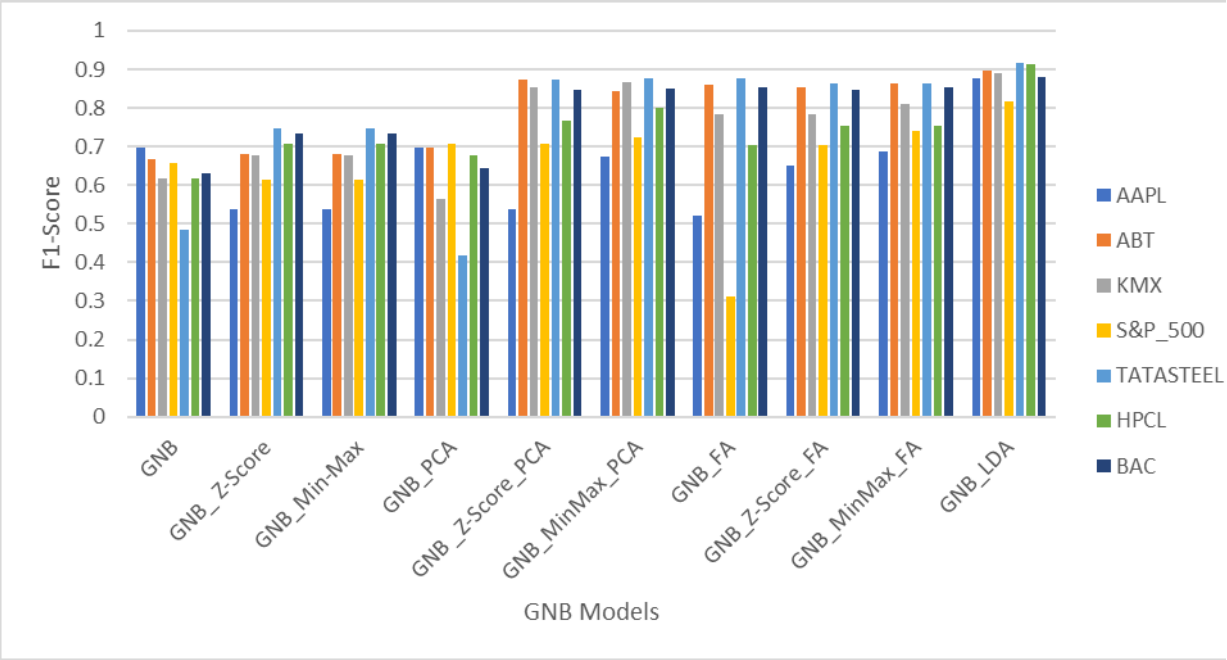

Figure 2: F1-score of the GNB models on the different stock data sets.

DataSets	GNB	GNB _ Z-Score	GNB _ MinMax	GNB _ PCA	GNB_ Z-Score_PCA
AAPL	0.6962	0.5365	0.5365	0.6964	0.5362
ABT	0.6662	0.6803	0.6803	0.6980	0.8740
KMX	0.6175	0.6766	0.6766	0.5629	0.8540
S&P_500	0.6583	0.6139	0.6139	0.7074	0.7058
TATASTEEL	0.4860	0.7461	0.7461	0.4166	0.8747
HPCL	0.6161	0.7074	0.7074	0.6777	0.7659
BAC	0.6304	0.7342	0.7342	0.6454	0.8454
Mean	**0.6244**	**0.6707**	**0.6707**	**0.6292**	**0.7794**

DataSets	GNB_ MinMax_PCA	GNB_ FA	GNB_ Z-Score_FA	GNB_ MinMax_FA	GNB_ LDA
AAPL	0.6740	0.5220	0.6494	0.6875	0.8783
ABT	0.8443	0.8590	0.8534	0.8635	0.8976
KMX	0.8653	0.7835	0.7839	0.8108	0.8891
S&P_500	0.7237	0.3114	0.7053	0.7398	0.8175
TATASTEEL	0.8760	0.8762	0.8626	0.8648	0.9167
HPCL	0.7990	0.7031	0.7532	0.7532	0.9119
BAC	0.8516	0.8546	0.8473	0.8542	0.8790
Mean	**0.8048**	**0.7014**	**0.7793**	**0.7963**	**0.8843**

Table 2: F1-scores recorded by the GNB model.

the models is 0.9142 generated by the GNB_LDA model on the TATASTEEL data. The least accuracy value recorded by any of the models is 0.5012 by GNB_ PCA model on the TATASTEEL stock data. The mean accuracy value of GNB_LDA model (0.8815) was the highest mean accuracy value, and GNB_ PCA produced the least mean accuracy value (0.5211). Figure 1 provides the column chart of the accuracy values produced by the GNB models on the different stock data.

Table 2 shows the outcome of F1-scores evaluation metric of the GNB models on the different stock data sets used. The F1-score of GNB_LDA model was better than all the other GNB models on each of the stock data. The highest F1-score recorded by any of the models was 0.9167 generated by the GNB_LDA model on the TATASTEEL data. The least F1-score value recorded by any of the models was 0.4166 produced by GNB_ PCA on the TATASTEEL stock data. The mean F1-score of GNB_LDA model (0.8815) was the highest mean F1-score among the GNB models, and the mean F1-score of the GNB model (0.6244) was the least mean F1-score among the various models. Figure 2 represents the column chart of the F1-score evaluation metric outcome for the GNB models on the different stock data.

Table 3 presents the specificity results of the models on the different stock data sets used. GNB_Z-Score_FA

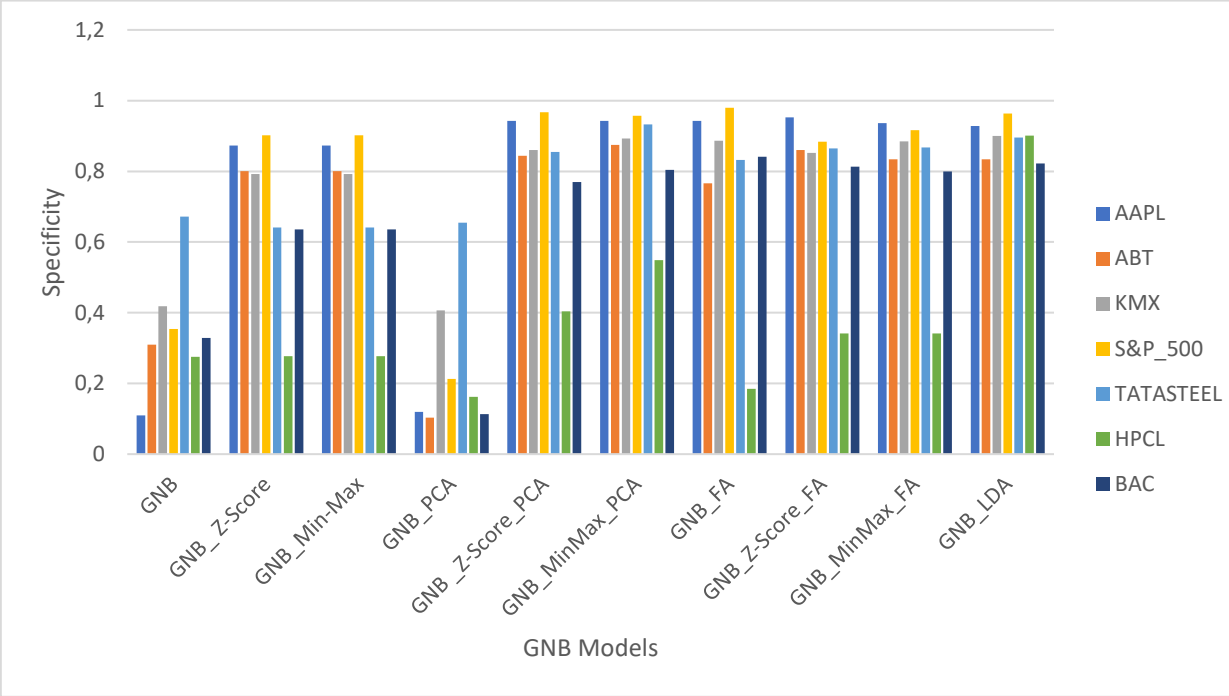

Figure 3: Specificity of the GNB models on the different stock data sets.

DataSets	GNB	GNB_ Z-Score	GNB_ Min-Max	GNB_ PCA	GNB_ Z-Score_PCA
AAPL	0.1099	0.8728	0.8728	0.1200	0.9423
ABT	0.3094	0.8004	0.8004	0.1030	0.8443
KMX	0.4184	0.7927	0.7927	0.4069	0.8599
S&P_500	0.3538	0.9018	0.9018	0.2131	0.9672
TATASTEEL	0.6717	0.6413	0.6413	0.6544	0.8544
HPCL	0.2754	0.2774	0.2774	0.1621	0.4037
BAC	0.3283	0.6359	0.6359	0.1132	0.7698
Mean	**0.3524**	**0.7032**	**0.7032**	**0.2532**	**0.8059**

DataSets	GNB_ Min-Max_PCA	GNB_ FA	GNB_ Z-Score_FA	GNB_ Min-Max_FA	GNB_ LDA
AAPL	0.9423	0.9424	0.9523	0.9363	0.9284
ABT	0.8743	0.7665	0.8603	0.8343	0.8343
KMX	0.8925	0.8868	0.8522	0.8848	0.9002
S&P_500	0.9571	0.9800	0.8834	0.9161	0.9632
TATASTEEL	0.9326	0.8326	0.8652	0.8674	0.8957
HPCL	0.5487	0.1843	0.3416	0.3416	0.9006
BAC	0.8038	0.8415	0.8132	0.8000	0.8226
Mean	**0.8502**	**0.7763**	**0.7955**	**0.7972**	**0.8921**

Table 3: Specificity values recorded by the GNB models.

model outperformed the other models on AAPL. GNB_Min-Max_ PCA model performed better than the rest of the models on ABT, and TATASTEEL stock data. GNB_LDA model recorded better specificity results than the rest of the models on KMX, and HPCL stock data. GNB_FA model produced better specificity results than the other models on S&P_500 and BAC stock data. The highest specificity value recorded by any of the GNB models was 0.9800 generated by the GNB_FA model on the S&P_500 data. The least specificity value recorded by any of the models was 0.1030 produced by GNB_ PCA on the ABT stock data. The mean specificity of GNB_LDA model (0.8921) was the highest mean specificity among the GNB models, and the mean specificity of GNB_PCA model (0.2532) was the least mean specificity among the GNB models. Figure 3 presents the column chart of the specificity results of the GNB models on the different stock data.

Table 4 provides the AUC results of the GNB models on the different stock data sets used. The performance of GNB_LDA model on each of the stock data was better than the rest of the GNB models. In general, the highest AUC value recorded by any of the models was 0.9743 generated by the GNB_LDA model on the TATASTEEL

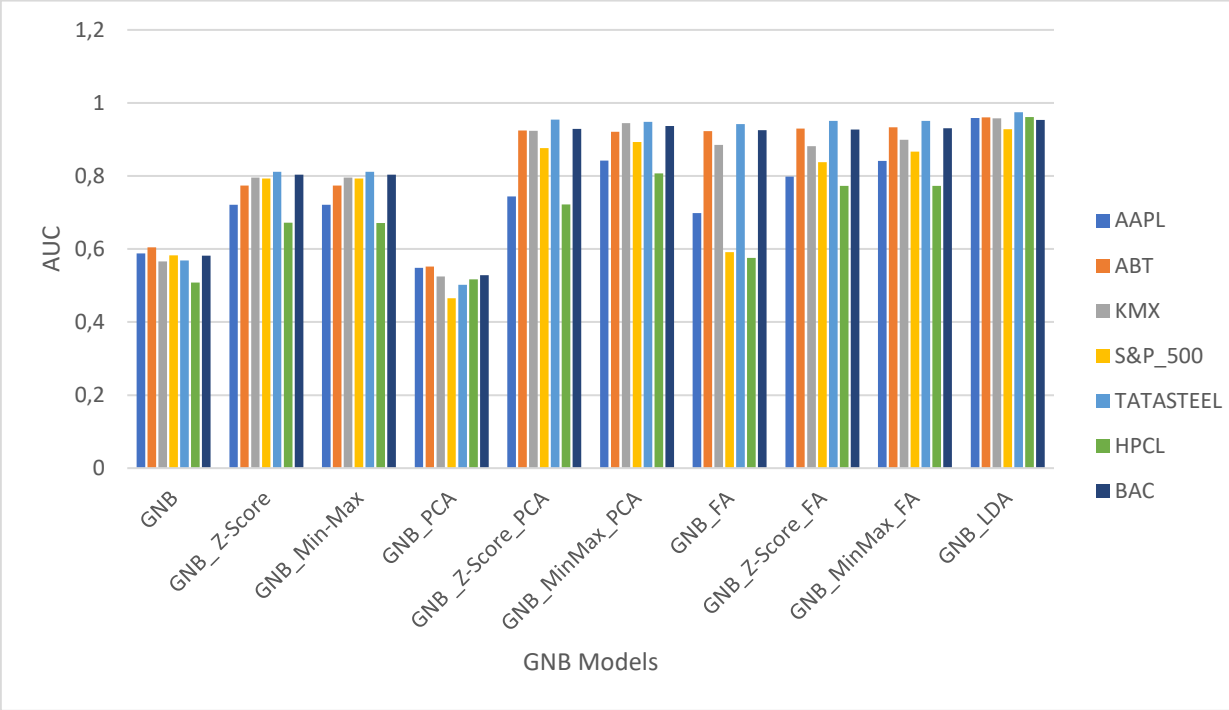

Figure 4: AUC values of the GNB models on the different stock data sets.

DataSets	GNB	GNB_ Z-Score	GNB_ Min-Max	GNB_ PCA	GNB_ Z-Score_PCA
AAPL	0.5883	0.7216	0.7216	0.5487	0.7438
ABT	0.6046	0.7736	0.7736	0.5517	0.9246
KMX	0.5661	0.7958	0.7958	0.5251	0.9238
S&P_500	0.5830	0.7927	0.7927	0.4649	0.8764
TATASTEEL	0.5688	0.8115	0.8115	0.5022	0.9540
HPCL	0.5085	0.6725	0.6708	0.5172	0.7225
BAC	0.5819	0.8037	0.8037	0.5281	0.9291
Mean	**0.5716**	**0.7673**	**0.7671**	**0.5197**	**0.8677**

DataSets	GNB_ Min-Max_PCA	GNB_ FA	GNB_ Z-Score_FA	GNB_ Min-Max_FA	GNB_ LDA
AAPL	0.8421	0.6986	0.7982	0.8416	0.9586
ABT	0.9211	0.9232	0.9296	0.9329	0.9603
KMX	0.9447	0.8850	0.8817	0.8994	0.9581
S&P_500	0.8927	0.5914	0.8382	0.8663	0.9282
TATASTEEL	0.9483	0.9421	0.9507	0.9511	0.9743
HPCL	0.8067	0.5757	0.7726	0.7726	0.9617
BAC	0.9364	0.9251	0.9268	0.9305	0.9532
Mean	**0.8989**	**0.7916**	**0.8711**	**0.8849**	**0.9563**

Table 4: AUC values recorded by the GNB models.

data. The smallest AUC value recorded by any of the models was 0.4649 by GNB_ PCA on the S&P_500 stock data. The mean AUC value of GNB_LDA model (0.9563) was the highest mean AUC recorded among the GNB models. The mean AUC of the GNB_PCA model (0.5197) was the least mean AUC among the various models. Figure 4 represents the column chart of the AUC evaluation metric result for the GNB models on the different stock data.

The ROC curves of the various GNB models on AAPL, ABT, KMX, S&P_500, TATASTEEL, HPCL, and BAC stock data sets are presented by Figure 5 to Figure 11 respectively.

Table 5 to Table 8 present the Kendall's coefficient of concordance rankings of the GNB models using accuracy, F1 score, specificity, and AUC evaluation results respectively. The study used a cutoff value of 0.05, and the Kendall's coefficient is considered significant and able to assign ranks to the models when $p < 0.05$ and $\chi^2 > 16.919$.

From Table 5, the Kendall's coefficient was significant to rank the GNB models using the accuracy

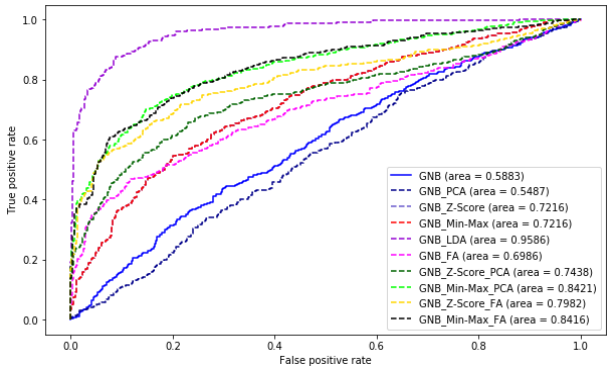

Figure 5: ROC Curves of the GNB models on the AAPL stock data set.

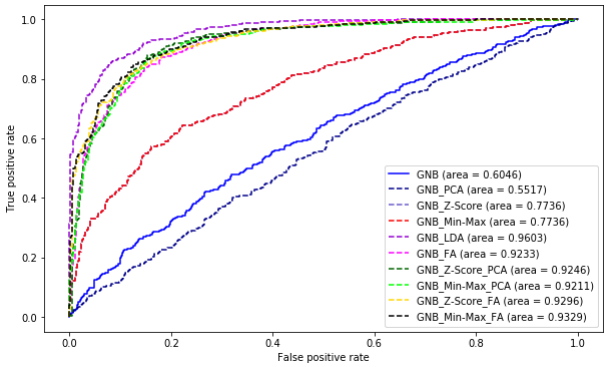

Figure 6: ROC Curves of the GNB models on the ABT stock data set.

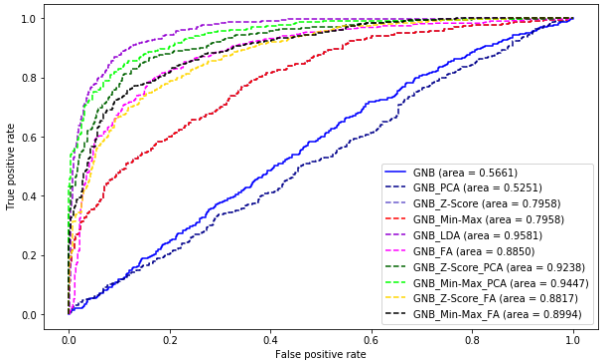

Figure 7: ROC Curves of the GNB models on the KMX stock data set.

results. GNB_LDA model attained the highest rank. The overall ranking of the models was:

GNB_LDA > GNB_Min-Max _PCA > GNB_Min-Max_FA > GNB_Z-Score_PCA > GNB_Z-Score_FA > GNB_FA > GNB_Z-Score = GNB_Min-Max > GNB > GNB_PCA

Table 6 shows that Kendall's coefficient was significant to rank the GNB models using the F1-Score metric. GNB_LDA model generated the highest rank. The overall ranking was given as:

GNB_LDA > GNB_Min-Max _PCA > GNB_Min-Max_FA > GNB_Z-Score_PCA > GNB_Z-Score_FA > GNB_FA > GNB_Z-Score = GNB_Min-Max > GNB_PCA > GNB

Table 7 indicates that Kendall's coefficient is significant to rank the GNB models using specificity

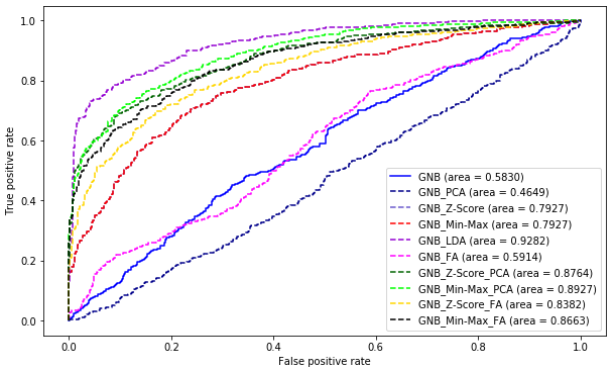

Figure 8: ROC Curves of the GNB models on the S&P_500 index stock data set.

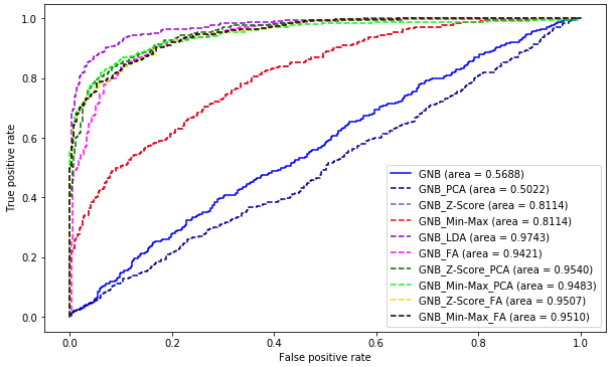

Figure 9: ROC Curves of the GNB models on the TATASTEEL stock data set.

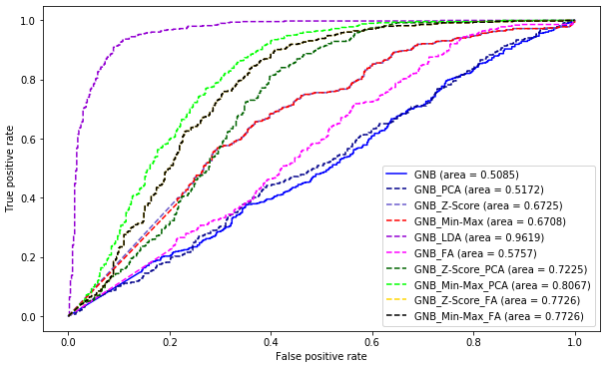

Figure 10: ROC Curves of the GNB models on the TATASTEEL stock data set.

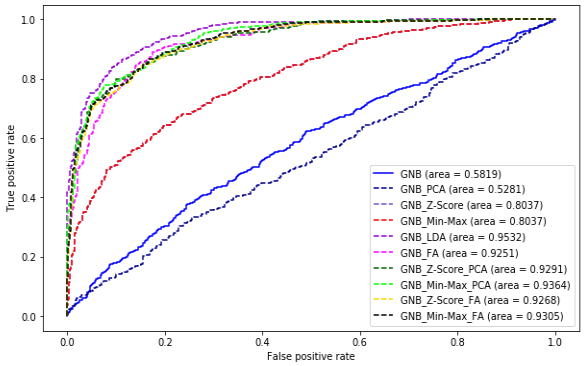

Figure 11: ROC Curves of the GNB models on the HPCL stock data set

Metric	W	$\chi 2$	p	Ranks					
Accuracy	0.84	53.07	0.00	**Technique**	GNB	GNB _ Z-Score	GNB_ Min-Max	GNB_ PCA	GNB _ Z-Score_ PCA
				Mean Rank	2.14	3.79	3.79	1.14	7.29
				Technique	GNB_ Min-Max _PCA	GNB_ FA	GNB_ Z-Score _FA	GNB_ Min-Max _FA	GNB_ LDA
				Mean Rank	7.86	5.29	6.07	7.50	10.00

Table 5: Kendall's W Test of Concordance Rankings of the GNB models using the accuracy metric.

Metric	W	$\chi 2$	p	Ranks					
F1-Score	0.62	39.22	0.00	**Technique**	GNB	GNB_ Z-Score	GNB_ Min-Max	GNB_ PCA	GNB_ Z-Score _PCA
				Mean Rank	2.71	3.36	3.36	3.14	6.43
				Technique	GNB_ Min-Max _PCA	GNB_ FA	GNB_ Z-Score _FA	GNB_ Min-Max _FA	GNB_ LDA
				Mean Rank	7.43	5.00	6.07	7.36	10.00

Table 6: Kendall's W Test of Concordance Rankings of the GNB models using the Specificity metric.

Metric	W	$\chi 2$	p	Ranks					
Specificity	0.70	43.76	0.00	**Technique**	GNB	GNB _ Z-Score	GNB _ Min-Max	GNB_ PCA	GNB _ Z-Score _PCA
				Mean Rank	2.29	3.64	3.64	1.43	7.07
				Technique	GNB_ Min-Max _PCA	GNB _ FA	GNB _ Z-Score _FA	GNB_ Min-Max _ FA	GNB _LDA
				Mean Rank	8.50	6.71	6.93	6.57	8.21

Table 7: Kendall's W Test of Concordance Rankings of the GNB models using the Specificity metric.

Metric	W	$\chi 2$	p	Ranks					
AUC	0.90	56.95	0.00	**Technique**	GNB	GNB _ Z-Score	GNB_ Min-Max	GNB _PCA	GNB _ Z-Score _PCA
				Mean Rank	1.86	4.00	3.86	1.14	7.29
				Technique	GNB_ Min-Max _ PCA	GNB _ FA	GNB_ Z-Score _FA	GNB_ Min-Max _FA	GNB _ LDA
				Mean Rank	8.00	4.43	6.64	7.79	10.00

Table 8: Kendall's W Test of Concordance Rankings of the GNB models using the AUC metric.

metric. GNB_Min-Max _PCA model produced the highest rank. The overall ranking of the models was: GNB_Min-Max _PCA > GNB_LDA > GNB_Z-Score_PCA > GNB_Z-Score_FA > GNB_Min-Max_FA > GNB_FA > GNB_Z-Score = GNB_Min-Max > GNB > GNB_PCA

Table 8 shows that Kendall's coefficient was significant to rank the GNB models using AUC metric. GNB_LDA generated the highest rank. The overall ranking of the GNB models was: GNB_LDA > GNB_Min-Max _PCA > GNB_Min-Max_FA > GNB_Z-Score_PCA > GNB_Z-Score_FA > GNB_FA > GNB_Z-Score > GNB_Min-Max > GNB > GNB_PCA.

The ROC curves of the various GNB models on AAPL, ABT, KMX, S&P_500, TATASTEEL, HPCL, and BAC stock data sets are presented by Figure 5 to Figure 11 respectively.

Table 5 to Table 8 present the Kendall's coefficient of concordance rankings of the GNB models using accuracy,

F1 score, specificity, and AUC evaluation results respectively. The study used a cutoff value of 0.05, and the Kendall's coefficient is considered significant and able to assign ranks to the models when $p < 0.05$ and $\chi^2 > 16.919$.

From Table 5, the Kendall's coefficient was significant to rank the GNB models using the accuracy results. GNB_LDA model attained the highest rank. The overall ranking of the models was:
GNB_LDA > GNB_Min-Max _PCA > GNB_Min-Max_FA > GNB_Z-Score_PCA > GNB_Z-Score_FA > GNB_FA > GNB_Z-Score = GNB_Min-Max > GNB > GNB_PCA

Table 6 shows that Kendall's coefficient was significant to rank the GNB models using the F1-Score metric. GNB_LDA model generated the highest rank. The overall ranking was given as:
GNB_LDA > GNB_Min-Max _PCA > GNB_Min-Max_FA > GNB_Z-Score_PCA > GNB_Z-Score_FA > GNB_FA > GNB_Z-Score = GNB_Min-Max > GNB_PCA > GNB

Table 7 indicates that Kendall's coefficient is significant to rank the GNB models using specificity metric. GNB_Min-Max _PCA model produced the highest rank. The overall ranking of the models was: GNB_Min-Max _PCA > GNB_LDA > GNB_Z-Score_PCA > GNB_Z-Score_FA > GNB_Min-Max_FA > GNB_FA > GNB_Z-Score = GNB_Min-Max > GNB > GNB_PCA

Table 8 shows that Kendall's coefficient was significant to rank the GNB models using AUC metric. GNB_LDA generated the highest rank. The overall ranking of the GNB models was:
GNB_LDA > GNB_Min-Max _PCA > GNB_Min-Max_FA > GNB_Z-Score_PCA > GNB_Z-Score_FA > GNB_FA > GNB_Z-Score > GNB_Min-Max > GNB > GNB_PCA.

5 Conclusion

This study assessed how the GNB algorithm performed with different feature scaling (i.e., standardization scaling, and Min-Max scaling techniques) and feature extraction techniques (i.e., PCA, LDA, and FA) in predicting the direction of movement of stock price using stock data randomly collected from different stock markets. The performance of the various GNB models were evaluated using accuracy, F1-Score, specificity and AUC evaluation metrics. Kendall's W test of concordance was used to generate ranks for the GNB models using the evaluation metrics.

The experimental results indicated that application of scaling techniques improved the performance of the GNB model. Models based on integration of GNB algorithm, feature scaling technique and feature extraction technique generated results which were superior to results produced by models based on either integration of GNB algorithm and feature scaling technique or GNB algorithm and feature extraction technique with the exception of GNB_LDA. In general, the model based on integration of GNB algorithm and Linear Discriminant Analysis (GNB_LDA) outperformed all the other models of GNB considered in three of the four evaluation metrics (i.e., accuracy, F1-score, and AUC). Similarly, the predictive model based on GNB algorithm, Min-Max scaling, and PCA produced the best rank using the specificity results. In addition, GNB produced better performance with Min-Max scaling technique than it does with standardization scaling techniques.

6 Acknowledgement

This work was supported by the NSFC-Guangdong Joint Fund (Grant No. U1401257), National Natural Science Foundation of China (Grant Nos. 61300090, 61133016, and 61272527), science and technology plan projects in Sichuan Province (Grant No. 2014JY0172) and the opening project of Guangdong Provincial Key Laboratory of Electronic Information Products Reliability Technology (Grant No. 2013A061401003)

7 Reference

[1] Fama E. F, Fisher L, Jensen M, Roll R (1969) The adjustment of stock price to new information. Int Eco Rev 10(1):1–21

[2] Yeh, I.-C., & Hsu, T.-K. (2014). Exploring the dynamic model of the returns from value stocks and growth stocks using time series mining. Expert Systems with Applications, 41, 7730–7743. https://doi.org/10.1016/j.eswa.2014.06.036

[3] Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. Journal of Computational Science, 2, 1–8. https://doi.org/10.1016/j.jocs.2010.12.007

[4] Smith, V. L. (2003). Constructivist and ecological rationality in economics. American Economic Review, 93, 465–508. https://doi.org/10.1257/000282803322156954

[5] Gandhmal, D. P., & Kumar, K. (2019). Systematic analysis and review of stock market prediction techniques. Computer Science Review, 34, 100190. https://doi.org/10.1016/j.cosrev.2019.08.001

[6] Huang, C.-J., Yang, D.-X., & Chuang, Y.-T. (2008). Application of wrapper approach and composite classifier to the stock trend prediction. Expert Systems with Applications, 34(4), 2870–2878. https://doi.org/10.1016/j.eswa.2007.05.035

[7] Huang, W., Nakamori, Y., &Wang, S.-Y. (2005). Forecasting stock market movement direction with support vector machine. Computers & Operations Research, 32(10), 2513–2522. https://doi.org/10.1016/j.cor.2004.03.016

[8] Maragoudakis M., Serpanos D. (2015). Exploiting Financial News and Social Media Opinions for Stock Market Analysis using MCMC Bayesian Inference. Computational Economics. DOI 10.1007/s10614-015-9492-9. https://doi.org/10.1007/s10614-015-9492-9

[9] Iqbal, N., & Islam, M. (2019). Machine learning for dengue outbreak prediction: A performance

evaluation of different prominent classifiers. Informatica, 43(3), 361 -371. https://doi.org/10.31449/inf.v43i3.1548

[10] Abaker, A. A., & Saeed, F. A. (2021). A Comparative Analysis of Machine Learning Algorithms to Build a Predictive Model for Detecting Diabetes Complications. Informatica, 45(1), 117 -125. https://doi.org/10.31449/inf.v45i1.3111

[11] Zhang, Y., & Wu, L. (2009). Stock market prediction of s&p 500 via combination of improved bco approach and bp neural network. Expert Systems with Applications, 36 (5), 8849–8854. https://doi.org/10.1016/j.eswa.2008.11.028
Meesad, P., & Rasel, R. I. (2013). Predicting stock market price using support vector regression. In Informatics, electronics & vision (iciev), 2013 international conference on informatics. IEEE, 2013, 1–6. https://doi.org/10.1109/iciev.2013.6572570

[12] Zhou, Z., Gao, M., Liu, Q., & Xiao, H. (2020). Forecasting stock price movements with multiple data sources: Evidence from stock market in China. Physica A: Statistical Mechanics and its Applications, 542, 123389. https://doi.org/10.1016/j.physa.2019.123389

[13] Ampomah, E. K., Qin, Z., Nyame, G., & Botchey, F. E. (2021). Stock market decision support modeling with tree-based AdaBoost ensemble machine learning models. Informatica, 44(4), 363 – 375 https://doi.org/10.31449/inf.v44i4.3159

[14] Kumar, M., & Thenmozhi,M. (2006). Forecasting Stock index movement: A comparison of support vector machines and random forest. SSRN Scholarly Paper. Rochester, NY: Social Science Research Network, January 24, 2006. https://doi.org/10.2139/ssrn.876544

[15] Ou, P., & Wang, H. (2009). Prediction of stock market index movement by ten data mining techniques. Modern Applied Science, 3(12), 28. https://doi.org/10.5539/mas.v3n12p28

[16] Subha, M. V., & Nambi, S. T. (2012). Classification of Stock Index movement using K-nearest neighbours (k-NN) algorithm. WSEAS Transactions on Information Science & Applications, 9(9), 261–270. P259.

[17] Saifan R, Sharif K, Abu-Ghazaleh M, Abdel-Majeed M. Investigating Algorithmic Stock Market Trading Using Ensemble Machine Learning Methods. Informatica. 2020 Sep 15;44(3), 311-325 https://doi.org/10.31449/inf.v44i3.2904

[18] Zikowski, K. (2015). Using volume weighted support vector machines with walk forward testing and feature selection for the purpose of creating stock trading strategy. Expert Systems with Applications, 42, 1797–1805. https://doi.org/10.1016/j.eswa.2014.10.001

[19] Patel, J., Shah, S., Thakkar, P., & Kotecha, K. (2015). Predicting stock and stock price index movement using Trend Deterministic Data Preparation and machine learning techniques. Expert Systems with Applications, 42, 259–268. https://doi.org/10.1016/j.eswa.2014.07.040

[20] Sun, S., Wei, Y., & Wang, S. (2018). AdaBoost-LSTM Ensemble Learning for Financial Time Series Forecasting. Computational Science – ICCS 2018, 590–597. https://doi:10.1007/978-3-319-93713-7_55

[21] Khan, W., Ghazanfar, M.A., Azam, M.A. et al. (2020). Stock market prediction using machine learning classifiers and social media, news. J Ambient Intell Human Comput. https://doi.org/10.1007/s12652-020-01839-w

[22] Bhandare, Y., Bharsawade, S., Nayyar, D., Phadtare, O., & Gore, D. (2020). SMART: Stock Market Analyst Rating Technique Using Naive Bayes Classifier. 2020 International Conference for Emerging Technology (INCET). https://doi:10.1109/incet49848.2020.9154002

[23] Saranya C. Manikandan G. (2013). A study on normalization techniques for privacy preserving data mining. International Journal of Engineering and Technology (IJET). 5(3):2701-2714

[24] Abdi H. Williams L. J. (2010). Principal component analysis. Wiley interdisciplinary reviews: computational statistics. 2(4):433-459 https://doi.org/10.1002/wics.101

[25] Bro R. Smilde A. K. (2014) Principal component analysis. Analytical Methods. 6(9):2812-2831

[26] Tharwat A. Gaber T. Ibrahim A. Hassanien A. E. Linear discriminant analysis: A detailed tutorial. AI communications. 2017, 30(2):169-190 https://doi.org/10.3233/aic-170729

[27] Maskey R. Fei J. Nguyen H. O. (2018), Use of exploratory factor analysis in maritime research. The Asian journal of shipping and logistics. 34(2):91-111 https://doi.org/10.1016/j.ajsl.2018.06.006

[28] Kendall, M. G. Babington, S. B. (1939). The Problem of m Rankings, The Annals of Mathematical Statistics, 10, 275-287.

8 Appendix

Data Set	Stock Market	Time Frame	Number of Sample
AAPL	NASDAQ	2005-01-01 to 2019-12-30	3773
ABT	NYSE	2005-01-01 to 2019-12-30	3773
BAC	NYSE	2005-01-01 to 2019-12-30	3773
S&P_500	INDEXSP	2005-01-01 to 2019-12-30	3773
HPCL	NSE	2005-01-01 to 2019-12-30	3278
KMX	NYSE	2005-01-01 to 2019-12-30	3773
TATASTEEL	NSE	2005-01-01 to 2019-12-30	3476

Table A1: Detail of stock data sets used.

Volume Indicator	Description
Chaikin A/D Line (ADL)	Estimates the Advance/Decline of the market.
Chaikin A/D Oscillator (ADOSC)	Indicator of another indicator. It is created through application of MACD to the Chaikin A/D Line
On Balance Volume (OBV)	Uses volume flow to forecast changes in price of stock

Table A2: Description of Volume Indicators used in the study.

Price Transform Indicator	Description
Median Price (MEDPRICE)	Measures the mid-point of each day's high and low prices.
Typical Price (TYPPRICE)	Measures the average of each day's price.
Weighted Close Price (WCLPRICE)	Average of each day's price with extra weight given to the closing price.

Table 3: Description of Price Transform Function.

Overlap Studies Indicators	Description
Bollinger Bands (BBANDS)	Describes the different highs and lows of a financial instrument in a particular duration.
Weighted Moving Average (WMA)	Moving average that assign a greater weight to more recent data points than past data points
Exponential Moving Average (EMA)	Weighted moving average that puts greater weight and importance on current data points, however, the rate of decrease between a price and its preceding price is not consistent.
Double Exponential Moving Average (DEMA)	It is based on EMA and attempts to provide a smoothed average with less lag than EMA.
Kaufman Adaptive Moving Average (KAMA)	Moving average designed to be responsive to market trends and volatility.
MESA Adaptive Moving Average (MAMA)	Adjusts to movement in price based on the rate of change of phase as determined by the Hilbert transform discriminator.
Midpoint Price over period (MIDPRICE)	Average of the highest close minus lowest close within the look back period
Parabolic SAR (SAR)	Heights potential reversals in the direction of market price of securities.
Simple Moving Average (SMA)	Arithmetic moving average computed by averaging prices over a given time period.
Triple Exponential Moving Average (T3)	It is a triple smoothed combination of the DEMA and EMA
Triple Exponential Moving Average (TEMA)	An indicator used for smoothing price fluctuations and filtering out volatility. Provides a moving average having less lag than the classical exponential moving average.
Triangular Moving Average (TRIMA)	Moving average that is double smoothed (averaged twice)

Table A3: Description of Overlap Studies Indicators used in the study.

Momentum Indicators	Description
Average Directional Movement Index (ADX)	Measures how strong or weak (strength of) a trend is over time
Average Directional Movement Index Rating (ADXR)	Estimates momentum change in ADX.
Absolute Price Oscillator (APO)	Computes the differences between two moving averages
Aroon	Used to find changes in trends in the price of an asset
Aroon Oscillator (AROONOSC)	Used to estimate the strength of a trend
Balance of Power (BOP)	Measures the strength of buyers and sellers in moving stock prices to the extremes
Commodity Channel Index (CCI)	Determine the price level now relative to an average price level over a period of time
Chande Momentum Oscillator (CMO)	Estimated by computing the difference between the sum of recent gains and the sum of recent losses
Directional Movement Index (DMI)	Indicate the direction of movement of the price of an asset
Moving Average Convergence /Divergence (MACD)	Uses moving averages to estimate the momentum of a security asset
Money Flow Index (MFI)	Utilize price and volume to identify buying and selling pressures
Minus Directional Indicator (MINUS_DI)	Component of ADX and it is used to identify presence of downtrend.
Momentum (MOM)	Measurement of price changes of a financial instrument over a period of time
Plus Directional Indicator (PLUS_DI)	Component of ADX and it is used to identify presence of uptrend.
Log Return	The log return for a period of time is the addition of the log returns of partitions of that period of time. It makes the assumption that return**s** are compounded continuously rather than across sub-periods
Percentage Price Oscillator (PPO)	Computes the difference between two moving averages as a percentage of the bigger moving average
Rate of change (ROC)	Measure of percentage change between the current price with respect to a at closing price n periods ago.
Relative Strength Index (RSI)	Determines the strength of current price in relation to preceding price
Stochastic (STOCH)	Measures momentum by comparing closing of a security with earlier trading range over a specific period of time
Stochastic Relative Strength Index (STOCHRSI)	Used to estimate whether a security is overbought or oversold. It measures RSI over its own high/low range over a specified period.
Ultimate Oscillator (ULTOSC)	Estimates the price momentum of a security asset across different time frames.
Williams' %R (WILLR)	Indicates the position of the last closing price relative to the highest and lowest price over a time

Table A4: Description of Momentum Indicators used in the study.