# Predicting the Direction of E-Commerce Stock Prices during COVID 19 Using Machine Learning

# التنبؤ باتجاه أسعار أسهم التجارة الالكترونية خلال كوفيد 19 باستخدام التعلم الآلي

**Bouriche Nouria[1], Benbouziane Mohamed[2]**

[1] University of Tlemcen, MIFMA Laboratory Faculty of Economics and Business - Tlemcen (Algeria), nouria.bouriche@univ-tlemcen.dz

[2] University of Tlemcen, MIFMA Laboratory Faculty of Economics and Business - Tlemcen (Algeria),
mohamed.benbouziane@univ-tlemcen.dz

**Received:** 10/01/2022 **Accepted:** 30/03/2022 **Published:**31/03/2022

**Abstract:**

The study investigates the predictive power of COVID 19 on E-Commerce stocks in the United States and China. Technical and COVID 19 indicators were used as input features to predict the stock prices trend using three machine learning classifiers RF, SVM and KNN. The performance of the classifiers is compared before and after performing feature selection technique with Random Forest Feature Importance on the dataset. The results indicate that COVID 19 indicators have a predictive power on stock prices movements in both the US and Chinese markets. The accuracy and F1 Score metrics improve when using feature selection. The performance of the classifiers with the selected features shows that RF outperforms the other classifiers with the highest accuracy and F1 Score followed by SVM and KNN respectively.

**Keywords:** COVID 19, E-Commerce, Feature Selection, Stock Prediction, Machine Learning.

**JEL Classification Codes:** I1, L81, C52, C53, G17

**ملخص:**

تهدف الدراسة إلى تقصي القدرة التنبؤية لفيروس كورونا على أسهم التجارة الإلكترونية في الولايات المتحدة الأمريكية و الصين. تم إستعمال مؤشرات التحليل التقني و حالات الإصابة و الوفاة المسجلة جراء الإصابة بالفيروس كمدخلات للتنبؤ باتجاه أسعار الأسهم بإستخدام خوارزميات التعلم الآلي: شعاع الدعم الآلي $SVM$، الغابة العشوائية $RF$ و الجار الأقرب $KNN$. تمت مقارنة أداء الخوارزميات قبل و بعد تطبيق طريقة الغابة العشوائية لإنتقاء أهم المدخلات. أظهرت نتائج الدراسة أن مؤشرات فيروس كورونا لها قوة تنبؤية على حركة أسعار أسهم السوق الأمريكي و الصيني. كل من مؤشر دقة التنبؤ و $F1$ تحسن بعد إستعمال خوارزمية الانتقاء. نموذج الغابة العشوائية $RF$ تفوق على باقي النماذج متبوعا ب $SVM$ و $KNN$ على التوالي.

**كلمات مفتاحية:** كوفيد 19، التجارة الالكترونية، اختيار العناصر، التنبؤ بأسعار الأسهم ، التعلم الآلي

**تصنيفات JEL** : I1, L81, C52, C53, G17

**Corresponding author**: Bouriche Nouria, **e-mail:** nouria.bouriche@univ-tlemcen.dz

## INTRODUCTION:

The COVID19 has infected 44.03 Million people worldwide and killed 1.16 Million by 27 October 2020 (JHU, 2021). The pandemic has caused a disruption in different sectors such as education, transportation, trade, healthcare while other industries have benefited from the situation with the help of digital technologies and the internet like the E-commerce industry. Lockdowns measures have shifted consumers to online shopping; their behavior was changed significantly especially when preparing for quarantine (Laato, Islam, Farooq, & Dhir, 2020). A study by (Bhatti et al., 2020) finds that 52% of consumers have avoided physical shopping and public gathering while 36% confirmed that they would continue the distancing until being vaccinated. Ecommerce revenues rose from $1.4 trillion in 2017 to $2.4 trillion in 2020. The largest markets are China, US, Japan, UK and Germany. Demand on goods and services increased for healthcare products, masks, sanitizers, gloves, household and technology (Alfonso C, Boar, Frost, Gambacorta, & Liu, 2021).

The share of E-commerce in total retail has increased significantly between the first two quarters of 2020 by 16% in the US and 31% in the UK while china has witnessed an increase of 24,6% from January to august 2020 (OECD, 2020). COVID19 increased the demand on online shopping; there has been a parallel change between E-Commerce net sales and stock prices. Positive high correlation was found between Amazon stock prices and consumer demand (İbiş, Işık, & Gulseven, 2021). The Virus spread has significantly affected the stock market, an increase in COVID19 cases is followed by an increase in stock prices (Mottaghi & Farhangdoost, 2021) . Predicting the stock market trend in the pandemic is a critical task for investors, understanding and analyzing the market variations caused by the COVID19 can help to better take investment strategies and minimize the risk exposure. Machine learning has been widely used in stock prediction; classification algorithms have demonstrated a remarkable performance in stock movement prediction while used with the appropriate features and techniques. (Xianya, Mo, & Haifeng, 2019) use four classification algorithms Random Forest, Decision Tree, Naive Bayes and Logistic Regression with Spark platform for stock prediction. The results of their study indicate that Random Forest and Decision Tree algorithms had the highest accuracy and the AUC and PR values. The classifiers demonstrate a good performance in stock prediction with a reduced error values. (Patel, Shah, Thakkar, & Kotecha, 2015) Predict the direction of Indian stock market using ANN, Random Forest, Naive Bayes and SVM. The study used two datasets to compare between two approaches, the first one by using a dataset that includes technical indicators as continuous values and the second by introducing these indicators as trend deterministic. Results show that the trend deterministic indicators dataset is better for stock trend prediction. The classifier algorithms perform almost identically and their accuracy improve when learning from trend deterministic technical indicators. (Kumar, Dogra, Utreja, & Yadav, 2018) Find that random forest classifier outperform KNN, SVM and Naive Bayes in term of prediction accuracy when using large technical indicator dataset. Naive Bayes classifier has the highest accuracy among other classifiers when the indicators size is reduced but the performance of all the classification algorithms drop in the case of using the small size dataset. (Khan, Ghazanfar, Assam, Ahmad, & Khan, 2016) compare the accuracy and error of KNN, SVM and Naive Bayes algorithms in London, New York and Karachi stock exchange markets. Principal Component Analysis (PCA) is applied to select the best features in the stock price datasets. The results confirm that feature selection is important in stock trend determination since the accuracy of the models increase and the MAE is reduced after using PCA. KNN and SVM are found to be the best classifiers with the highest accuracy and the lowest MAE.

The current study investigates the following problematic:

Do COVID 19 cases and deaths have a predictive power on stock movements?

Does the feature importance technique improve the accuracy of the ML classifiers?

Which of the ML classifiers has the highest prediction accuracy when applied as a tool for stock price prediction?

We use in our research feature selection technique and machine learning classifiers with technical indicators and COVID-19 attributes as features to predict the movement of E-Commerce stocks in two developed markets US and China. The accuracy of the models before and after feature selection is evaluated to select the classifier with the best performance.

## 2- Methodology:

We first import the dataset of E-Commerce stocks and COVID 19 cases and deaths. Our first analysis approach was to extract eight technical indicators and two COVID 19 indicators to predict the trend of the selected stocks in US and China using machine learning classifiers RF, SVM and KNN. In the second approach, we apply feature selection technique on the selected dataset using Random Forest Feature Importance. The results of the prediction before and after Feature Selection are compared using two model evaluation metrics accuracy and F1 Score. We finally evaluate the prediction performance of the classifiers with the same metrics.

Data ⇨ Preprocessing ⇨ ML Classifiers ⇨ Model Evaluation

Data ⇨ Preprocessing ⇨ Feature Selection ⇨ ML Classifiers ⇨ Model Evaluation

## 2-1- Machine Learning Models Description:

### 2-1-1- Random Forest:

Random Forest is a technique of combination of many decision tree algorithms referred as ensemble modeling (Breiman, 2001). It handles large datasets with higher dimensionality. Random Forest learns from a random sample of data and train each decision tree on a different set of data points by splitting the dataset into subsets to predict the target variable. The nodes in each tree are split with a limited number of features until the maximal depth of the tree is reached and the final prediction in classification problems is the majority of votes generated from all the trees. The technique of improving the results by combining different learning models is called bagging; it reduces the variance and overfitting and improves the accuracy of the classifier. Random forest algorithm also identifies the most contributing features in the dataset by calculating the relevance score of each feature in the training set.

The node importance indicator is Gini Index (mean decrease in node impurity), it measures how each feature decreases the impurity of the split. The value is averaged on each single tree and the average of all the values from all the trees is the feature importance.

$$Gini = \sum_{i=1}^{n} p_i\,(1 - p_i) = 1 - \sum_{i=1}^{n} {p_i}^2$$

$p_i$: Probability of choosing class i element

n: number of classes.

Features at the top of the trees are considered more important. The impurity of the node before and after the split is compared to find how much improvement is made in the process. The best split is reached with lowest impurity.

### 2-1-2- Support Vector Machine:

Support Vector Machine is a supervised machine learning technique that is used for classification and regression tasks (V, 1998). The SVM algorithm separates the data points into two classes with a hyperplane. The closest points of each class to the hyperplane are called support vectors. They separate the two classes of the data with parallel hyperplanes from each side of the data. The distance between these closets points also referred to as (margin). It should be maximum to obtain the best hyperplane. (Trafalis, 1999) propose dual Interior Point Methods (IPM) to solve the quadratic program described below:

Quadratic Program (Primal Form):

$$\min(P) = \frac{1}{2}||w||^2 + C\sum_{1}^{n} \xi_i \tag{1}$$

s.t $y_i[w^T\phi(x_i) + b] \geq 1 - \xi_i \quad \forall i$
$\xi_{i\geq 0}$

The Lagrange multiplier is introduced to the formal primal minimization program as follow:

$$L(w, b, \xi, \alpha) = \frac{1}{2}w^2 + C\sum_{i=1}^{n} \xi_i - \sum_{i=1}^{n} \alpha_i\,(y_i(w^T\phi(X_i) + b) - 1 + \xi i \tag{2}$$

The dual form of the Quadratic Primal becomes:

$$\max(D) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} y_i \, \alpha_i y_j \alpha_j K(X_i, X_j) \qquad (3)$$

$0 \leq \alpha_i \leq C$
S.t $\sum_i y_i \, \alpha_i = 0$ where $K(X_i, X_j)$ is the Kernel function
$\forall i$

### 2-1-3- KNN:

K nearest neighbor is a supervised Machine Learning algorithm that learns from a training set and classifies new data points based on similarity. When new data points are given, KNN calculates the distance between these data points and their neighbors from the available classes, based on K values and the least distance. The algorithm generates a response by sorting the training records and assigns the class of the new data points according to majority votes for classification problems (Cover & Hart, 1967).

The most widely used similarity metric to calculate the distance between point A and B is the Euclidean distance:

$$D = \sqrt{(X_b - X_a)^2 - (Y_b - Y_a)^2} \qquad (4)$$

## 2-2- Data Preprocessing:

**2-2-1- Import dataset**: Two datasets are imported from January 14, 2020 to May 13, 2020. The first one consists of daily stock prices (open, high, low, close, Volume) of the trading day of the largest E-commerce companies in the US and Chinese markets. The second dataset represents daily statistics on the coronavirus cases and deaths. Stock prices were imported from yahoo finance while COVID-19 statistics were retrieved from Johns Hopkins University COVID-19 data repository.

### 2-2-2- Feature Extraction:

Technical analysis has been extensively used in stock prediction. Eight technical indicators are extracted in this study (SMA, EMA, Stochastic Oscillator, RSI, MACD, Williams's %R). In addition to that, we have calculated the daily change of coronavirus new cases and new deaths from the COVID dataset. Values of 0 and 1 were assigned to the target variable to represent the upward and downward trend of the daily closing stock prices.

### 2.2.3. Data cleaning:

After calculating the technical indicators and the COVID-19 daily changes in new cases and deaths, the unnecessary rows and columns were deleted from the datasets, all the missing values were handled and the two datasets with the technical indicators and COVID-19 statistics were combined in a single dataset.

**Table (1): Variables of the study**

| |
|---|
| Simple moving average (SMA):<br>SMA= $\sum^{n}_{i=1} (C / n)$ |
| Exponential moving average (EMA):<br>EMA = K × (C - PEMA) + PEMA |
| Stochastic Oscillator:<br>%K = ( (C - L14) / (H14 - L14) ) × 100<br>%D = 3 day SMA (%K) |
| Relative Strength Index (RSI):<br>RSI = 100 - (100 / (1+RS) ) |
| Moving average convergence divergence (MACD):<br>MACD = 12 days EMA(C ) - 26 days EMA(C )<br>Signal Line = 9 days EMA (MACD) |
| Williams %R = ( ( H14 - C ) / ( H14 - L14 ) ) |
| Target variable: Y ( Stock trend)<br>PCT (Close) = $(C_t – C_{t-1}) / C_t \times 100$<br>PCT $> 0 \rightarrow$ Y = 1<br>PCT $< 0 \rightarrow$ Y = 0 |

**Note:** C : closing price of the stock in the trading day, K: exponential smoothing constant, PEMA: exponential moving average of the previous period, L14: lowest price of the stock in the last 14 days, H14: highest price of the stock in the last 14 days, RS: (average gain/average loss), PCT: percentage change of the close price, $C_{t-1}$: first lag of the closing price .

### 2-2-4- Data splitting:

Before the application of the machine learning classifiers, data are split into training and testing sets. 70% of our data are used as a training set while the resting 30 % are kept for testing the ability of our models in predicting the outcomes.

### 2-2-5- Feature Selection:

The input variables used to train the Machine Learning model has a huge impact on the performance of the model. The feature selection process consists of removing the less important features from the dataset and identify the variables that most contribute to the prediction of our target variable which helps to reduce overfitting problems and achieve a better prediction accuracy. We use tree based feature importance technique with random forest classifier to find the most relevant features to the target variable according to higher mean values.

### 2-2-6- Feature scaling:

Scaling the input features within a specific range is important before applying the machine learning classifier. Normalization of the independent variables in a common ground reduces the superiority of high range variables in the dataset when training the model hence, all the features contribute proportionally to the final output. In this Study, feature scaling is applied only on SVM and KNN since they are algorithms that compute distance and do not rely on rules as decision tree algorithms (random forest). MinMax scaler was calculated for each feature in the dataset as follow:

$$X\,scaled = \frac{(\mathrm{X} - min)}{(max - min)} \qquad (5)$$

X: The original value

The feature range is [0 - 1], that will generate smaller standard deviation and reduce the effect of outliers.

## 2-3- Model Evaluation Metrics:

To evaluate the performance of each classifier in the prediction, two evaluation metrics are used, Accuracy and F1 score. Accuracy represents the proportion of true outcomes from the total investigated cases; it demonstrates the correct predictions generated by the model among total prediction outputs.

$$Accuracy = \frac{(TP+TN)}{(TP+FP+FN+TN)} \qquad (6)$$

TP: True Positive FP: False Positive FN: True Negative

F1 score combines precision, recall in one metric, it identifies a class (upward) from a class (downward) in our stock trend prediction case and specify each class from both classes. It is the weighted average of precision and recall.

$$Precision = \frac{TP}{(TP+FP)} \qquad (7)$$

$$Recall = \frac{TP}{(TP+FN)} \qquad (8)$$

$$F1\ Score = 2 \times \frac{(Percision \times \mathrm{Recall})}{(Percision + \mathrm{Recall})} \qquad (9)$$

## 3- Results and Discussion:

Table (2) shows the results of feature selection technique on the chosen datasets. Technical indicators along with COVID data were found to be the most important predictors of stock movements. The COVID-19 indicators were selected as most contributing features by the RF algorithm in all the datasets, which indicates an existing effect of the pandemic on E-

Commerce stock movements in both US and Chinese markets.

**Table (2): Results of Feature Selection with Random Forest Feature Importance**

| **Dataset** | **Δ cases** | **Δ deaths** |
|---|---|---|
| **AMZN** | | × |
| **EBAY** | | × |
| **TGT** | × | |
| **KR** | × | |
| **BABA** | × | |
| **JD** | × | |
| **PDD** | × | × |
| **3690.HK** | × | |

**Source**: Python output

Table (3) and Figure (1) show the comparison of the accuracy scores of RF, SVM and KNN models before and after feature selection. The accuracy of the selected classifiers improved with the application of feature selection technique in the majority of the datasets.

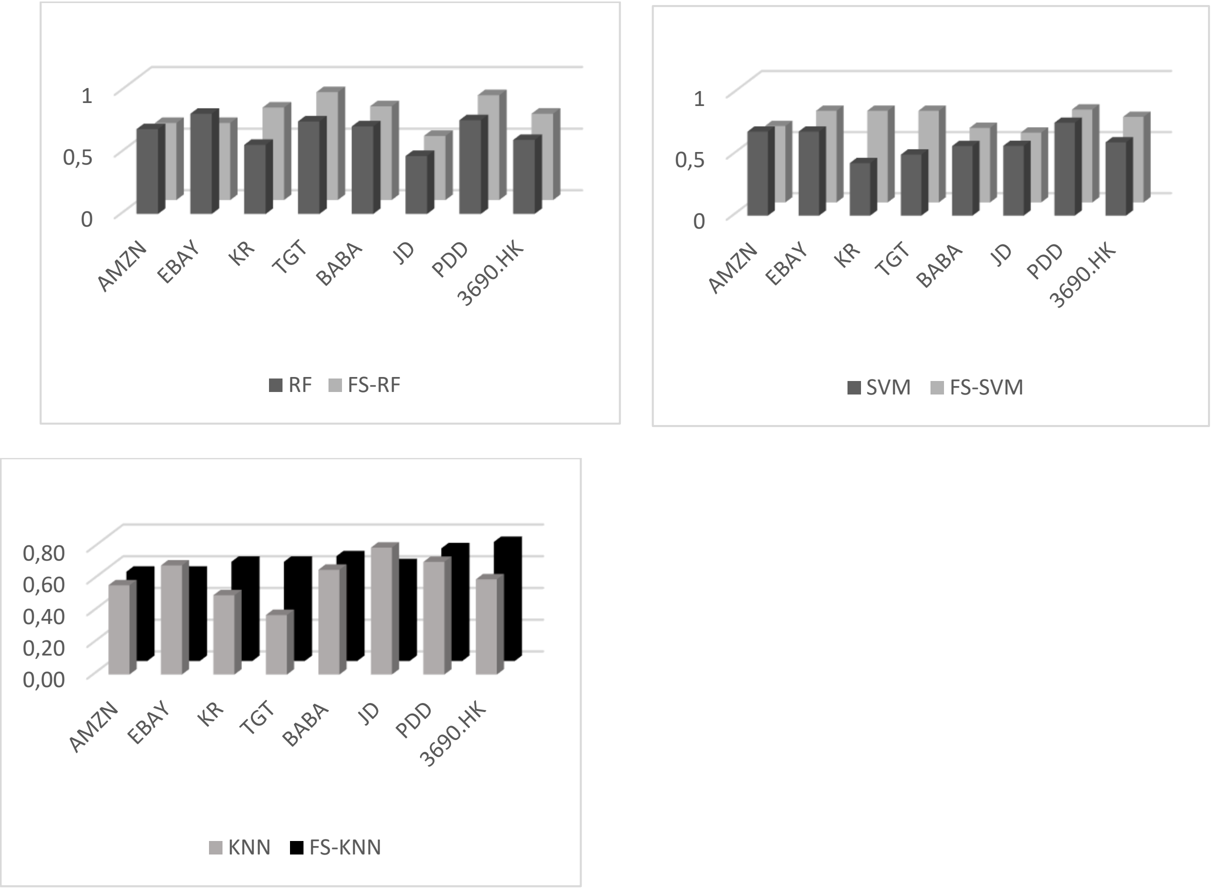
**Table (3): Results of Accuracy Score before and after Feature Selection**

| **Dataset** | **RF** | **FS-RF** | **SVM** | **FS-SVM** | **KNN** | **FS-KNN** |
|---|---|---|---|---|---|---|
| **AMZN** | **0,69** | 0,63 | **0,69** | 0,63 | 0,56 | 0,56 |
| **EBAY** | **0,81** | 0,63 | 0,69 | **0,75** | **0,69** | 0,56 |
| **KR** | 0,56 | **0,75** | 0,43 | **0,75** | 0,50 | **0,63** |
| **TGT** | 0,75 | **0,88** | 0,50 | **0,75** | 0,38 | **0,63** |
| **BABA** | 0,71 | **0,76** | 0,57 | **0,61** | 0,66 | 0,66 |
| **JD** | 0,47 | **0,52** | 0,57 | 0,57 | **0,80** | 0,61 |
| **PDD** | 0,76 | **0,85** | 0,76 | 0,76 | 0,71 | 0,71 |
| **3690.HK** | 0,60 | **0,70** | 0,60 | **0,70** | 0,60 | **0,75** |

**Source:** Python output

**Note:** RF: random forest, FS-RF: random forest with selected features, SVM: support vector machine, FS-SVM: support vector machine with selected features, KNN: k nearest neighbors, FS-KNN: k nearest neighbors with selected features.

**Figure (1): Accuracy before and after feature selection**



**Source**: Excel output

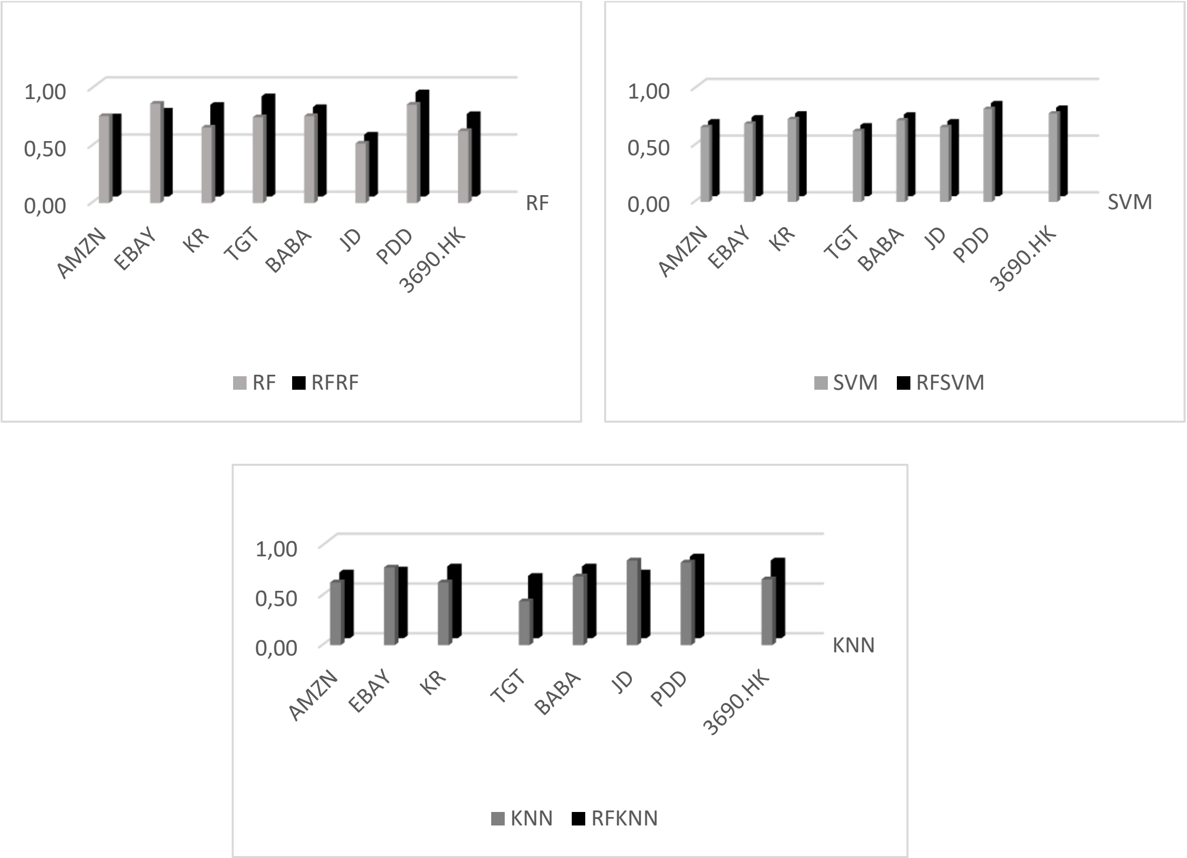
Table (4) and Figure (2) demonstrate the comparison of F1 Score of RF, SVM and KNN before and after feature selection. The F1 Scores were better in the majority of the datasets when the RF feature importance technique is used. Since the COVID-19 indicators are selected as the most important features, they represent relevant predictors of stock trend and hence improve the prediction accuracy when included in the analysis.

**Table (4): Results of F1 Score before and after Feature Selection**

| Dataset | RF | FS-RF | SVM | FS-SVM | KNN | FS-KNN |
|---|---|---|---|---|---|---|
| **AMZN** | **0,76** | 0,70 | 0,66 | 0,66 | 0,63 | **0,66** |
| **EBAY** | **0,87** | 0,75 | 0,69 | **0,70** | **0,78** | 0,69 |
| **KR** | 0,66 | **0,80** | 0,73 | 0,73 | 0,63 | **0,72** |
| **TGT** | 0,75 | **0,88** | 0,63 | 0,63 | 0,44 | **0,63** |
| **BABA** | 0,76 | **0,78** | 0,72 | 0,72 | 0,69 | **0,72** |
| **JD** | 0,52 | **0,54** | 0,66 | 0,66 | **0,85** | 0,66 |
| **PDD** | 0,86 | **0,91** | 0,82 | 0,82 | **0,83** | 0,82 |
| **3690.HK** | 0,63 | **0,72** | 0,78 | 0,78 | 0,66 | **0,78** |

**Source**: Python output

**Figure (2): F1 score before and after feature selection**



**Source:** Excel output

Table (5) shows the prediction accuracy and F1 Score of the three Machine Learning classifiers. We compare the models using the dataset with feature selection technique. The results confirmed that RF is the best model with the highest accuracy and F1 Score followed by SVM and KNN respectively.

**Table (5): Results of accuracy and F1 score**

| | **Accuracy Score** | | | **F1 Score** | | |
|---|---|---|---|---|---|---|
| **Datasets** | RF | SVM | KNN | RF | SVM | KNN |
| **AMZN** | **0,63** | **0,63** | 0,56 | **0,7** | 0,66 | 0,66 |
| **EBAY** | 0,63 | **0,75** | 0,56 | **0,75** | 0,7 | 0,69 |
| **KR** | **0,75** | **0,75** | 0,63 | **0,8** | 0,73 | 0,72 |
| **TGT** | **0,88** | 0,75 | 0,63 | **0,88** | 0,63 | 0,63 |
| **BABA** | **0,76** | 0,61 | 0,66 | **0,78** | 0,72 | 0,72 |
| **JD** | 0,52 | 0,57 | **0,61** | 0,54 | **0,66** | **0,66** |
| **PDD** | **0,85** | 0,76 | 0,71 | **0,91** | 0,82 | 0,82 |
| **3690.HK** | 0,7 | 0,7 | **0,75** | 0,72 | **0,78** | **0,78** |

**Source:** Python output

## Conclusion:

The current study investigates the predictive power of COVID 19 on the movement of E-Commerce stocks from two developed markets US and China. Technical indicators along with COVID 19 indicators are used as features for stock trend prediction. The results of feature selection using random forest feature importance indicate that COVID 19 indicators have a predictive power on E-Commerce stocks in both the Chinese and US markets. The selected features improve the accuracy and F1 Score of the machine learning classifiers compared to the datasets without feature selection. RF and SVM are found to be the best classifiers with an average of accuracy and F1 Score of 72%, 76% for RF and 69%, 71% for SVM respectively in the US and Chinese stocks. KNN has the lowest performance with 64% accuracy and 71% F1 Score. Ecommerce business has been affected by the COVID 19 due to worldwide government prevention measures to control the spread of the virus (lockdowns, ban of public gathering, quarantine). As a result, consumers demand has shifted to online shopping. This change in consumer’s behavior was supported by the strong and developed financial and technological sectors in the developed countries.

The current study uses only technical and COVID-19 indicators to predict the stock prices trend in developed markets. Further studies may use fundamental and sentiment analysis along with technical indicators to capture other factors impacting the price movements and conduct the analysis in different environment rather than developed economies.

## Bibliography List:

Alfonso C, V., Boar, C., Frost, J., Gambacorta, L., & Liu, J. (2021). E-commerce in the pandemic and beyond: Bank for International Settlements.

Bhatti, A., Akram, H., Basit, H., Khan, A., Mahwish, S., Naqvi, R., & Bilal, M. (2020). E-commerce trends during COVID-19 Pandemic. *International Journal of Future Generation Communication and Networking, 13*.

Breiman, L. (2001). Random Forests. *Machine Learning, 45*(1), 5-32. doi: 10.1023/a:1010933404324

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory, 13*(1), 21-27. doi: 10.1109/tit.1967.1053964

İbiş, H., Işık, S., & Gulseven, O. (2021). *The Impact of the COVID-19 Pandemic on Amazon's Business*.

JHU. (2021). COVID-19 Data Repository, from https://coronavirus.jhu.edu/

Khan, W., Ghazanfar, M. a., Assam, M., Ahmad, S., & Khan, J. (2016). PREDICTING TREND IN STOCK MARKET EXCHANGE USING MACHINE LEARNING CLASSIFIERS.

Kumar, I., Dogra, K., Utreja, C., & Yadav, P. (2018, 20-21 April 2018). *A Comparative Study of Supervised Machine Learning Algorithms for Stock Market Trend Prediction.* Paper presented at the 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT).

Laato, S., Islam, A. K. M. N., Farooq, A., & Dhir, A. (2020). Unusual purchasing behavior during the early stages of the COVID-19 pandemic: The stimulus-organism-response approach. *Journal of Retailing and Consumer Services, 57*, 102224. doi: https://doi.org/10.1016/j.jretconser.2020.102224

Mottaghi, N., & Farhangdoost, S. (2021). Stock Price Forecasting in Presence of Covid-19 Pandemic and Evaluating Performances of Machine Learning Models for Time-Series Forecasting: arXiv.org.

OECD. (2020). E-commerce in the times of COVID-19 Retrieved June 5, 2021, from https://www.oecd.org/coronavirus/policy-responses/e-commerce-in-the-time-of-covid-19-3a2b78e8

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*(1), 259-268. doi: https://doi.org/10.1016/j.eswa.2014.07.040

Trafalis, T. B. (1999). *Primal-Dual Optimization Methods In Neural Networks And Support Vector Machines Training*.

V, V. (1998). *Statistical Learning Theory*. New York: Wiley.

Xianya, J., Mo, H., & Haifeng, L. (2019). Stock Classification Prediction Based on Spark. *Procedia Computer Science, 162*, 243-250. doi: https://doi.org/10.1016/j.procs.2019.11.281