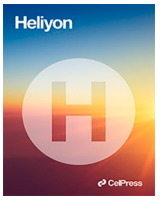
Heliyon 10 (2024) e33190

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon



Research article

COVID-19's influence on Karachi stock exchange: A comparative machine learning algorithms study for forecasting

Tahir Munir [a,*], Rabia Emhamed Al Mamlook [b,c], Abdu R. Rahman [d], Afaf Alrashidi [e], Aqsa Muhammad Yaseen [f]

[a] Department of Anaesthesiology, The Aga Khan University, Karachi, 74800, Pakistan
[b] Department of Business Administration, Trine University, Angola, IN, 49008, USA
[c] Department of Mechanical and Industrial Engineering, University of Zawia, Al Zawiya City, P.O. Box 16418, Libya
[d] Institute for Global Health and Development, The Aga Khan University, Karachi, 74800, Pakistan
[e] Department of Statistics, College of Science, University of Tabuk, Saudi Arabia
[f] Department of Sociology, University of Karachi, Pakistan

ARTICLE INFO

Keywords:
Karachi stock exchange
KSE-100 index
COVID-19
Machine learning
Performance metrics

ABSTRACT

The COVID-19 pandemic has great effects for economies internationally. This study studies the interconnection between COVID-19 metrics and Pakistan's premier stock exchange, the Karachi Stock Exchange (KSE) with the object of identifying the most effective machine learning (ML) model for predicting KSE developments in the pandemic. Our investigation periods the peak COVID-19 period from March 1, 2020, to November 26, 2021, applying data from both the KSE 100 index and COVID-19 associated variables. Five various ML methods were applied involving Linear Regression (LR), K-Nearest Neighbors (KNN), Random Forest (RF), Regression Tree (Rtree), and Support Vector Machine (SVM) and measured their performance employing critical accuracy metrics such as Mean Absolute Percentage Error (MAPE), Mean Squared Error (MSE), Mean Absolute Error (MAE), and R-squared (R^2). The outcomes discover that the RF model outperformed its equivalents realizing an R^2 of 0.91 with $k = 5$. These results conflict with a previous study that supported a negative impact of COVID-19 on improved stock markets. The visions from this study can assist investors in managing strategic investment decisions and assist policymakers in making measures to reduce the pandemic's effects on the stock market.

1. Introduction

The international increase of COVID-19 has significantly reshaped economic environments with demonstrated effects on stock exchanges worldwide. Early demonstrations pointed to severe effects on trade, tourism, and employment, indicating a pandemic-induced economic crisis [1–3]. As an important component of Pakistan's economy, the Karachi Stock Exchange (KSE) encapsulates the financial tumult, asserting a thorough investigation into the pandemic's impact. Stock markets, integrally volatile have challenged amplified during the pandemic with significant conflict mirrored in the performance of the KSE-100 Index. An overload of research has investigated its effects on financial markets, noting changes in returns, volatility, and risk dynamics [4–9].

The financial tremors caused by COVID-19 are unparalleled with significant effects on economic activities such as travel, trade, and

* Corresponding author.
E-mail addresses: tahirmunir677@yahoo.com, tmunir@stat.qau.edu.pk (T. Munir).

https://doi.org/10.1016/j.heliyon.2024.e33190
Received 16 September 2023; Received in revised form 10 May 2024; Accepted 16 June 2024
Available online 17 June 2024

production. The pandemic's attack has suggested extreme instabilities across the world's financial models with some studies suggesting its impact on stock markets is without comparison with any prior health crisis [10]. The unique extent and preservation of COVID-19's economic outcome have been decorated when compared with other natural devastations [11]. Stock market instability extended exceptional altitudes in several regions including the United States (US) where the instability eclipsed that of historic economic downturns. This period of extreme change has sparked extensive investigation into the pandemic's effect on stock markets, examining aspects like government interventions, and the influence of pandemic-related metrics [12–22]. Stock market prediction remains a challenging yet interesting field due to the difficulty and volatility of the market environment. The literature integrates a host of macroeconomic and global market indicators such as exchange rates, production indices, and commodities like gold and oil, which all bear influence on market dynamics [23–25].

The Karachi Stock Exchange 100 Index (KSE-100) is a critical economic indicator for Pakistan, encompassing the nation's top companies and completing as a comparative tool for market performance. Studies have highlighted how COVID-19 has made worse market volatility and liquidity issues, combining the pandemic to heightened systemic risk in financial sectors worldwide [26–38]. This study advances existing research by employing machine learning (ML) to assess the impact of COVID-19 on the KSE-100. Leveraging ML's proficiency in revealing hidden market trends and its adaptability to complex, nonlinear patterns offer a cutting-edge approach to understanding the pandemic's influence on stock prices [39,40]. Our research contributes a novel perspective by using these advanced techniques to navigate the intricate relationships between COVID-19 metrics and stock market behaviors.

Navigating this complex scenario demands robust analytical tools capable of deciphering intricate patterns within financial data. Machine learning (ML) presents a promising avenue, yet its application introduces challenges, such as selecting the appropriate model and configuring it to capture the nuances of pandemic-related market shifts. The aim of this research is to harness ML to model and predict the KSE-100 Index, focusing on COVID-19's influence. By assessing various ML algorithms, we aim to identify the most accurate forecasting method in this context.

This study extends the frontiers of financial market analysis by employing ML models to dissect the multifaceted effects of COVID-19 on the KSE-100 Index. Through a comparative study of ML techniques including Linear Regression, Decision Tree, Random Forest, K-Nearest Neighbors, and Support Vector Regression, we provide novel insights into their predictive efficacy, highlighted by the standout performance of the Random Forest model [32]. Our work contributes to informed decision-making for investors and policymakers in the face of pandemic-induced volatility.

The organization of this paper is presented as follows: Section 2 evaluates related work, Section 3 shows the Stock exchange and COVID-19 datasets, and the pre-processing method employed, Section 4 explains the different ML models, Section 5 presents the results and discussion, and finally, Section 6 examines future work and concludes the research.

2. Literature review

The literature on machine learning (ML) in financial market forecasting is rich and diverse, reflecting an array of methodologies tailored to address the sector's complex challenges [41]. demonstrated the efficacy of ML in predicting the Bombay Stock Exchange, while [42] evaluated Kalman filters, XGBoost, and ARIMA models, illuminating their predictive strengths and limitations [43]. expanded this landscape by predicting stock market volume with algorithms like AdaBoost and Multilayer Perceptron, thereby enriching our understanding of ML applications in finance. Similarly [44,45], explored tree-based and neural network models, offering insight into the versatility of ML techniques across varying market conditions. ML's robustness was also harnessed by [46,47] for COVID-19 time series and spread pattern analysis, with LSTM networks showing superior performance in forecasting. Echoing this [48], leveraged a deep neural network (DNN) model to forecast pandemic-induced changes in urban environments, outstripping traditional ML methods.

The pandemic's effect on financial markets attracted focused research, evidenced by [49] who utilized natural language processing to correlate COVID-19 news sentiment with market fluctuations. Similarly [50], examined the interrelation between COVID-19, gold prices, and the Karachi Stock Exchange-100 Index, while [51] applied DNN-based multivariate regression to decode the pandemic's complex impact on currency markets [52]. introduced a comprehensive MLSF framework to predict Chinese market trends, and [53] employed PCA to dissect COVID-19's impact on the Egyptian Stock Exchange.

Further studies [54,55] offered innovative AI-driven frameworks to understand financial stress patterns and the influence of COVID-19 vaccination programs on carbon markets, respectively. The SHAPley Additive exPlanations method in [55]'s study elucidated the interpretive power of ML in analyzing economic phenomena. Reflecting on stock prices, the pandemic's immediate economic consequences precipitated widespread downturns [56,57]. documented significant volatility and declines in global markets. The multifaceted response to COVID-19, as shown in studies [58,59], was shaped by government responses, outbreak severity, and vaccination efforts.

While the existing literature has made valuable contributions, there are several limitations that warrant further investigation. A major gap is the lack of widespread, high-quality datasets that accurately obtain marketplace dynamics in reaction to the developing pandemic position. Data scarcity is a problem not just in terms of quantity, but also quality as the accuracy of predictions heavily relies on the reliability and relevance of the input data. Model complexity can be a double-edged weapon in previous studies. While highly developed models with high degrees of freedom may better display the market's volatile behavior, they also risk overfitting to the training data and failing to generalize well to unseen scenarios. Additionally, complex models often lack interpretability which is crucial for gaining trust in the financial community.

Existing research tends to focus narrowly on individual markets without accounting for the global interconnectivity between

different financial systems. The contagion effects of market sentiment, especially during global crises, necessitate a broader perspective that captures these intricate inter-market relationships. The "black box" nature of certain advanced ML algorithms used in prior work can be problematic, as the financial sector's regulatory environment demands transparency and accountability. Models with opaque inner workings undermine this key requirement. By highlighting these limitations, this study aims to address the gaps and advance the field by employing innovative techniques and a holistic approach to financial forecasting amidst global health crises like COVID-19.

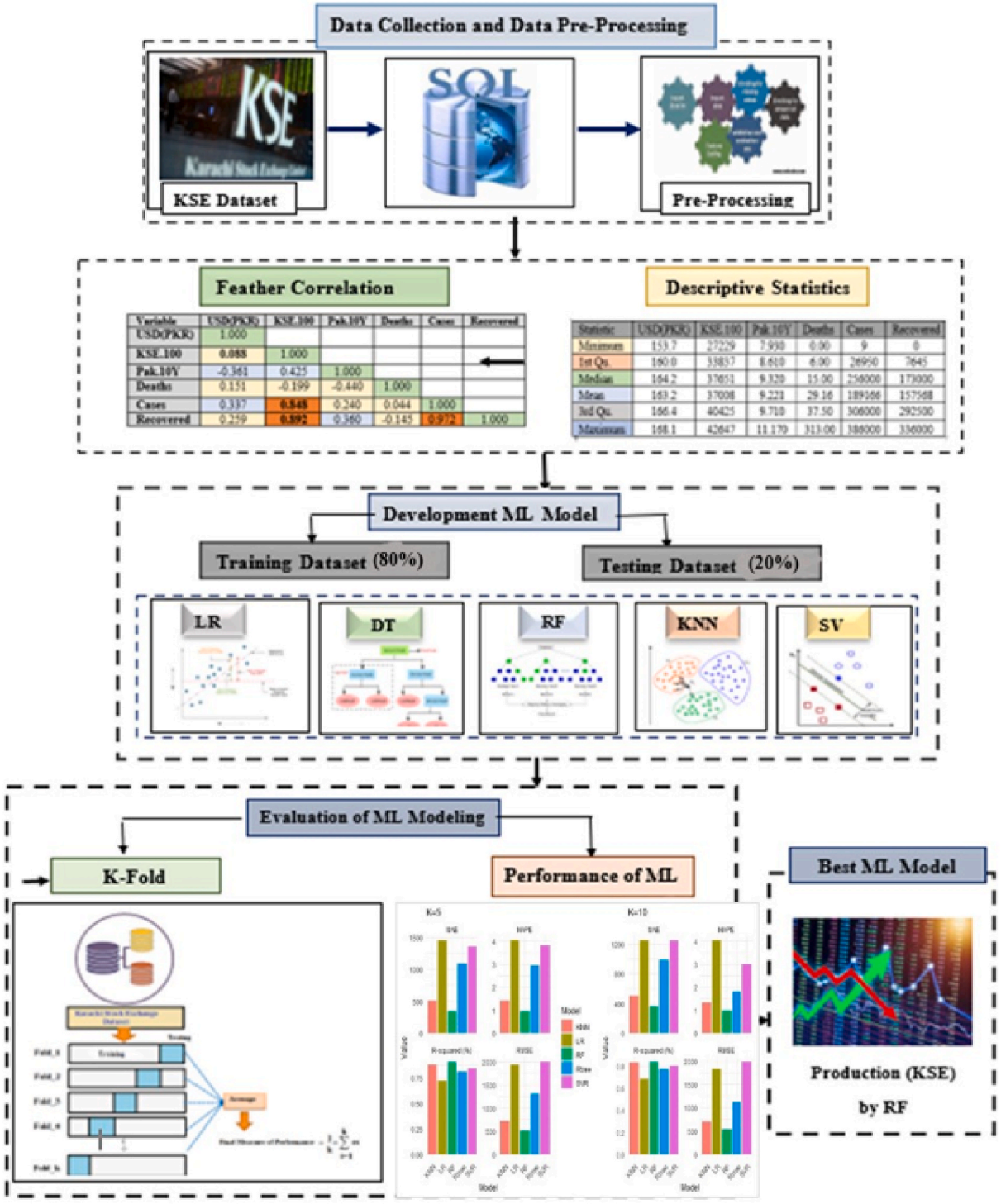


Fig. 1. The framework of the machine learning model proceeds.

Overcoming these limitations requires innovative data collection methods, model development that balances accuracy with explainability, and approaches that embrace the global interconnectedness of financial markets. Only through such holistic advancements can ML truly tackle the formidable challenge of financial forecasting amidst the complexities of global health crises like COVID-19.

3. Materials and methods

This section briefly demonstrates that data collection, data preprocessing, ML models, model performance evaluation, and validation are the four primary aspects of the suggested methodology. The framework of the ML model proceeds is shown in Fig. 1.

3.1. Data collection and feature definition

Precise prediction of the KSE-100 index necessitates meticulous recording and consideration of relevant variables. Our methodology commenced with the systematic collection of data, exploring the relationship between COVID-19 and the KSE-100 index across various scenarios. COVID-19 data was sourced from the official World Health Organization (WHO) website, while KSE-100 index data was retrieved from the Pakistan Stock Exchange (https://www.psx.com.pk/). We meticulously gathered data spanning from March 1st, 2020, to November 26th, 2021, excluding official holiday values. The variable descriptions crucial for the KSE-100 index model are concisely outlined in Table 1.

3.2. Data pre-processing

Data pre-processing is a critical precursor to developing accurate ML models. Prior to model development, meticulous data preparation was conducted, comprising multiple phases including noise reduction, outlier detection, standardization, normalization, feature selection, and encoding. A comprehensive assessment of the data was undertaken to address inconsistencies in stock codes, followed by the removal of duplicate instances and application of noise removal techniques to the COVID-19 dataset. Feature aggregations were also performed to optimize the dataset for analysis. Notably, the datasets contained numerical attributes with disparate ranges, necessitating standardization for efficient model training. Z-score normalization was employed to transform the data to have a mean of zero and standard deviation of one.

3.3. Machine learning models implementation for KSE-100 forecasting

This section outlines the methodology employed for evaluating the accuracy of combined models relative to individual models in forecasting the KSE-100 during the COVID-19 period. The method implicated training and testing methods to evaluate the efficacy of the offered ML algorithms. The training phase utilized 80% of the dataset to train the models while the remaining 20% was utilized for testing and validation. To ensure toughness and effectiveness, 5-fold and 10-fold cross-validation methods were employed.

Five prediction algorithms were selected for evaluation: Linear Regression (LR), Decision Tree (DTree), Random Forest (RF), K-Nearest Neighbors (KNN), and Support Vector Regression (SVR). These algorithms were chosen based on their recent advancements, efficiency in ML-based prediction, scalability, accuracy, speed, flexibility, and regularization capabilities to mitigate overfitting. Detailed insights into the features and implementation procedure of each model are provided in the subsequent section.

3.3.1. Quantile regression (QR)

QR is an expansion of linear regression used when the conditions of linear regression are not met. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables, they are considering, and the number of independent variables being used. A linear form for general quantile regression is described by [60]:

$$y_i = x_i\beta_\theta + \varepsilon_\theta, \tag{1}$$

for $i = 1,2,3,\ldots,n$, where β is a $(k \times 1)$ vector of coefficients in Eq. (1), $\{x_i\}$ is the column vector corresponding to the transposition of the i^{th} row of the $\{X_{n\times k}\}$ matrix of explanatory variables, $\{y_i\}$ is the i^{th} dependent variable observation, and ε_θ is the unknown error term in the presence of $\{x\}$, the i^{th} conditional quantile of $\{y\}$ can be rewritten as

Table 1
Variable descriptions for the KSE-100 index model.

No	Variable Label	Data Type	Description of variable
1.	KSE-100	Numerical	The closing price of the index.
2.	USD (PKR)	Numerical	US dollar (USD) to Pakistani rupee (PKR) exchange rate
3.	P10YBY	Numerical	Pakistan 10-year bond yield according to over the counter
4.	Deaths	Numerical	Frequency of daily COVID-19 deaths
5.	Reported	Numerical	Frequency of daily reported COVID-19
6.	Recovered	Numerical	Frequency of daily recovered COVID-19

$$\text{Qunant}_{\theta} = (y_i|x_i) = x'\beta_{\theta}. \tag{2}$$

The continuous increase in the conditional distribution of $\{y\}$ given $\{x\}$ is traced out in Eq. (2). The conditional quantile of $\{y_i\}$, conditional on $\{x_i\}$, is assumed to satisfy $\text{Qunant}_{\theta} = x'\beta_{\theta}$, for several different values of θ, $\theta\epsilon(0,1)$, resulting in $\text{Qunant}_{\theta} = (y_i|x_i) = 0$. This allows for parameter heterogeneity across various types of regressors utilizing quantile regression. As a result, in Eq. (3), the quantile regression estimator can be employed to solve the minimization problem described below:

$$\min_{\beta\epsilon\mathscr{R}^K}\left[\sum_{i\epsilon\left\{i:y_i\right\}>x_i\beta}\theta|y_i - x'\beta| + \sum_{i\epsilon\left\{i:y_i\right\}>x_i\beta}(1-\theta)|y_i - x'\beta|\right]. \tag{3}$$

The quantile function is a weighted sum of the absolute value of the residuals. Where the weights are symmetric for the median regression case in $\theta = 1/2$, the minimization problem above reduces to $\min_{\beta\epsilon\mathscr{R}^K}\sum_i^n|y_i - x'\beta|$, otherwise, it's asymmetric.

3.3.2. Support Vector Machine (SVM)

SVM is a flexible supervised ML process to analyze data amid at both classification, regression, and other purposes like outlier detection. To achieve said objectives, it is constructed a hyperplane or set of hyperplanes in a high- or infinite-dimensional space [61–63]. SVR has some key features to work; 1) Kernel function is used for mapping lower-dimensional data into higher-dimensional data. 2) Hyperplane is a line that draws the separation between two classes in general SVM. While SVR helps to predict the continuous variables and covers most of the data points. 3) Boundary lines are the two lines apart from the hyperplane, which creates a margin for data points. Finally, in 4) Support vectors are the data points that are nearest to the hyperplane and opposite class. The hyperplane gives intuitively optimal separation which has the largest distance to the nearest training data point of any class and is used for minimizing an error. In this case, it is termed as SVR, if used for regression analysis. Generally, the effort in SVR is to consider the maximum data points within the boundary lines, and the hyperplane (best-fit line) must contain a maximum number of data points (cf., Fig. 2).

3.3.3. Random Forest (RF)

RF is a well-established supervised learning technique used for classification, regression, and other tasks which are based on the ensemble learning method. It works by developing multiple decision trees to control the over-fitting process at the time of training and outputting classes (classification) as a pattern or average prediction (regression) of each decision tree. To train each decision tree, the training data is split into n bags which are used to train their respective decision trees. At the final stage predict by regressing or taking the average of all tresses (cf., Fig. 3).

3.3.4. K-Nearest Neighbors (KNN)

KNN is one of the simplest and easy-to-implement supervised machine learning processes that can be used to solve both classification and regression problems. KNN is robust to the noise as well as more effective if the training data is large. It assumes the similarity between the new case/data and available cases and put the new case into the category that is most like the available categories. This technique stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a good suite category by using the KNN algorithm. In the present study, the actual Manhattan distance was enhanced using weighting (cf., Fig. 4).

3.3.5. Regression Decision Tree (RTree)

RTree is considered a predictive modeling technique in machine learning. It uses a decision tree (as a predictive model) to go from observations about an item (represented in the branches) to conclusions about the item's target value (represented in the leaves). Decision trees where the target variable can take continuous values are called RT. They are among the most popular machine learning algorithms for their intelligence and simplicity (cf., Fig. 5).

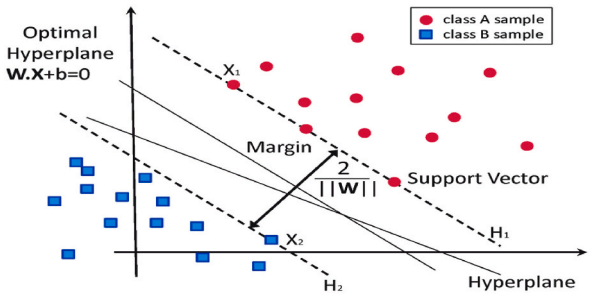


Fig. 2. Support vector machine.

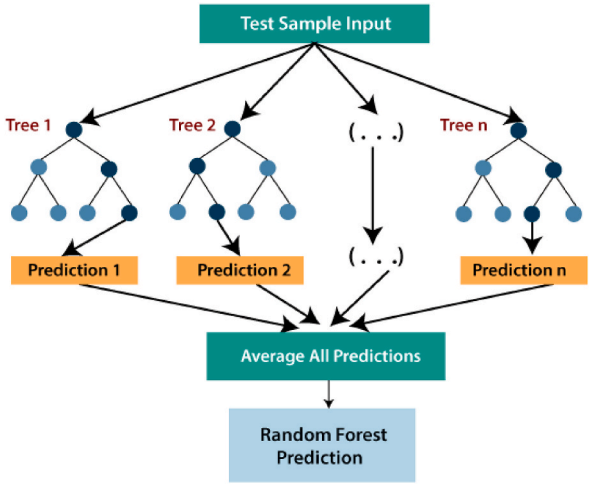


Fig. 3. Flowchart of the RF's architecture.

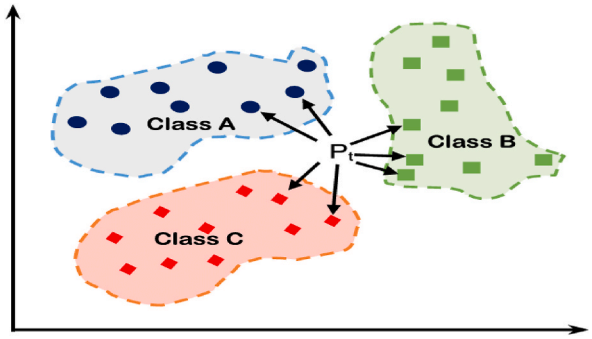


Fig. 4. K-nearest neighbors.

3.4. Model building and validation (K-fold cross-validation)

Datasets were divided into two portions (i.e., training and testing sets). This phase is critical to regulating the efficiency of the utilized machine learning procedures. The adopted algorithms are trained using the training portion of the datasets, and then the remaining portion is used for testing purposes, which is vital to demonstrate the developed model's response toward new data being processed for the first time. In the current research, the suggested prediction models were tested for robustness and effectiveness using 5-fold as well as 10-fold cross-validation. To follow [64,65], the training portion (i.e., 80 % of the dataset) is used to train the suggested

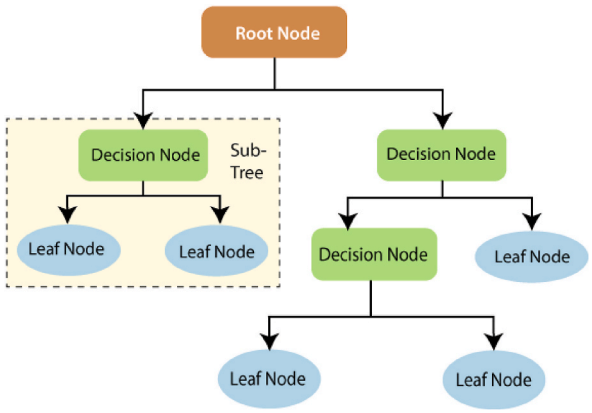


Fig. 5. Regression decision tree.

model, and the testing portion (i.e., 20 % of the dataset) is used to conduct the tests. The performance of our proposed model through K-fold cross-validation and hold-out validation was evaluated in Fig. 6 with the metrics of accuracy. The K-fold cross-validation-based model is important for fine-tuning the parameters of a prediction model. In this study, the proposed system uses 5-fold and 10-fold cross-validation. A cross-validation model allows you to find the best-performing parameters set for any ML technique without overfitting.

In Fig. 6, the training data is divided into five equal subsets in which one subset represents the validation set and the other four subsets represent the training. Five-fold cross-validation, iteratively repeats the process five times, each time using a different subset as the validation set and the rest as training data.

After the iteration is completed, the hit rate of all five iterations is considered and the average of the results is taken. In this study, the results are the rate and the trading strategy outcome. If the average maximum result is achieved in that test, the input parameters for the machine learning technique are captured and stored for the best-performing parameters set. Various cross-validation tests are performed to achieve the best-performing parameters using different variables. All the best-performing parameters of each machine learning technique for the average rate and profit are used once on the testing data. A few advantages and disadvantages are reported of the considered ML models in Table 2.

3.5. Performance measurements

In this subsection, the prediction performance of this study is measured through multiple metrics to assess the quality of the learning methods. Four measures were used that are common measures to evaluate models as accuracy metrics to assess the model's performance. The modeling process outputs' accuracy was then measured and compared using three performance indecencies (i.e., the Mean Absolute Error (MAE), the Mean Squared Error (MSE), and the mean absolute percentage error (MAPE). The mathematical representation of the four implemented measures is given in Table 3.

4. Results and discussion

4.1. Descriptive statistics of the dataset

In the provided dataset, we analysed five predictor variables and one dependent variable. The predictors include USD to PKR exchange rate, 10-year bond yield (P10YBY), and COVID-19-related data deaths, reported cases, and recoveries. The dependent variable is the KSE-100 index. Descriptive statistics for these variables, including range, mean, median, and quantiles, were compiled and are detailed in Table 4. From the analysis, we observe that during the pandemic, the average value of the KSE-100 index stabilized at approximately 37,008 points. Initially, in 2020, the daily average for reported COVID-19 cases was 189,166; the average number of deaths stood at 15, and the average recoveries were reported at 157,568. These statistics provide a foundational understanding of the dataset's characteristics and the pandemic's impact on the financial and heath sectors captured within our data.

4.2. Meeting assumptions

The normality test is applied to ascertain whether the dataset under consideration adheres to a normal distribution. This testing is crucial for the correct application of numerous statistical analysis techniques and involves evaluating the data's skewness and excess kurtosis. The Jarque-Bera test, which also hinges on skewness and kurtosis to determine normality, was employed. The results, as displayed in Table 4 and illustrated in Fig. 7, indicate that the KSE-100 index distribution deviates from normality, evidenced by a p-value less than 0.05, suggesting significant skewness and kurtosis.

Additionally, a Normal Q-Q test was conducted to further assess the distribution's normality. This test utilizes a Q-Q Plot, which juxtaposes the quantiles of the dataset against those expected under a normal distribution, providing a visual assessment of normality. Ideally, if the data points form a straight line on the plot, it indicates that the residuals are normally distributed. However, as observed in Fig. 8, the plot exhibits deviations from a straight line, confirming the non-normal distribution of the dependent variable. These

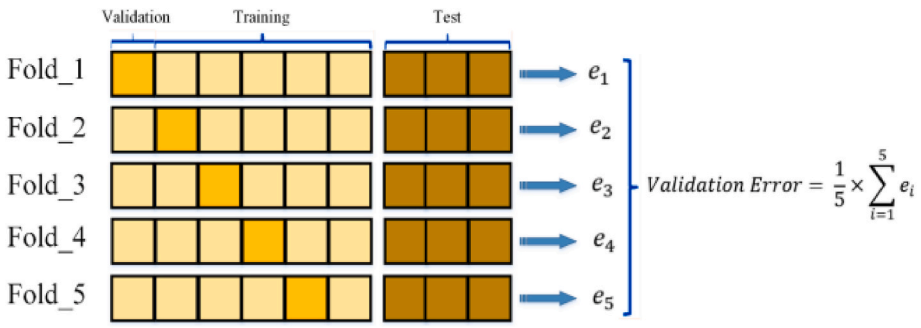


Fig. 6. Schemes of k-fold Cross-validation.

Table 2
Advantages and disadvantages of the considered ML models.

Method	Advantages	Disadvantages
LR	Simplicity and interpretability	Assumes linear relationships, may not capture complex patterns
Rtree	Can model complex relationships	Prone to overfitting, sensitive to small variations
RF	Reduces overfitting by averaging multiple trees	Complexity, harder to interpret than individual trees
KNN	Simple to implement, works well with small datasets	Computationally expensive for large datasets, sensitive to outliers
SVR	Effective in high-dimensional spaces, robust to outliers	Requires tuning of hyperparameters, less effective with noisy data

Table 3
Mathematical illustration for performance metrics.

Measure Models	Formula	Variables Description	
Root Mean Squared Error RMSE	$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	n	Sample size
		y_i	is the actual value represented
		$\hat{y}_i$	the predicted value represented by the ith case.
Mean Absolute Error—MAE	$MAE = \frac{1}{n}\sum_{i=1}^{n}\lvert e_i\rvert$	n	Sample size
		e_i	$= (y_i - \overline{y}_i)$
R-square		y_i	is the actual value represented
	$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$	n	Sample size
		$\hat{y}$	the predicted value
Mean Absolute Percentage Error—MAPE	$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left\lvert\frac{y_i - \hat{y}_i}{y_i}\right\rvert \times 100$	n	Sample size
		y_i	The actual value is represented by ith case.
		$\hat{y}_i$	The predicted value is represented by the ith case.

Table 4
Descriptive statistical information.

Features	USD(PKR)	KSE-100	P10YBY	Deaths	Reported	Recovered
Min.	153.7	27229	7.93	0	9	0
Max.	168.1	42647	11.17	313	386000	336000
1st Qu.	160	33837	8.61	6	26950	7645
3rd Qu.	166.4	40425	9.71	37.5	306000	292500
Mean ± StDev	163.15 ± 0.22	37008 ± 237	9.22 ± 0.04	29.13 ± 2.19	189166 ± 827	157568 ± 815
Skew	0.6	−0.8	0.7	0.3	0.24	0.5
Kurtosis	−1.34	−0.42	−1.22	−1.27	−1.21	−0.92
Jarque-Bera	51.03	49.06	37.36	34.04	30.34	32.68
p-value	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001

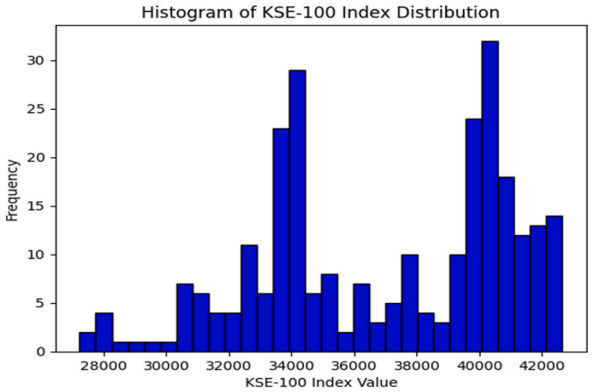


Fig. 7. Histogram of KSE-100 index distribution.

findings underscore the challenges in modeling the KSE-100 index using methods that assume normal distribution and highlight the need for alternative statistical techniques that can accommodate non-normal data distributions from that data would not have a normal distribution. As likely, the Q-Q-Plot did not show that the residuals from the model were normally distributed. To obtain a normal distribution, the data would need to be transformed Therefore, to determine whether our dataset is suitable for modeling with

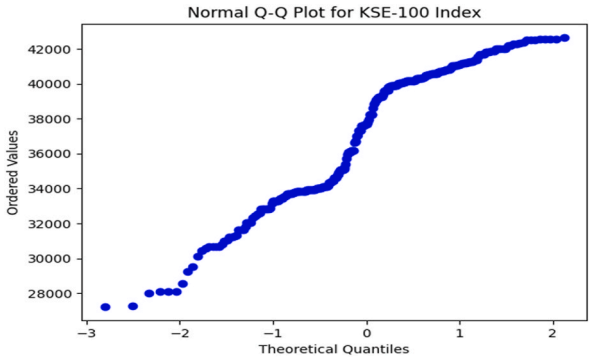


Fig. 8. Normal Q-Q plot.

ML, we applied the following criteria to the to be transformed. The fourth step of data preprocessing was transforming the data to be more suitable for applying ML approaches. To ensure that all the feature values are on the same scale and treated with equal weight, normalization is important. A Box-Cox normalization transformation was performed to transform each numerical attribute (e.g. KSE-100) to improve the ML model's performance.

4.3. Statistical analysis

In this section, results for the stock forecasting are presented. Likewise, the contour plot shows the relation between recovered versus P10YBY, KSE-100. The recovered cases increased as KSE-100 increased, as demonstrated in Fig. 9. Darker regions indicate higher recovered cases. These higher response values seem to form a ridge running from the upper middle to the lower of the graph. Furthermore, this shows that the maximum expected KSE-100 occurs at recovered more than 300000. The lowest values reported cases are in the lower left corner of the plot, which corresponds with the low values of both P10YBY and KSE-100.

Fig. 10 illustrates the relationship between the P10YBY and KSE-100 by deaths. Small darker regions indicate fewer deaths, which are less than zero. These higher response values seem to form a ridge running from the upper middle to the lower right of the graph. The valleys in the lower left of the graph represent P10YBY and KSE-100 that result in 100–150 death.

The contours are curved because the model contains quadratic terms that are statistically significant. The lowest values of reported cases are in the lower down corner of the plot, which corresponds with low values of both P10YBY and KSE-100. Furthermore, Fig. 11 shows that the maximum expected KSE-100 index occurs at reported cases of more than 300.

4.4. Correlation matrix analysis

Correlation matrix among and in between selected attributes to designate the pairs with high negative or positive correlation and

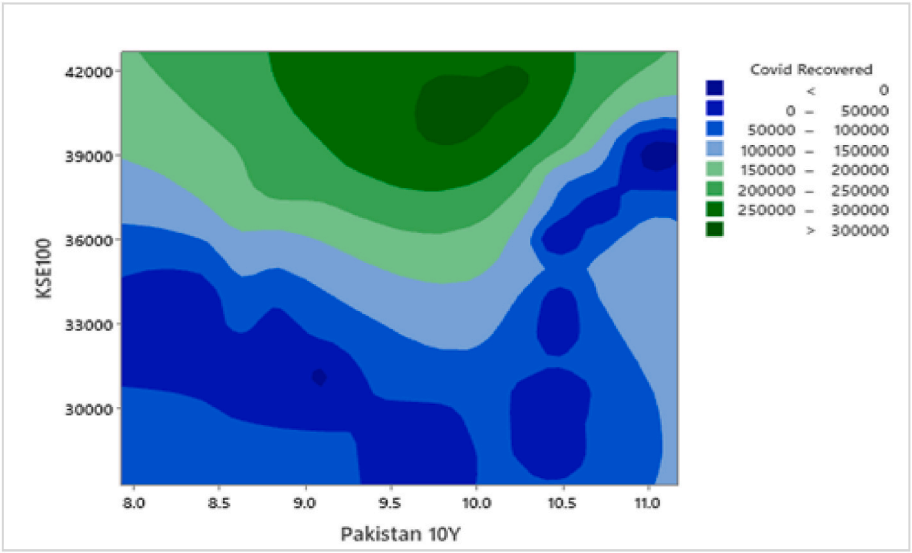


Fig. 9. Contour plot that shows recovered cases between P10YBY vs KSE-100.

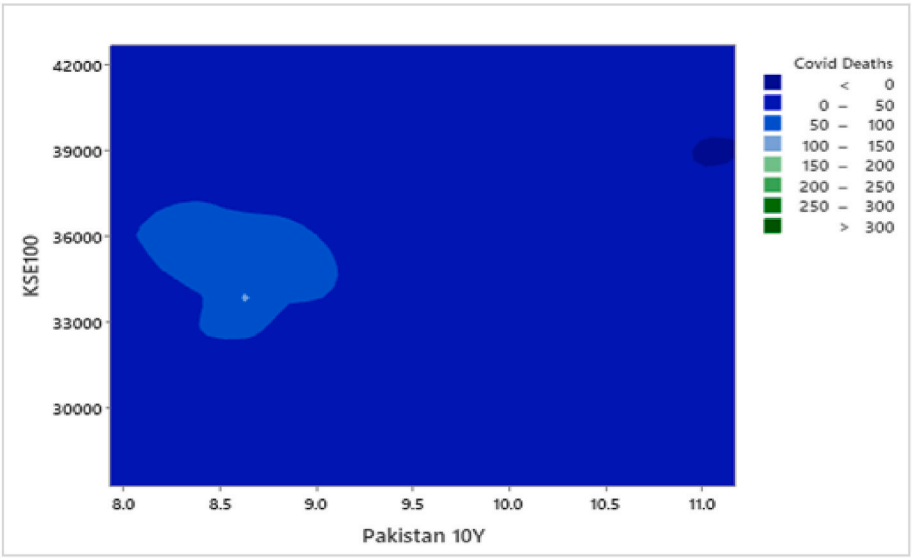


Fig. 10. Contour plot of deaths between P10YBY and KSE-100.

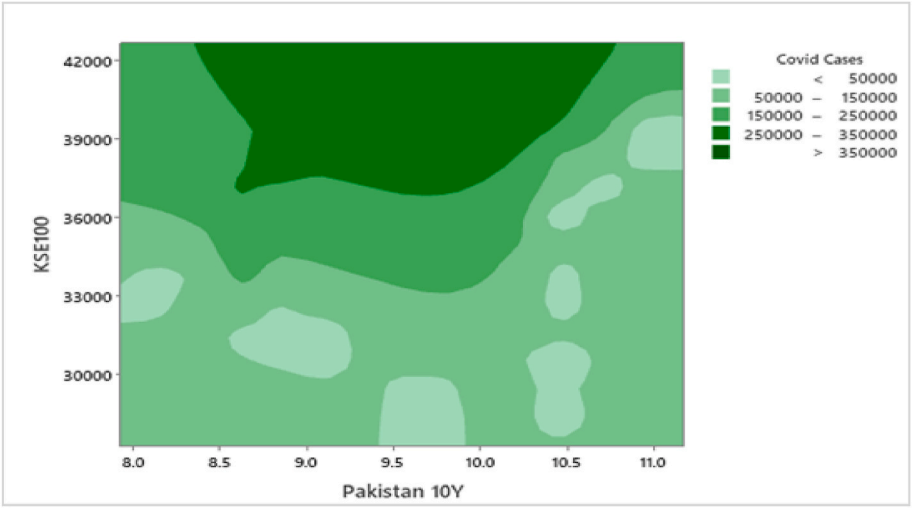


Fig. 11. Contour plot of reported cases between P10YBY vs KSE-100.

the KSE-100 was applied to evaluate the impact of these features, as shown in Fig. 12. The correlation analysis among the variables USD (PKR), KSE-100, P10YBY, deaths, reported and recovered cases to gauge the strength of their association. The correlation matrix indicates a minor level of a positive relationship between KSE-100 and USD (PKR) e.g., 0.1. There were three pairs of features with correlations of more than 0.5. One pair was recovered and reported cases (correlation coefficient = 0.8), and the other pair was KSE-100 and reported cases (correlation coefficient = 0.7).

In addition to KSE-100 and recovered (correlation coefficient = 0.8). Highly correlated features added useless information and noise to the models. Furthermore, this study shows that the predictive ability of the models lowered when the feature of recovered was involved. Therefore, taking the feature importance ranking and model performance into consideration, five independent features, including reported, recovered, P10YBY, deaths, and USD(PKR), were used as input features in the RF model.

4.5. Comparison performance of ML algorithms

The proposed system compares five machine learning techniques, namely (i.e., LR, DT, RF, KNN, and SVM). All six techniques are tested with the same data using different cross-validation based models. Each ML procedure includes three tests having different history stock prices as training data. Each test is trained with a K-fold cross-validation based model. Different models are created by selecting different values of K-fold ($K = 5$, and $K = 10$) and another model is also created in which training data (80 %) are randomly

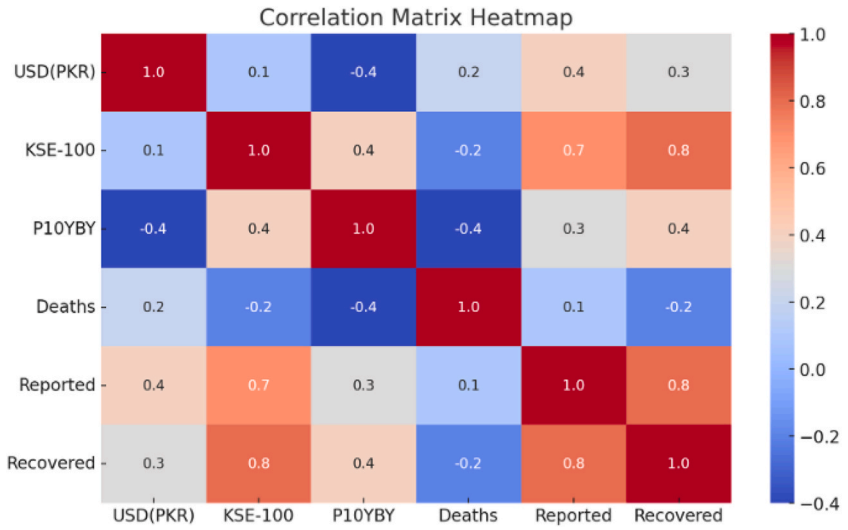


Fig. 12. Correlation matrix.

selected while hypermeters are reported in Table 5. A detailed description in Table 6 of the best-performing set generated during the cross-validation of training data. The efficiency of proposed ML algorithms has been evaluated using (i.e., MAE, MSE, RMSE, MAPE, and R^2) indicators to predict the KSE-100 prediction, refer to Fig. 13.

The comparison outcomes present RF and KNN have higher measures of R^2 value and lower MAE, RMSE, and MAPE values than the traditional models (i.e., LR, DT, RF, KNN, and SVM)) in the KSE-100 predictions. As for other metrics, the RF model showed a higher R^2 and a lower MAE, MSE, and RMSE. Moreover, the predicted results of the proposed demonstrate that the RF estimate values are very high R^2 measures with less MSE. Performance of the proposed models at K = 5, and 10, the RF model exhibited better overall performance with an R^2 of 0.91 and 0.84, and the KNN model achieved acceptable performance in predicting the KSE-100 with an R^2 of 0.88 and 0.83, respectively. Both can achieve higher predictive accuracy compared with other ML models. As a result, RF bested the other models at 5-fold and 10-fold and was the most efficient in predicting the KSE-100.

4.6. *Features importance analysis*

An improved understanding of the model's features assists investors effectively in evaluators' trends. Therefore, feature importance assessment has been performed using RF models to determine the importance mark of each variable involved in predicting the KSE-100. In the present study, RF models contained reported, recovered, and P10YBY, in the top-ranking features (Fig. 14).

Thus, the feature score plot has been performed to provide a relative score for each variable as shown in Fig. 14. Features' significance was in down order as follows: reported, recovered, P10YBY, deaths, and USD (PKR). It presents the top three most important rankings of the selected features used in RF predictive models, respectively. Reported cases were the first important feather in predicting KSE-100. In addition, recovered cases and P10YBY were the second and third most important features in predicting the RF model respectably, but another feather was the less important feather.

4.7. *Limitations, methodological scope, and practical implications*

The study may not consider external factors influencing stock market behavior during pandemics, such as socio-political dynamics, government interventions, and global economic trends, which could potentially affect outcomes beyond the scope of the analysis. Addressing these factors would bolster the robustness and applicability of the findings, yielding more comprehensive insights into stock market dynamics during pandemics. Through the integration of COVID-19 variables into ML models, these methodologies offer valuable insights into forecasting stock market behavior amidst global crises. These findings not only assist investors in making well-

Table 5
Hyperparameters values of the various ML models.

Algorithm	Hyperparameters	Values
LR	C	1.0
RTree	Max_epth	5
RF	n_estimators	100
KNN	k	5
SVR	cost (C)	1

Table 6
Regression model's validation and performance evaluation.

K-fold Cross-Validation	Regression Model	Performance Evaluation Metrics			
		RMSE	MAE	MAPE	R-squared (%)
$K = 5$	LR	1939.463	1446.226	4.023991	0.72
	Rtree	1314.119	1081.647	2.931309	0.81
	RF	**512.4574**	**347.9247**	**0.959033**	**0.91**
	KNN	720.9236	508.0206	1.415234	0.88
	SVR	2011.759	1351.021	3.808549	0.84
$K = 10$	LR	1829.4429	1245.723	4.023991	0.68
	Rtree	1113.4653	981.6789	1.834545	0.77
	RF	534.2294	362.6565	1.005125	0.84
	KNN	699.2345	508.9987	1.35098	0.83
	SVR	1998.1127	1245.0209	2.98763	0.80

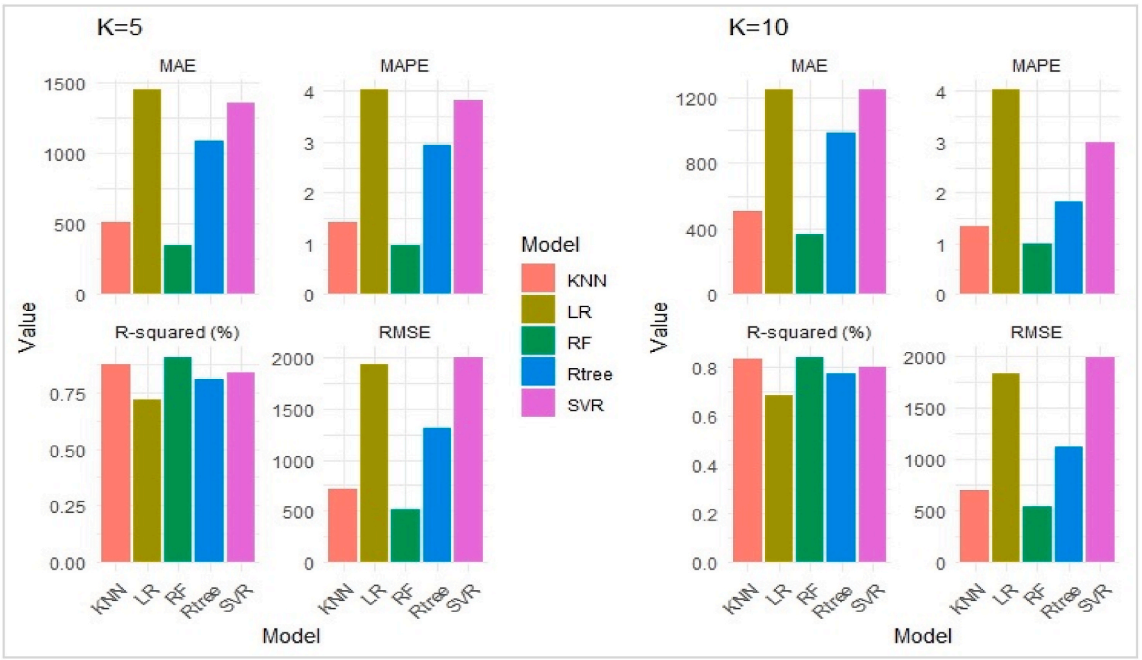


Fig. 13. Plot compared the predicted performance of models.

informed decisions but also equip policymakers worldwide with actionable strategies to mitigate the financial repercussions of pandemics.

5. Conclusion

The study effectively applied five machine learning models—Linear Regression, Decision Tree, Random Forest, K-Nearest Neighbors, and Support Vector Machines—to predict the KSE-100 index amid the COVID-19 pandemic. Utilizing detailed performance evaluations through metrics like MAE, MSE, APE, and R2, along with 5-fold and 10-fold cross-validation, the Random Forest model was found to be the most accurate. This insight is crucial for investors, offering them valuable predictions of stock market trends during health crises. This paper suggests future research should include advanced machine learning techniques like deep neural networks and Bayesian methods to improve prediction accuracy. Furthermore, analyzing the importance of each feature in the models highlights the critical variables influencing market behavior during crises, underscoring the importance of strategic feature selection to enhance the understanding of market dynamics in challenging conditions.

Funding statement

There is no specific funding for this study.

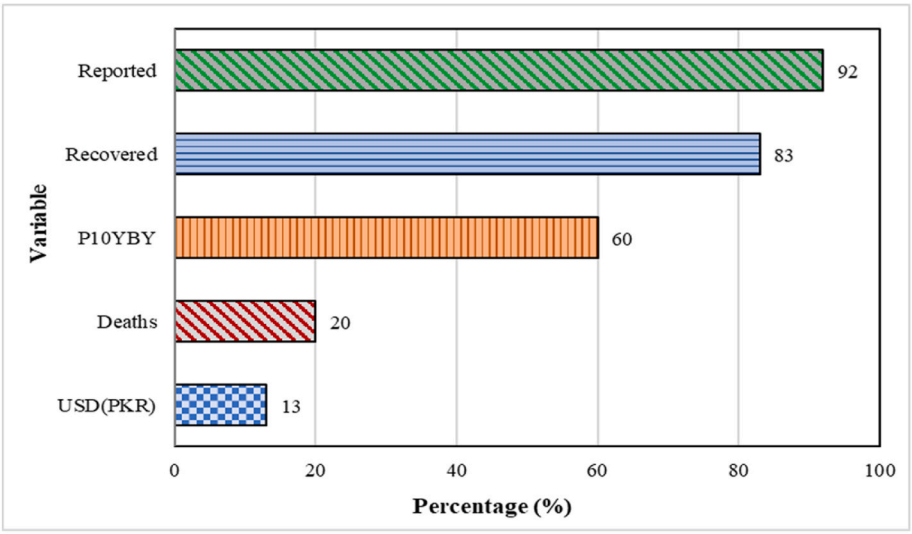


Fig. 14. Features importance analysis.

Data availability statement

Data will be made available on request.

CRediT authorship contribution statement

Tahir Munir: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Methodology, Investigation, Conceptualization. **Rabia Emhamed Al Mamlook:** Writing – original draft, Visualization, Methodology, Funding acquisition, Data curation, Conceptualization. **Abdu R. Rahman:** Writing – original draft, Visualization, Investigation, Data curation. **Afaf Alrashidi:** Writing – review & editing, Writing – original draft, Visualization, Validation, Investigation, Data curation. **Aqsa Muhammad Yaseen:** Writing – original draft, Investigation, Formal analysis, Data curation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

I would like to thank all the members of the research team for their valuable efforts.

References

[1] T. Walmsley, A. Rose, D. Wei, The impacts of the coronavirus on the economy of the United States, Econ. Disasters Climate Change 5 (1) (2021) 1–52.
[2] E.A. Mack, S. Agrawal, S. Wang, The impacts of the COVID-19 pandemic on transportation employment: a comparative analysis, Transp. Res. Interdiscip. Perspect. 12 (2021) 100470.
[3] S. Maital, E. Barzani, The global economic impact of COVID-19: a summary of research, Samuel Neaman Inst. National Policy Res. (2020) 1–12.
[4] A. Fernandez-Perez, A. Gilbert, I. Indriawan, N.H. Nguyen, COVID-19 pandemic and stock market response: a culture effect, J. Behav. Exp. Finance 29 (2021) 100454.
[5] M.N. Alam, M.S. Alam, K. Chavali, Stock market response during COVID-19 lockdown period in India: an event study, J. Asian Finance Econ. Business 7 (7) (2020) 131–137.
[6] C. Alexakis, K. Eleftheriou, P. Patsoulis, COVID-19 containment measures and stock market returns: an international spatial econometrics investigation, J. Behav. Exp. Finance 29 (2021) 100428.
[7] T.A. Kusumahadi, F.C. Permana, Impact of COVID-19 on global stock market volatility, J. Econ. Integrat. 36 (1) (2021) 20–45.
[8] R. Chaudhary, P. Bakhshi, H. Gupta, Volatility in international stock markets: an empirical study during COVID-19, J. Risk Financ. Manag. 13 (9) (2020) 208.
[9] Y. Liu, Y. Wei, Q. Wang, Y. Liu, International stock market risk contagion during the COVID-19 pandemic, Finance Res. Lett. 45 (2022) 102145.
[10] S. Baek, S.K. Mohanty, M. Glambosky, COVID-19 and stock market volatility: an industry level analysis, Finance Res. Lett. 37 (2020) 101748.
[11] J. Goodell, COVID-19 and finance: agendas for future research, Finance Res. Lett. 35 (2020) 101512.
[12] B.N. Ashraf, Economic impact of government interventions during the COVID-19 pandemic: international evidence from financial markets, J. Behav. Exp. Finance 27 (2020) 100371.
[13] A. Zaremba, R. Kizys, D.Y. Aharon, E. Demir, Infected markets: novel coronavirus, government interventions, and stock return volatility around the globe, Finance Res. Lett. 35 (2020) 101597.

[14] O. Haroon, S.A.R. Rizvi, Flatten the curve and stock market liquidity – an inquiry into emerging economies, Emerg. Mark. Finance Trade 56 (10) (2020) 2151–2161.
[15] M. Akhtaruzzaman, S. Boubaker, A. Sensoy, Financial contagion during COVID–19 crisis, Finance Res. Lett. 101604 (2020).
[16] D.I. Okorie, B. Lin, Stock markets and the COVID-19 fractal contagion effects, Finance Res. Lett. 101640 (2020).
[17] A.M. Al-Awadhi, K. Alsaifi, A. Al-Awadhi, S. Alhammadi, Death and contagious infectious diseases: impact of the COVID-19 virus on stock market returns, J. Behav. Exp. Finance 27 (2020) 100326.
[18] B.N. Ashraf, Stock markets' reaction to COVID-19: cases or fatalities? Res. Int. Bus. Finance 54 (2020) 101249.
[19] M.A. Harjoto, F. Rossi, R. Lee, B.S. Sergi, How do equity markets react to COVID-19? Evidence from emerging and developed countries, J. Econ. Business 105966 (2020).
[20] C.-O. Cepoi, Asymmetric dependence between stock market returns and news during COVID-19 financial turmoil, Finance Res. Lett. 36 (2020) 101658.
[21] M. Ambros, M. Frenkel, T.L.D. Huynh, M. Kilinc, COVID-19 pandemic news and stock market reaction during the onset of the crisis: evidence from high-frequency data, Appl. Econ. Lett. (2020) 1–4.
[22] S.R. Buckman, A.H. Shapiro, M. Sudhof, D.J. Wilson, News sentiment in the time of COVID-19, FRBSF Economic Lett 19 (2020).
[23] J.L. Ticknor, A Bayesian regularized artificial neural network for stock market forecasting, Expert Syst. Appl. 40 (14) (2013) 5501–5506.
[24] A. Srivastava, Relevance of macro-economic factors for the Indian stock market, Decision 37 (3) (2010).
[25] S. Akter, M.S. Rana, T.H. Anik, The dynamic relationship between stock market returns and macroeconomic variables: an empirical study from Bangladesh, J. Management, Econ., Industrial Organization (2020) 40–62.
[26] H.Y. Mohammed, A.A. Abu Rumman, The impact of macroeconomic indicators on Qatar stock exchange: a comparative study between Qatar exchange index and Al Rayyan Islamic index, J. Transnat. Manag. 23 (4) (2018) 154–177.
[27] A. Jain, P. Biswal, Dynamic linkages among oil price, gold price, exchange rate, and stock market in India, Resour. Pol. 49 (2016) 179–185.
[28] E. Bouri, A. Jain, P. Biswal, D. Roubaud, Cointegration and nonlinear causality amongst gold, oil, and the Indian stock market: evidence from implied volatility indices, Resour. Pol. 52 (2017) 201–206.
[29] S. Basher, A.A. Haug, P. Sadorsky, Oil prices, exchange rates and emerging stock markets, SSRN Electron. J. (2011), https://doi.org/10.2139/ssrn.1852828.
[30] R. Shahani, A. Bansal, Gold vs. India VIX: a comparative assessment of their capacity to act as a hedge and/or safe haven against stocks, crude and rupee-dollar rate, SSRN Electron. J. (2020), https://doi.org/10.2139/ssrn.3597889.
[31] W. Mensi, S. Hammoudeh, S. Yoon, M. Balcilar, Impact of macroeconomic factors and country risk ratings on GCC stock markets: evidence from a dynamic panel threshold model with regime switching, Appl. Econ. 49 (13) (2016) 1255–1272.
[32] A.S. Baig, H.A. Butt, O. Haroon, S.A.R. Rizvi, Deaths, panic, lockdowns and US equity markets: the case of COVID-19 pandemic, Finance Res. Lett. 38 (2020) 101701.
[33] M.S. Rizwan, G. Ahmad, D. Ashraf, Systemic risk: the impact of COVID-19, Finance Res. Lett. 36 (2020) 101682.
[34] D. Chen, H. Hu, C.P. Chang, The COVID-19 shocks the stock markets of oil exploration and production enterprises, Energy Strategy Rev. 38 (2021) 100696.
[35] C.D. Utomo, D. Hanggraeni, The impact of COVID-19 pandemic on stock market performance in Indonesia, J. Asian Finance, Econ. Business 8 (5) (2021) 777–784.
[36] I. Abdelkafi, Y. Ben Romdhane, S. Loukil, F. Zaarour, Covid-19 impact on Latin and Asian stock markets, Manag. Finance 49 (1) (2023) 29–45.
[37] H. Sakawa, N. Watanabel, The impact of the COVID-19 outbreak on Japanese shipping industry: an event study approach, Transport Pol. 130 (2023) 130–140.
[38] H.J. Bartolome, R.P. Bautista, M.A. Sansalian, R. Cabauatan, The effects of the COVID-19 pandemic on the philippine stock exchange index, UJoST-Universal J. Sci. Technol. 2 (1) (2023) 127–165.
[39] I. Bhattacharjee, P. Bhattacharja, Stock price prediction: a comparative study between traditional statistical approach and machine learning approach, in: Proc. 2019 4th Int. Conf. Electr. Inf. Commun. Technol. (EICT), 2019, pp. 1–6, https://doi.org/10.1109/EICT48899.2019.9068850.
[40] J. Stanković, I. Marković, M. Stojanović, Investment strategy optimization using technical analysis and predictive modeling in emerging markets, Procedia Econ. Finance 19 (2015) 51–62, https://doi.org/10.1016/S2212-5671(15)00007-6.
[41] M. Usmani, S.H. Adil, K. Raza, S.S.A. Ali, Stock market prediction using machine learning techniques, in: Proc. 2016 3rd Int. Conf. Comput. Inf. Sci. (ICCOINS), 2016, pp. 322–327, https://doi.org/10.1109/ICCOINS.2016.7783260.
[42] V.V. Prasad, et al., Prediction of stock prices using statistical and machine learning models: a comparative analysis, Comput. J. 65 (5) (2022) 1338–1351.
[43] M.A. Ghazanfar, et al., Using machine learning classifiers to predict stock exchange index, Int. J. Mach. Learn. Comput. 7 (2) (2017) 24–29.
[44] E.K. Ampomah, Z. Qin, G. Nyame, Evaluation of tree-based ensemble machine learning models in predicting stock price direction of movement, Information 11 (6) (2020) 332.
[45] U. Demirel, H. Çam, R. Ünal, Predicting stock prices using machine learning methods and deep learning algorithms: the sample of the Istanbul Stock Exchange, Gazi Univ. J. Sci. 34 (1) (2022) 63–82.
[46] M. Azarafza, M. Azarafza, J. Tanha, COVID-19 infection forecasting based on deep learning in Iran, MedRxiv (2020) 2020-05.
[47] Y.A. Nanehkaran, et al., The predictive model for COVID-19 pandemic plastic pollution by using deep learning method, Sci. Rep. 13 (1) (2023) 4126.
[48] M. Azarafza, M. Azarafza, H. Akgün, Clustering method for spread pattern analysis of corona-virus (COVID-19) infection in Iran, medRxiv (2020) 2020-05.
[49] M. Costola, et al., Machine learning sentiment analysis, COVID-19 news and stock market reactions, Res. Int. Bus. Finance 64 (2023) 101881.
[50] Y. Shah, et al., COVID-19 and commodity effects monitoring using financial & machine learning models, Sci. African 21 (2023) e01856.
[51] H.M. Naveed, et al., Artificial neural network (ANN)-based estimation of the influence of COVID-19 pandemic on dynamic and emerging financial markets, Technol. Forecast. Soc. Change 190 (2023) 122470.
[52] C. Yuan, et al., COVID19-MLSF: a multi-task learning-based stock market forecasting framework during the COVID-19 pandemic, Expert Syst. Appl. 217 (2023) 119549.
[53] H.M. Ezzat, The effect of COVID-19 on the Egyptian exchange using principal component analysis, J. Humanities Appl. Soc. Sci. 5 (5) (2023) 402–416.
[54] I. Ghosh et al., "Modelling financial stress during the COVID-19 pandemic: prediction and deeper insights," Int. Rev. Econ. Finance, doi: 10.1016/j. iref.2024.102953.
[55] C. Yang, et al., Effects of COVID-19 vaccination programs on EU carbon price forecasts: evidence from explainable machine learning, Int. Rev. Financ. Anal. 91 (2024) 102953.
[56] B. Rakshit, Y. Neog, Effects of the COVID-19 pandemic on stock market returns and volatilities: evidence from selected emerging economies, Stud. Econ. Finance 39 (4) (2022) 549–571.
[57] Z. Li, et al., A comparative analysis of COVID19 and global financial crises: evidence from US economy, Econ. Res. Ekonomska Istraživanja 35 (1) (2022) 2427–2441.
[58] C.P. Chang, et al., Government fighting pandemic, stock market return, and COVID-19 virus outbreak, Emerg. Mark. Finance Trade 57 (8) (2021) 2389–2406.
[59] I. Fasanya, O. Periola, A. Adetokunbo, On the effects of Covid-19 pandemic on stock prices: an imminent global threat, Qual. Quantity 57 (3) (2023) 2231–2248.
[60] M. Buchinsky, Recent advances in quantile regression models: a practical guideline for empirical research, J. Hum. Resour. (1998) 88–126.
[61] S. Hamida, et al., Optimization of machine learning algorithms hyper-parameters for improving the prediction of patients infected with COVID-19, in: Proc. 2020 IEEE 2nd Int. Conf. Electron., Control, Optim, Comput. Sci. (ICECOCS), 2020, pp. 1–6.
[62] N.A. Mahoto, et al., Machine learning based data modeling for medical diagnosis, Biomed. Signal Process Control 81 (2023) 104481.
[63] R.E. Al Mamlook, et al., Utilizing machine learning models to predict the car crash injury severity among elderly drivers, in: Proc. 2020 IEEE Int. Conf. Electro Inf. Technol. (EIT), 2020, pp. 105–111.
[64] R.E. Al Mamlook, et al., Machine learning to predict freeway traffic accidents-based driving simulation, in: Proc. 2019 IEEE Natl. Aerosp. Electron. Conf. (NAECON), 2019, pp. 630–634.
[65] R.E. AlMamlook, et al., Comparison of machine learning algorithms for predicting traffic accident severity, in: Proc. 2019 IEEE Jordan Int. Joint Conf. Electr. Eng. Inf. Technol. (JEEIT), 2019, pp. 272–276.