Forecasting the Pakistan’s Leading Stock Exchange During Covid-19 Using Machine learning (ML) Algorithms: Model Development and Validation

Tahir Munir (✉ tahirmunir677@yahoo.com )
The Aga Khan University

Rabia Emhamed Al Mamlook
Al Zawiya University (Seventh of April University

Abdu R Rahman
The Aga Khan University

Sujeet Shrestha
Wayne State University

Mohamed Bzizi
Western Michigan University

Abeer Aljohani
Taibah University

Research Article

Keywords: Karachi stock exchange (KSE), KSE-100 index Covid-19, Machine learning, Performance Metrics

Posted Date: January 24th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2498599/v1

License: ☺️ ⏳ This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License
Abstract

During COVID-19, marketing shows sharp fluctuation in upward and downward trends. Forecasting price actions is one of the most challenging problems in this situation. It is challenging to build an accurate model, which integrates economic and Covid-19 variables as input for KSE index prediction. To tackle this problem, our proposal comprises applying machine learning (ML) techniques to predict the KSE during Covid-19. The principal aim of this study is to examine accuracy of combined models with individual models to forecast the Karachi Stock Exchange during COVID-19. This study has analyzed the indices of KSE from March 1st, 2020, to November 26th, 2021. Therefore, this study is keen to find the best-fitted model that forecasts more accurately during the pandemic. To select the most suitable machine learning technique, the six inferred models (i.e., Linear regression (LR), Artificial Neural Network (ANN), Regression Tree (RT), Random Forests (RF), (KNN), and Support Vector Regression (SVR)) are selected to forecast the Karachi Stock Exchange During Covid-19. Performance metrics (i.e., MAE, MSE, MAPE, and R²) are applied to measure and compare accuracy. The modeling outputs presented the RF model provided the best performance of 0.98 versus the other models in predicting the KSE100 index. Thus, the addition of ML methods improves the exchange indications and the competitiveness of future trading guidelines. These projections helped the government to make strategies for the stock exchange KSE-100 and fight against a pandemic disease. The results suggest that the performance of the KSE-100 index can be predicted with machine-learning techniques.

1 Introduction

In the world of today, a significant role is played by the stock markets in determining the economy of a country. The Karachi Stock Exchange 100 Index (KSE-100) is a major stock market index that tracks the performance of the largest companies by market capitalization from each sector of the Pakistani economy. The KSE-100 Index is designed to measure the performance of 100 companies comprising the largest as well as the highest market capitalization. The primary objective of the KSE-100 index is to have a benchmark using which the stock price performance can be compared over a period. In particular, the KSE-100 is designed to deliver investors with the nous of how the Pakistan equity market is performing. Thus, the KSE-100 is similar to other indicators that track various sectors of the country's economic activity such as the gross national product, consumer price index, etc. The market value of firms and stock prices may significantly be affected by numerous factors. Exchange rate fluctuations are one such significant factor. The infectious disease of Novel Coronavirus (2019-nCOV) or COVID-19 first emerged in December 2019 in Wuhan City, Hubei Province of China (WHO, 2020). It quickly got the attention of the world due to its vast spread in January 2020 due to its quick spread compared to other kinds of viruses. The economic losses due to COVID-19 have also affected the global stock markets. The contagion effect of the pandemic on global stock markets has been observed in almost every continent. Pakistan's stock market is also one of the exchanges that have been influenced as a result of COVID-19.

Studies have found that even in the early stages of COVID-19, its impact on the real economy has already been reflected [1], the pandemic has negatively affected trade, tourism, and transportation, and it
increased the unemployment rate [2]; even some studies showed that the global spread of COVID-19 had made a similar impact to an economic crisis [3]. COVID-19 not only has a great impact on economic and social development but also influences the operation of financial markets. Some studies showed the impact of COVID-19 had been reflected in the return of financial markets [4–8], the volatility of financial markets [9–12], and the risk contagion among financial markets [13]. Baig et al. investigated the impact of COVID-19 on the liquidity and volatility of the stock market and found that the increase in confirmed cases and deaths due to COVID-19 had significantly aggravated the lack of liquidity and volatility in the market, and strict closure measures had also deteriorated the liquidity and stability of the market [14]. Rizwan et al. studied the banking systemic risk in eight major countries which were seriously affected by COVID-19; it was found that the financial systemic risk of each country increases significantly during the pandemic period [15].

It is challenging to develop a classified model for the Karachi Stock Exchange prediction during Covid-19 due to the complexity and the interrelated relationships among attributes affecting the outcome. Thus, to ensure modeling accuracy, it is vital to implement a Machine learning (ML) model to provide reliable, accurate, functional, and robust modeling. The current research aims to investigate the impact of COVID-19 on the KSE – 100 Index in Pakistan by considering the other major factors that influence the performance of stock markets such as interest rate and exchange rate. Most importantly, we consider the measures taken by the government in the early response to a pandemic such as the closure of business activities and implementing smart lockdown later by relaxing the restrictions. Several economic measures such as fiscal stimulus and easing of monetary policy to mitigate the slowdown in economic activities are also considered. The Pakistani stock market posted record losses due to jerks to the sentiments of investors facing the lowest six-year intra-day value, that is, a reduction of 28 percent this year Pakistan Institute of Development Economics (2020).

2 Literature Review

As for a prediction, several researchers have identified several predictors that are useful to predict future stock returns. Those include (but are not limited to) dividend yield and dividend-price ratio (Fama and French 1988; Campell and Shiller 1988), price-earnings ratio (Campell and Shiller 1988; Welch and Goyal 2008), short interest rate (Campbell 1987; Ang and Bekaert 2007), term and default spreads (Campbell 1987; Fama and French 1989), and consumption-wealth ratio (Lettau and Ludvigson 2001). Besides predictors, forecasting techniques also play an important role in determining forecast accuracy. According to Mallikarjuna and Rao (2019), traditional regression techniques generally outperform others including artificial intelligence and frequency domain models in providing accurate forecasts. In terms of stock volatility, academic researchers used to make forecasts by traditional GARCH models using indicators based on the past behavior of stock price and volatility (Gokcan 2000; Emenogu et al. 2020). More recent studies become aware of issues such as parametric assumptions, leverage, asymmetric effects, power transformations, and long memory (e.g., Brooks 2007; Bandi and Reno 2012; Hou 2013). In this paper, we introduce GARCH models for volatility forecasting because we aim to test for the instability of the volatility process, which is primarily built upon those modeling techniques.
In addition to a growing number of papers linking COVID-19 and finance relates to stock markets. The literature in this scope contains works published even before the outbreak of the pandemic but suitable for explaining investor behaviors during the COVID period as well as works completed during the pandemic. In the first group, one may find papers addressing the issues of contagion [33], spillovers between markets during shocks [34] as well as the impact of bad news on the time-varying betas [35]. The second, there are papers related to issues of dependencies between global factors and markets [36] or links between individual stock market reactions and the severity of the outbreak of the pandemic in various countries [37]. Moreover, one can list some other works, e.g., related to price movements of stock during the pandemic. Singh [38] found that investors become more attentive to corporate fundamentals and ESG that support the long-run sustainability of firms during turbulence. Fundamental aspects of investments were also pointed out by Mirza et al. [39] who found that social entrepreneurship investment funds outperformed their counterparts during the outbreak of the pandemic. In the field of stock pricing and price tendencies, Shehzad et al. [40] found that the pandemic has influenced the variance of the US, Germany, and Italy’s stock markets stronger than the global financial crisis. Against this background, Narayan et al. [41] and Phan & Narayan [42] found positive effects of lockdowns, travel bans, and economic stimulus packages on stock markets, and Sharif et al. [43] found that in the US the pandemic outbreak has a greater effect on the geopolitical risk and economic uncertainty than the stock market itself.

Moreover, researchers argued that the stock markets are always affected by major events (Haque & Sarwar, 2013; Waheed, Wei, Sarwar, & Lv, 2018). However, as this virus becomes a global pandemic, it starts affecting the businesses which are reflecting in world stock markets. Some studies have examined the impact of COVID-19 on developed stock return (Al-Awadhi, Al-Saifi, Al-Awadhi, & Alhamadi, 2020; Kowalewski & Śpiewanowski, 2020), which reported that the Hang Seng index and Shanghai stock exchange, United States and European stock markets reflect negative returns. In March, the United States marwas ket hita by circuit break mechanism, four times in 10 days. Similarly, the United Kingdom stock market index, FTSE, has a decline of more than 12% worse after 1987 (Al-Awadhi et al., 2020). In Pakistan, the first case of COVID-19 is reported on February 26, 2020, which has crossed the figure of 13,000, till conducting the study. However, the recovery rate is better as compared to developed countries, like Italy, France, and United States. The impact of this pandemic situation on Pakistan's economy depends on the time taken in taking preventive measures and the intensity of spreading the disease. According to the Asian Development Bank (ADB), this pandemic situation can cost the Pakistani economy approximately $16.38 million to $4.95 billion, nearly 1.57% of the overall GDP. The report also mentioned that this pandemic cost more than 946,000 job losses. In this way, a country that is at the recovery stage, in the last 2 years, is affected badly.

This research developed different modeling with several advantages over predictable statistical analysis approaches. Furthermore, the existing traditional model, such as predictive ML is more useful for inferencing patterns or training data rules. Statistical distribution assumptions or postulated functional models provide a statistical regression model arbitrarily and differ from model to model. In contrast, the ML model is extracted from an algorithm centered on accessible data and requires limited user
intervention in model development. Moreover, the adopted modeling techniques can be epitomized using a computational model to consider complicated relationships and configurations via indexes, algorithms, and data structures. Another advantage is that the utilized models could be envisioned in an imaginative manner that suits human understanding. Furthermore, the used data mining-based models are fixable in adapting to the required changes as it their algorithms are developed as an automatic method in a computer framework that can be modified in real-time after upgrading the data sources.

3 Materials And Methods

The main objective of this study is to develop a predictive model for better forecast accuracy of KSE during Covid-19. Moreover, data collection, data preprocessing, ML model comparison, and model performance evaluation and validation are the four primary aspects of the suggested methodology. The framework of the machine learning model proceeds is shown in Fig. 1.

3.1 Data collection and Feature Definition

For accurate KSE-100 index predations, variables need to be recorded and considered. The first step in development of a machine-learning model is collecting data. In this study, the relationship between COVID-19 and KSE was studied.

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

<table>
<thead>
<tr>
<th>No</th>
<th>Variable Label</th>
<th>Data Type</th>
<th>Description of variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>KSE-100</td>
<td>Numerical</td>
<td>The closing price of the index.</td>
</tr>
<tr>
<td>2</td>
<td>USD PER</td>
<td>Numerical</td>
<td>US Dollar (USD) to Pakistani Rupee (PKR) exchange rate</td>
</tr>
<tr>
<td>3</td>
<td>PAK. 10Y</td>
<td>Numerical</td>
<td>Pakistan 10-Year Bond Yield according to over the counter</td>
</tr>
<tr>
<td>4</td>
<td>Covid-19 Deaths</td>
<td>Numerical</td>
<td>the frequency of daily death cases</td>
</tr>
<tr>
<td>5</td>
<td>Covid-19 Reported Cases</td>
<td>Numerical</td>
<td>the frequency of daily reported Covid-19</td>
</tr>
<tr>
<td>6</td>
<td>Covid-19 Recovered Cases</td>
<td>Numerical</td>
<td>the frequency of daily recovered Covid-19</td>
</tr>
</tbody>
</table>

The data of COVID-19 from World Health Organization (WHO) was collected and the KSE-100 index data is taken from Karachi stock exchange (https://www.psx.com.pk/). Thus, the first patient of COVID-19 was reported on February 26, 2020. However, the data of COVID-19 and KSE from March 1st, 2020, to November 26th, 2021, were used by omitting the value of official holidays. Variable descriptions for KSE-100 index model was represented in Table 1. All the data was saved as a numerical value, demonstrating the variables used in the linear regression model (Quantile Regression (QR)).
3.2 Data Pre-Processing

Data pre-processing is a vital stage for developing ML models. The data preparation process was accomplished before each model development to improve the model's prediction. The method comprises several phases that embrace data noise, outlier cleaning, standardization, normalization, characteristic assortment, and conversion. Data assessment was done to tackle inconsistent stock codes. Duplicate instances were removed from the dataset. Cleaning and noise removal procedures were also applied to COVID-19 dataset. In addition, feature aggregations were done to make dataset ready for analysis purpose. The datasets used comprehend numerical values for representing the attributes. The numerical inputs are in very different ranges; therefore, it is essential to standardize the dataset in a close range to enable the model's faster teaching. A z-score was used to convert data, which is the zero mean and standard deviation of the data.

3.3 Machine Learning Models Implementation for KSE Prediction

In this section, briefly describe the methods of ML were used to examine the accuracy of combined models with the individual models in terms of forecasting the Karachi Stock Exchange during COVID-19. Training and testing processes were also conducted to check the suggested ML algorithm's efficiency. The training applied 90% of the database to train the proposed model, and the testing part applied 10% of the database to carry out the test process. 5-fold cross-validation to ensure the robustness and effectiveness of the suggested prediction models was utilized. Six prediction algorithms (i.e., Linear Regression (LR), Artificial Neural Network (ANN), Regression Tree (RT), Random Forests (RF), Regression (KNN), and Support Vector Regression (SVR)) to forecast the KSE-100 are considered among the most recent and efficient machine learning-based prediction algorithms. These algorithms were selected mainly, among other reasons, because they are scalable, accurate, relatively fast, flexible, and more provide regularized model formalization to control over-fitting. The model features and implementation procedure are represented in the next 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 Bachinsky (1998):

\[ y_i = x_i \beta + \epsilon \]
for \( i = 1, 2, 3, \ldots, n \), where \( \beta \) is a \((k \times 1)\) vector of coefficients, \( \{ x_i \} \) 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 \( \epsilon_\theta \) is the unknown error term in the presence of \( \{ x \} \), the \( i^{th} \) conditional quantile of \( \{ y \} \) can be rewritten as

\[
Q_{\text{quant}_\theta} = (y_i | x_i) = x' \beta_\theta.
\]

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

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

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 \( \min_{\beta \in \mathcal{R}^K} \sum_i |y_i - x' \beta| \), otherwise, it's asymmetric.

### 3.3.2 Support Vector Machine (SVM)

SVM is a flexible supervised machine learning process to analyze data amid at both classification, regression, and the other purposes like outlier's detection. To achieve said objectives, it is constructed a hyperplane or set of hyperplanes in a high- or infinite-dimensional space. SVR having some key features to work; 1) Kernel function is used for mapping a lower-dimensional data into higher dimensional data. 2) Hyperplane is a line that draws the separation amid two classes in general SVM. While in SVR, helps to predict the continuous variables and cover most of the datapoints. 3) Boundary lines are the two lines apart from hyperplane, which creates a margin for datapoints. Finally, in 4) Support vectors are the datapoints which 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 used for minimizing an error. In case, it is termed as support vector regression (SVR), if use for regression analysis. Generally, effort in SVR, to consider the maximum data-points within the boundary lines and the hyperplane (best-fit line) must contain a maximum number of data-points.

### 3.3.3 Random Forest (RF)
RF is a well-established supervised learning technique used for classification, regression and other tasks which based on ensemble learning method. It works through 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 trees. To train each decision trees, the training data is splitting into n bags which are used to train their respective decision trees. At the final stage to predict by regressing or taking the average of all trees.

### 3.3.4 K-Nearest Neighbors (KNN)

KNN is one of the simplest and easy-to-implement supervised machine learning process that can be used to solve both classification and regression problems. KNN is robust to the noisy 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 well suite category by using KNN algorithm. In the present study, the actual Manhattan distance was enhanced using weighting. The weighted Manhattan distance is determined by the following:

### 3.3.5 Regression Decision Tree (RDT)

RDT is considered as 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. Decision trees are among the most popular machine learning algorithms given their intelligibility and simplicity.

### 3.3.6 Artificial Neural Network (ANN)

ANN is a commonly used ML model in regression. ANNs deemed in the study involve of three node layers: input, hidden, and output. Each node contains of a transfer function applied to the weighted sum of the previous layer’s nodes and a bias. Linear activation function was evaluated and analysis with others like ReLU, SoftMax activation function. Since our study was linear, SoftMax was the best activation function. An example of an ANN can be seen in Fig. 3, which displays the input (attributes), hidden, and output layers.

$$f(x) = \beta \left( b + \sum_{n=1}^{n} (X_i \cdot w_i) \right) \quad (4)$$

where \( \beta \) the Activation function was defined as

$$\beta(x) = \begin{cases} 0, & x < 0 \\ 1, & x \geq 0 \end{cases} \quad (5)$$
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 towards new data being processed for the first time. In the current research, the suggested prediction models were tested for robustness and effectiveness using 10-fold cross-validation. The training portion (i.e., 80% of the dataset) is used to train the suggested 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 were evaluated Fig 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 a 5-fold cross-validation. A cross-validation model allows you to find the best performing parameters set for any machine learning technique without over fitting.

In Fig. 7, 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. A five-fold cross-validation, iteratively repeats the process for five times, each time using a different subset as validation set and the rest as training data. After the iteration is completed, the hit rate of all five iterations are 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 particular 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.

4.1.1 Performance Measurements

In this section, the prediction performance of our studied 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 indicators (i.e., the Mean Absolute Error (MAE), the Mean Squared Error (MSE), and the mean absolute percentage error (MAPE)). The mathematical representation of four implemented measures is given as below:
Table 2
Mathematical illustration for performance metrics

<table>
<thead>
<tr>
<th>Measure Models</th>
<th>Formula</th>
<th>Variables Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>The Root Mean Squared Error—RMSE</td>
<td>$MSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$</td>
<td>$n$ Sample size</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$y_i$ is the actual value represented by ith case.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\hat{y}_i$ the predicted value represented by the ith case.</td>
</tr>
<tr>
<td>The Mean Absolute Error—MAE</td>
<td>$MAE = \sum_{i=1}^{n}</td>
<td>e_i</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$e_i = (y_i - \hat{y}_i)$</td>
</tr>
<tr>
<td>R-square</td>
<td>$R^2 = 1 - \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$</td>
<td>$y_i$ is the actual value represented by ith case?</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$n$ Sample size</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\hat{y}_i$ the predicted value represented by the ith case.</td>
</tr>
<tr>
<td>The Mean Absolute Percentage Error—MAPE</td>
<td>$MAPE = \frac{1}{n} \sum_{i=1}^{n} \left</td>
<td>\frac{y_i - \hat{y}_i}{y_i} \right</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$y_i$ The actual value represented by ith case.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\hat{y}_i$ The predicted value represented by the ith case.</td>
</tr>
</tbody>
</table>

5 Results And Discussion

5.1 Descriptive Statistics of the Dataset

In the dataset offered, there were five independent (predictor) variables and one dependent (response) variable. The independent variables were USD(PKR), Pak.10Y, Deaths, Auto Liquidated Damage Indicator, Deaths, Cases and Recovered, and the one dependent variable was the KSE.100 prediction. The following tables illustrate the results based upon regression model. The KSE.100-related features data indices’ descriptive statistics, such as such as range, mean, median and quantiles were measured, are calculated, and presented in Tables. According to the descriptive statistics mean; it can be observed that the average price of KSE-100 index has persisted on 37,008 points during the pandemic situation. The average of
positive cases at the beginning of year 2020 is 189166, death cases is 15.00 and the recoveries is 157568 centred on daily basis.

Table 3
Descriptive Statistics Analysis

<table>
<thead>
<tr>
<th>Features</th>
<th>Statistical Methods</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Minimum</td>
</tr>
<tr>
<td>USD(PKR)</td>
<td>153.7</td>
</tr>
<tr>
<td>KSE.100</td>
<td>27229</td>
</tr>
<tr>
<td>Pak.10Y</td>
<td>7.930</td>
</tr>
<tr>
<td>Deaths</td>
<td>0.00</td>
</tr>
<tr>
<td>Cases</td>
<td>9</td>
</tr>
<tr>
<td>Recovered</td>
<td>0</td>
</tr>
</tbody>
</table>

5.2 Meeting Assumptions

The normality test is implemented to test whether the data into attention is normally distributed or not. These tests are provided for the analysis of two numerical procedures, the shape skewness, and the surplus kurtosis. The information sets are normally distributed if those measures are near zero. The acceptance of Jurque-Bera test also focused on skewness and kurtosis. Hence, the test of normality consists of checking the skewness and kurtosis on which the Jurque-Bera test relies. As shown in Figure.1, the distribution of the KSE-100 index is perfectly normal.

Normal Q-Q test is also implemented to test whether the data into consideration is normally distributed or not. The degree of normality can be imagined using a Q-Q-Plot, which plots the histogram of a normal distribution critical of the ML model's predictions. If this gives a straight line, the residuals have a normal distribution. Based on the few skewed distributions of the dependent variable in Fig. 2, it was likely that the residuals 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 ML, we applied the following criteria to the to be transformed.

5.3 Statistical Analysis

In this section, results for the stock forecasting prediction were present. Likewise, contour plot which shows the relation between COVID recovered versus Pakistan 10 Y, KSE 100. The COVID recovered increase as KSE 100 increase, as demonstrated in ‘Figure 10’. Darker regions indicate higher covid recovered. These higher response values seem to form a ridge running from the upper middle to the lower
of the graph. Furthermore, the ‘Figure 10’ shows that the maximum expected KSE 100 index occur at COVID cases more than 300000. The lowest values of COVID cases are in the lower left corner of the plot, which corresponds with low values of both Pakistan 10 Y and KSE100.

This contour plot illustrates the relationship between the Pakistan 10 Y and KSE100 Pakistan 10 Y and KSE100 by covid deaths. Small darker regions indicate less 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 Pakistan 10 Y and KSE100 that result in 100–150 covid death.

The contours are curved because the model contains quadratic terms that are statistically significant. The highest values of rating for line resistance of cotton cloth are in the upper right corner of the plot, which corresponds with high values of both formaldehyde Pakistan 10 Y and KSE100. The lowest values of COVID cases are in the lower down corner of the plot, which corresponds with low values of both Pakistan 10 Y and KSE100. Furthermore, the ‘Figure 10’ shows that the maximum expected KSE 100 index occur at COVID cases more than 300.

### 5.4 Correlation matrix analysis

Correlation matrix among and in between selected attributes to designate the pairs with high negative or positive correlation and the KSE.100 was applied to evaluate the impact of these features, as shown in Figure .3. The current study has further carried out the correlation analysis among the variables USD (PKR), KSE-100, Pak-10Y, Deaths, Cases and Recovered to gauge the strength of their association. The correlation matrix indicates a minor level of positive relationship between KSE-100 and USD (PKR). There were three pairs of features with correlations more than 0.5. One pair was recovered and cases (correlation coefficient = 0.972), and the other pair was KSE.100and Cases (correlation coefficient = 0.848). In addition to KSE.100and recovered (correlation coefficient = 0.892). Highly correlated features added useless information and noise to the models. Furthermore, this study create 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 of eleven independent features, including Covid Cases, Covid Recovered, Pakistan 10Y, Covid Deaths, and USD(PKR), were used as input features in the RF model.
Table 4
Correlation matrix

<table>
<thead>
<tr>
<th>Variable</th>
<th>USD(PKR)</th>
<th>KSE.100</th>
<th>Pak.10Y</th>
<th>Deaths</th>
<th>Cases</th>
<th>Recovered</th>
</tr>
</thead>
<tbody>
<tr>
<td>USD(PKR)</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KSE.100</td>
<td>0.088</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pak.10Y</td>
<td>-0.361</td>
<td>0.425</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Deaths</td>
<td>0.151</td>
<td>-0.199</td>
<td>-0.440</td>
<td>1.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cases</td>
<td>0.337</td>
<td>0.848</td>
<td>0.240</td>
<td>0.044</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>Recovered</td>
<td>0.259</td>
<td>0.892</td>
<td>0.360</td>
<td>-0.145</td>
<td>0.972</td>
<td>1.000</td>
</tr>
</tbody>
</table>

5.5 Comparison Performance of ML algorithms

The proposed system compares six 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 model. 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

Table 5
Regression model's validation and performance evaluation

<table>
<thead>
<tr>
<th>k-fold Cross-Validation</th>
<th>Regression Model</th>
<th>Performance Evaluation Metrics</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RMSE</td>
<td>MAE</td>
</tr>
</tbody>
</table>

\[ k = 5 \]

<table>
<thead>
<tr>
<th>LR</th>
<th>1939.463</th>
<th>1446.226</th>
<th>4.023991</th>
<th>0.751</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rtree</td>
<td>1314.119</td>
<td>1081.647</td>
<td>2.931309</td>
<td>0.870</td>
</tr>
<tr>
<td>RF</td>
<td>512.4574</td>
<td>347.9247</td>
<td>0.959033</td>
<td>0.980</td>
</tr>
<tr>
<td>KNN</td>
<td>720.9236</td>
<td>508.0206</td>
<td>1.415234</td>
<td>0.979</td>
</tr>
<tr>
<td>SVR</td>
<td>2011.759</td>
<td>1351.021</td>
<td>3.808549</td>
<td>0.859</td>
</tr>
</tbody>
</table>

\[ k = 10 \]

<table>
<thead>
<tr>
<th>LR</th>
<th>1829.4429</th>
<th>1245.723</th>
<th>4.023991</th>
<th>0.733</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rtree</td>
<td>1113.4653</td>
<td>981.6789</td>
<td>1.834545</td>
<td>0.824</td>
</tr>
<tr>
<td>RF</td>
<td>534.2294</td>
<td>362.6565</td>
<td>1.005125</td>
<td>0.902</td>
</tr>
<tr>
<td>KNN</td>
<td>699.2345</td>
<td>508.9987</td>
<td>1.35098</td>
<td>0.899</td>
</tr>
<tr>
<td>SVR</td>
<td>1998.1127</td>
<td>1245.0209</td>
<td>2.98763</td>
<td>0.859</td>
</tr>
</tbody>
</table>
created in which training data (90%) are randomly selected Table 5 shows a detailed description of the best performing parameters set generated during the cross-validation on 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. The comparison outcomes present RF and KNN have higher measures of $R^2$ value and lower MAE, RMSE, MAPE values than the traditional models (i.e., LR, DT, RF, KNN and SVM)) in the KSE-100 predictions. Overall, the results suggested that the RF model exceeded on our data set in terms of RMSE, with values of 512.4574 and 720.9236, respectively. The mean accuracy of the RF model was 0.98, which was the same excellent as that of the KNN model. As for other metrics, the RF model showed a higher $R^2$ and a lower MAE, MSE, RMSE. Moreover, the predicted results of the proposed demonstrate that the RF estimate values are very high $R^2$ measures with the less MSE. Therefore, the plots compare the predicted performance of the proposed models at K = 5, and 10. Overall, the RF model exhibited better overall performance with an $R^2$ of 0.98, and the KNN model achieved acceptable performance in predicting the KSE100 with an $R^2$ of 0.979. Both can achieve higher predictive accuracy compared with other ML models. As a result, the RF model showed a higher $R^2$ and a lower MAE, MSE, RMSE. Moreover, the predicted results of the proposed demonstrate that the RF estimate values are very high $R^2$ measures with the less MSE. Therefore, the plots compare the predicted performance of the proposed models at K = 5, and 10. Overall, the RF model exhibited better overall performance with an $R^2$ of 0.98, and the KNN model achieved acceptable performance in predicting the KSE100 with an $R^2$ of 0.979. Both can achieve higher predictive accuracy compared with other ML models. As a result, the $RF$ model is the most efficient and proficient model in predicting the KSE100. On the other hand, the results also indicate that $RF$ had remaining prediction ability compared to KNN.

5.6 Features importance analysis

An improved understanding of the model's features assistances investors effectively 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. The present study, RF models contained Covid Cases, Covid Recovered, Pakistan 10Y, 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. Features' significance was in down order as follows: Covid Cases, Covid Recovered, Pakistan 10Y, Covid Deaths, and USD(PKR)'. Figure 14 presentations the top three most importance ranking of the selected features used in RF predictive models, respectively. Covid Cases was the first important feature in predicting of KSE.100. In addition, Covid Recovered and Pakistan 10Y were the second and third most important feature in predicting of the RF model as respectfully, but another feature was the less important feature.

Further studies would be helpful to improve the predictive accuracy by inducing more ML models, such as support vector machine (SVM) and deep neural network (DNN). The rank of feature importance showed the contribution of each factor to the forecast prediction. Because a more considerable value suggests a more significant correlation between the features and the outcome.

6 Conclusion
In this study, six machine learning models (i.e., LR, DT, RF, KNN, and SVM)) are used to enhance prediction accuracy for KSE100 using related features. Performance metrics of the (i.e. MAE, MSE, APE, and $R^2$) under the 5-fold cross-validation process have been applied to enhance the strength and efficiency of our models. The results of this study confirm that RF method is efficient of predicting the stock market performance model has better than other methods. The relationship between each of these factors and market behavior was studied individually using Correlation. Covid case was the most important factors. Furthermore, the created models are suitable for studying impact of COVID-19 on KSE – 100 Index by considering the other major factors that influence the performance of stock markets such as interest rate and exchange rate. The predicted stock prices assist investors to make smart investment decisions as well as aid analysts to predict and study trends market stocks.

**Declarations**

**Conflicting Interests**

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

**Funding**

The authors received no financial support for the research, authorship, and/or publication of this study.

**Credit Authorship Contribution Statement**

**Tahir Munir**: Conceptualization, Methodology, Formal analysis, Writing - original draft, Writing - review & editing. **Rabia Al Mamlook**: Methodology, Formal analysis, Writing- original draft. **Abdu R Rahman**: Data analysis, Results interpretations. **Sujeet Shrestha**: Methodology, Formal analysis, Writing -review. **Mohamed Bzizi**: Writing - original draft, Data curation. **Abeer Aljohani**: Formal analysis and Writing -original draft.

**References**


Figures
Figure 1

Framework of machine learning model proceed
Figure 2

Support Vector Machine
Figure 3

Flowchart of the RF’s architecture
Figure 4

K-Nearest Neighbors
Figure 5

Regression Decision Tree
Figure 6

Multi-Layer Perceptron with two hidden layers
Figure 7

Schemes of k-fold Cross-validation
Figure 8

Histogram of KSE-100 index distribution

Figure 9

Normal Q-Q
Figure 10

Contour plot that shows Covid between Pakistan vs KSE
Figure 11

Contour plot between Pakistan and KSE
Figure 12

Contour plot between Pakistan 10Y vs KSE100
Figure 13

Plot compared the predicted performance of models
Figure 14
Features importance analysis