



Research article

Determinants of Egypt's unemployment rate with machine learning algorithms

Mohamed F. Abd El-Aal*

Economics Department, Faculty of Commerce, Arish University, North Sinai, Egypt

* **Correspondence:** Email: Mohammed.fawzy@comm.aru.edu.eg.

Abstract: This research endeavored to harness the capabilities of machine learning algorithms, specifically gradient boosting (GB) and random forest (RF), to ascertain the optimal algorithm for precisely analyzing the determinants influencing Egypt's unemployment rate. Among the algorithms employed, the study identified the gradient boosting (GB) algorithm as the most accurate in this context. The investigation discerned vital variables that substantially influence Egypt's unemployment rate. The analysis revealed that the industry sector value notably impacts the unemployment rate, accounting for 37.3%, and the GDP growth contributes 23.7%. Additionally, variables such as the value of imports (11.7%) and labor force participation (9.5%) emerged as significant determinants. Moreover, the research demonstrated that the unemployment rate exhibits negative correlations with factors including the value added in the agriculture sector, foreign direct investment (FDI), GDP growth, and gross fixed capital formation. Conversely, the unemployment rate positively correlates with variables such as inflation, human development index, imports, labor force participation, industry sector value added, and services sector value.

Keywords: unemployment rate; machine learning algorithms; gradient boosting; random forest

JEL Codes: E24, C80, C81, C87

1. Introduction

Unemployment is a pivotal socio-economic issue that nations strive to comprehend and address due to its far-reaching consequences on individual well-being, economic stability, and societal harmony. Egypt, a country of historical significance and diverse economic dynamics, faces the intricate challenge of managing its unemployment rate. The persistent pursuit of strategies to tackle unemployment and foster sustainable growth has led to a growing interest in unraveling the factors contributing to Egypt's unemployment landscape. Applying machine learning algorithms to this inquiry presents an innovative and promising avenue for understanding these complex interactions in the era of data-driven insights and advanced computational techniques.

The determinants of unemployment are multifaceted, encompassing a wide array of economic, social, and demographic factors. Traditional economic models and statistical analyses have provided valuable insights into these determinants, but their ability to capture intricate nonlinear relationships and patterns may be limited. Here, machine learning algorithms offer a distinct advantage. By leveraging historical data from 1991 to 2022, this research seeks to harness the power of machine learning to explore the determinants contributing to Egypt's unemployment rate. Egypt's unemployment rates witnessed various variations from 1991 to 2022. During this period, the unemployment rate reached its highest in 2013, at 13.2%. During the past years, the rate decreased until it reached its lowest, 6.9%, in 2022 (World Bank data).

Machine learning, a branch of artificial intelligence, identifies hidden patterns within vast and complex datasets. Its capacity to discern nonlinear relationships, detect interactions, and handle high-dimensional data makes it a compelling tool for unraveling the nuanced factors influencing unemployment. This paper aims to provide a more granular and holistic understanding of Egypt's unemployment landscape through machine learning algorithms, identifying key determinants that may have previously gone unnoticed.

This research bridges the gap between traditional economic analysis and modern computational techniques. By combining machine learning algorithms with economic theory, we seek to uncover the determinants of Egypt's unemployment rate and provide a deeper comprehension of the intricate interplay between these determinants. This holistic approach is anticipated to provide valuable insights for policymakers, researchers, and stakeholders, facilitating the design of targeted interventions to reduce unemployment and promote sustainable economic development.

Below is a summary of this research paper's structure: Section 2 gives a thorough analysis. The existing literature highlights key determinants of unemployment and the potential of machine learning in this context. Section 3 includes the theoretical and empirical framework, and Section 4 outlines the methodology employed, detailing data collection, preprocessing, feature selection, and the selection of machine learning algorithms. Section 5 presents the results of the analysis, discusses the implications of the findings, and provides insights into the importance of various determinants. Finally, Section 6 concludes the study by summarizing the key takeaways, discussing the limitations, and suggesting directions for future research.

Through this research, we aim to contribute to the ongoing unemployment discourse by harnessing the power of machine learning algorithms. By gaining a deeper understanding of the determinants of Egypt's unemployment rate, we aspire to provide evidence-based insights that can inform effective policy decisions and contribute to Egypt's economy's sustainable growth and prosperity.

2. Literature review

Gross fixed capital formation, wage proxy, and trade openness were all analyzed to see how they affected employment in Pakistan (Imran & Sial, 2013) using historical data from 1997–2010. Researchers employed the cointegration and unit root tests to investigate the qualities of time series data and establish any underlying long-term relationships. Two cointegrating vectors appeared below the research period, as the results indicated. The results also showed that gross fixed capital formation significantly affects full employment. Using historical data from 1980 and 2014, Megbowon and Mushunje (2016) analyzed the connection between gross capital formation, foreign direct investment inflow, and the unemployment rate in South Africa. This research utilized cointegration and causality tests on time series data, and the results of the study indicated that the variables in the employment model share a long-term relationship.

Using time series data from 1976 to 2012, Sattar and Bhalli (2013) investigated the root causes of unemployment in Pakistan. The correlation between GDP growth, population size, FDI, unemployment, and inflation rate was also analyzed. Autoregressive distributed lag (ARDL) was used in this study to determine what caused unemployment in Pakistan. The study found that the long-term and short-term reasons for unemployment in Pakistan were the country's growing population, rising inflation, stagnating GDP, and a lack of FDI. Hence, Qadar & Muhammad (2013) looked into this hypothesis. Time series data were used in the study, which ranged over time from 1976–2010. The study used the OLS technique and the unit root test to investigate the link between GDP and unemployment. The data showed that a 1% rise in the unemployment rate will result in a drop of 0.36% in real GDP growth.

Some Arab countries' unemployment rates were compared to their economic growth rates (see Abdul-Khaliq & Shihab, 2014). The overarching goal of this article was to measure the correlation between the unemployment rate and GDP growth rate in a few Arab countries. Statistics from 1994 to 2010 were combined. According to the research, economic expansion significantly and negatively impacts unemployment. In other words, a 1% increase in GDP during the research period was associated with a 0.16% drop in unemployment. The causes of unemployment in Nigeria's metropolitan areas were studied by Bakare et al. in 2011. Time series data were analyzed for the years 1978–2008. Unit root testing and OLS regression were performed to ensure the variables were stationary. In other words, the results showed that rising unemployment in Nigeria throughout the research period was caused by population expansion.

The effects of Nigeria's ballooning public debt on the country's job market have been investigated (Ogonna et al., 2016), where they used a data set of annual numbers from 1980 to 2015. Autoregressive models and other econometric analytic tools were employed. The findings demonstrated a link between foreign debt and unemployment in the long term. In addition, the ARDL model found that an increase of only 1% in the national debt was associated with a 1.6% rise in the jobless rate. According to Dalmar, Ali, and Ali (2017), the relationship between population, GDP, and foreign debt correlates positively with the unemployment rate while negatively with gross capital formation and the exchange rate.

Abugamea (2018) proved that GDP has a negative impact on unemployment, inflation, and labor, and constraints on the labor force have been shown to have a beneficial effect on the unemployment rate. The research also shows that international commerce does not affect unemployment. Salama and Judit (2019) proved that unemployment among men and women has been negatively correlated with

economic freedom. Furthermore, the effects of the financial crisis of 2008 have not substantially added to the overall unemployment rate.

Another paper used annual time series data ranging from 1982 to 2013 for the analysis (Azhar & Ibrahim, 2021). The unit root hypothesis was examined using the augmented Dickey-Fuller (ADF) test, which used college enrollment, inflation, GDP, and population data. The primary findings revealed that tertiary education enrollment and inflation significantly affect unemployment, but the gross domestic product and population have insignificant effects. Siddiqa (2021) shed new light on the factors contributing to unemployment in developing countries. The World Bank's database was examined for data on ten developing countries between 2000 and 2019. All of the factors were found to have statistically significant results. Unemployment is positively influenced by population and external debt and, in contrast, negatively by GDP, inflation, remittances, the currency rate, and education spending.

Digvijay (2021) analyzed information covering 2000 and 2019 and used regression analysis to examine the relationships between India's unemployment, GDP, and inflation rates. The result showed that GDP growth significantly affects the unemployment rate. The unemployment rate in India has been decreasing along with the country's GDP. The country's inflation rate does not impact India's unemployment rate.

Gogas et al. (2022) aimed to make broad predictions about joblessness in the Eurozone. No other research has made any predictions about the unemployment rate in the Euro area to the greatest extent of our knowledge. The dataset, which runs from 1998:4 to 2019:9 with monthly frequency, includes the rate of unemployment and 36 explanatory variables based on theoretical and empirical recommendations. Three machine learning methods—decision tree (DT), random forest (RF), and a support vector machine (SVM)—were then used to feed the data into econometrics-based elastic-net and logistic regression models. The results showed that the best RF model had an out-of-sample accuracy of 85.4% and a full-dataset accuracy of 88.5%.

3. Theoretical and empirical framework

3.1. Theoretical framework

3.1.1. Neoclassical thoughts

According to the neoclassical approach of thinking about the labor market, labor demand is derived demand, which is determined by the marginal revenue product of labor (McCann, 2001). For a fixed price on output and stock of capital, the marginal product of labor falls as labor input rises (the "law of diminishing marginal productivity"). So, if we plot wage on the y -axis and the number of workers on the x -axis, we get a demand curve that slopes downward.

Furthermore, the neoclassical approach assumes that workers' decisions about their labor supply, the income they want to earn, and the quantity of commodities they want to consume are all influenced by absolute wage levels (McCann, 2001). If the substitution effect on labor supply is more significant than the income effect, then the labor supply will increase as real wages rise. To rephrase Bowles and Gintis (1975), neoclassical economists think that the link between pay and labor may be viewed in the same way as the relationship of any other commodity, since they view labor as a good or service.

Companies would raise their demand for labor if the price of labor cost decreased, which is the same as a salary decrease (McCann, 2001).

3.1.2. Keynes's thoughts

Keynes's method of intervening in the economy and lowering unemployment worked until the late 1960s. Constant state interference, however, has increased public spending, contributing to budget deficits. The rising rate of inflation has highlighted the severity of the economic crisis. After the Keynesian approach failed to resolve these issues, the monetarist school of economics developed. *Stagflation* is the term given to describe the 1970s period in which inflation and unemployment co-occurred for no apparent reason. The situation did not improve despite the implementation of demand-boosting policies. This is because monetary policies meant to combat unemployment also contributed to inflation, and vice versa.

The monetarist perspective and the idea of natural unemployment gained widespread recognition thanks to developments in the 1970s. Significant shifts in economic thought have resulted from the monetarist perspective and the prevalence of natural unemployment. Given this reality, the pursuit of full employment has been largely abandoned in favor of price stability as the primary objective of economic policy. This argument holds that using fiscal policy to distinguish the unemployment rate has no effect other than to raise prices across the board (Turgut, 2021).

3.1.3. The new Keynesian thoughts

The new Keynesian economic approach diverges from the Keynesian school by advocating for more government involvement in economic policy and management. The new Keynesian perspective does not rule out the possibility of state intervention. Still, it does hold that such action should be taken only when necessary and at a pace commensurate with the severity of the problem. The new Keynesian explanation for the persistence of unemployment makes sense when considering wage stickiness, the primary challenge facing the labor market. Wage stickiness is studied by the new Keynesian economics, which attempts to explain it by focusing on three main theories: inside outs, wage efficiency, and implicit contracts (Turgut, 2021).

3.2. Empirical framework

The unemployment rate is affected by many economic variables, especially the country's gross domestic product growth, the labor force growth rate, the foreign direct investments inflow, gross fixed capital formation, the country's exports, the country's imports, the level of human development, inflation rate, economic sectors value added, and wages (complete data on this is not available). The following equation represents the factors that contribute to Egypt's unemployment rate:

$$U_{ne} = GDPg + LFG + FDI + GFCf + EX + IM + HDI + INF + AGV + SRV + INV \quad (1)$$

where:

U_{ne} = Unemployment rate

$GDPg$ = Gross domestic product growth

LFg = Labor force growth
FDI = Foreign direct investment
GFcf = Gross fixed capital formation
EX = Exports
IM = Imports
HDI = Human development index
INF = Inflation rate
AGV = Agriculture sector value added
SRV = Services sector value added
INV = Industry sector value added

4. Data and methodology

4.1. Data

This paper contains many economic variables collected from the World Bank and UNESCO database from 1991 to 2022. To determine the factors affecting the unemployment rate in Egypt, the following variables were relied upon: gross domestic product growth, the labor force growth rate, the foreign direct investments inflow, gross fixed capital formation, the country's exports, the country's imports, the level of human development, inflation rate, and economic sectors value added. Table 1 shows the descriptive statistics of the variables:

Table 1. Description and descriptive statistics of the variables.

Variables	Abbrev.	Source of data	mean	mode	median	dispersion	min	max
Unemployment rate	UNe	World Bank data	9.95	6.95	9.61	0.17	6.95	13.15
Gross domestic product growth	GDP	World Bank data	4.39	1.12	4.42	0.35	1.12	7.15
Labor force participation	LFp	World Bank data	46.18	41.31	46.84	0.004	41.31	49.37
Foreign direct investment	FDI	World Bank data	2.3	-0.2	1.48	0.96	-0.2	9.34
Gross fixed capital formation	GFcf	World Bank data	18.46	12.44	18.04	0.2	12.44	27.06
Exports	EX	World Bank data	31.80	10.26	33.28	0.53	10.26	71.92
Imports	IM	World Bank data	44.69	12.93	39.67	0.59	12.93	104.38
Human development index	HDI	UNESCO	0.64	0.55	0.64	0.08	0.55	4.70
Inflation rate	INF	World Bank data	9.77	2.26	9.39	0.58	2.26	29.50
Agriculture sector value added	AGV	World Bank data	13.61	10.69	13.64	0.14	10.69	16.98
Services sector value added	SRV	World Bank data	48.63	44.69	47.61	0.058	44.69	54.48
Industry sector value added	INV	World Bank data	33.51	28.40	33.04	0.092	28.40	39.89

Source: Compiled by the author.

4.2. Methodology

This investigation uses gradient boosting and random forest algorithms as machine learning tools. Each model is a supervised machine learning model, meaning it learns from historical data and uses that information to create a predictive function for new data. The Scikit-Learn Python package is utilized to implement all of the machine learning methods used in this paper.

4.2.1 Machine learning algorithms (ML)

(1) Gradient boosting (GB)

The GB algorithm, developed by Friedman (2001), is an ensemble machine learning approach. Combining numerous inefficient learners into a single strong one is the core notion behind the gradient-boosting paradigm.

A single leaf is created to construct regression trees in the gradient-boosting model. Instead of serving as a classifier, regression trees (a subset of decision trees) are built to provide estimates for continuous real-valued functions. Through repeated iterations, the data is partitioned into smaller and smaller subsets as the regression tree is built. All of the data points are pooled together at the outset. The information is then divided in half using every split on every predictor that can be used. This study uses the Friedman MSE established by Friedman (2001) to quantify residual error, and the predictor that splits the tree is the one that best divides the observations into two separate sets.

The GB algorithm builds a new tree based on the preceding tree's error. It keeps doing so until the desired number of trees has been reached or the model can no longer be improved. To avoid overfitting, the gradient-boosting method gradually increases the weight provided to the new tree based on the learning rate.

The algorithm of the GB includes the following phases for the input data, $[(x_i, y_i)]_{i=1}^n$, and a loss function that can be differentiated: $L(y_i, F(x))$.

Step 1: Set a fixed value as the model's starting point.

$$F_0(x) = \operatorname{argmin}_y \sum_{i=1}^n L(Y, \hat{y}) \quad (2)$$

where Y is a value as observed, and \hat{y} is a value as predicted. $F_0(x)$ is the average of the observed values.

Step 2: For $m = 1$ to M :

- Calculate

$$\hat{y}_{im} = - \left[\frac{\partial L(Y, F(x_i))}{\partial F(x_i)} \right]_{F(x)=F_{m-1}(x)} \quad \text{for } i = 1, \dots, n \quad (3)$$

- Adjust a tree for regression to the \hat{y}_{im} values and generate terminal regions R_{jm} for $j = 1, \dots, jm$.
- For $j = 1, \dots, jm$, calculate

$$\hat{y}_{jm} = \operatorname{argmin}_y \sum_{x_i \in R_{ij}} L(Y_i, F_{m-1}(x_i) + \hat{y}) \quad (4)$$

- Modify

$$F_m(x) = F_{m-1}(x) + \alpha \sum_{j=1}^{J_m} \hat{y}_m I(x \in R_{jm}) \quad (5)$$

where α is the learning rate.

By adjusting the learning rate, α , the user can alter the behavior of the employed loss functions. This feature enhances the model's flexibility while reducing the overfitting issue through the lessons learned during the slower iterations (Hastie et al., 2009).

Step 3: Output

$$\hat{F}(x) = F_M(x) \quad (6)$$

Once all M iterations have been completed, the $F_M(x)$ function will have been updated, and the final model, $\hat{F}(x)$, will approximate the connection between the independent and the dependent variables.

(2) Random forest algorithm (RF)

Breiman (2001) presented the RF algorithm, another ensemble technique analogous to boosting models. As Dietterich (2000) stated, random forest is one of the most influential ensemble algorithms for machine learning. The RF algorithm employs regression trees like the gradient boosting model. In contrast to the GB algorithm, the regression trees in the random forest model are trained individually, and the trees' predictions are averaged.

The random forest model consists of the following primary procedures:

Step 1. For $m = 1$ to M :

- To do this, take the training data and generate a bootstrapped sample set, Z , with a size of N .
- Follow these methods for each terminal node in the tree to grow a random T_m forest for the bootstrapped data, with a minimum node size of n_{\min} .
 - To do this, randomly choose x of the p variables.
 - The point will be divided among the x variables; select the best one.
 - Create a new node by dividing the existing one in half. The MSE is minimized by making the following calculations to determine how to divide the data:

$$F_0(x) = \frac{1}{n} \sum_{i=1}^n (Y - \hat{y})^2 \quad (7)$$

where Y is an observed value, and \hat{y} is a predicted value.

In addition to bootstrapping specific data for each tree predictor, we introduce randomness by allocating variables to divide each node. This random approach significantly lessens inter-tree dependencies and increases robustness against overfitting. If a fully grown tree fits the model too well, it is said to be overfitting. Adding additional data could cause a model with almost ideal tree fits to fail to predict future outcomes reliably. To prevent this issue, an RF model may reduce the size of the forest by cutting down on the number of trees or nodes.

Step 2. Output of the ensemble of trees, $\{T_m\}_{m=1}^M$:

$$\hat{F}_{rf}^M(x) = \frac{1}{M} \sum_{m=1}^M T_m(x). \quad (8)$$

The final output, $\hat{F}_{rf}^M(x)$, is determined by averaging the results from each tree. The predictive performance of trees is stabilized, and the variance is reduced when several forecasts are averaged.

4.2.1. Cross-validation

The ML algorithms use several hyperparameters in this investigation. K-fold cross-validation, widely used in studies like this one, is used to fine-tune the hyperparameters. The k-fold cross-validation technique splits the training set into k equal halves to test how well a model fits the data. K-fold cross-validation uses a training set and a test set with varying start and end times to account for the fact that data may evolve. Future data cannot be used to assess previous data since the forecasting model should not include any information about events that occur after the events used to fit the model (Tashman, 2000). Following previous works such as Molinaro et al. (2005), we set $k = 10$ and split the training data into 10 subgroups for model training and fitting.

It is possible that some people believe the cross-validation method is unnecessary because random forest models use trees that are obtained during the bagging process. The out-of-bag technique is comparable to cross-validation, so it is possible that RF models do not require cross-validation. Our study is primarily motivated by comparing the effectiveness of GB and RF models. To make sure the comparisons were as objective as possible, this study applied the cross-validation approach on the RF model. The out-of-sample data for the GB and RF models were set to be the same in order to further ensure a fair comparison (El-Aal, 2024).

The cross-validation procedure aims to pick the hyperparameters that result in the slightest mean squared errors across all ten testing subsets. In other words, predictions will be made using the test data set and the hyperparameters that were determined through the cross-validation procedure. Previous research employed a grid search approach to find the optimal settings for the provided hyperparameters (Probst et al., 2019). All predictors were taken into account, and in both the GB model and the RF model, the depth of the trees was adjusted by varying the number of splits. Cross-validation aims to identify the hyperparameters that, when combined, result in the slightest mean squared error (MSE).

5. Empirical results

5.1. Model evaluation

The most frequent metrics used in regression analysis to assess error rates and best accuracy in forecasts are the mean absolute error (MAE), the mean squared error (MSE), the root mean squared error (RMSE), and the coefficient of determination (R_2).

Calculated using equation (9), the MAE represents the discrepancy between observed and predicted values by averaging the absolute difference throughout the data set.

$$\text{MAE} = \frac{1}{n} \sum_1^n |y_i - \hat{y}_i| \quad (9)$$

The MSE is the average squared error, and it is calculated by multiplying the average difference between the actual and projected values by four. The following formula (10) can be used to determine this value:

$$MSE = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y})^2 \quad (10)$$

Multiply the error rate by the square root of the MSE to get the RMSE. You can figure it out with this formula (11):

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y})^2} \quad (11)$$

The R^2 value expresses the degree of fit between the values and the initial values. R^2 is calculated using formula (12)—a larger value denotes a more accurate model—and it is the proportion of numbers that is neither zero nor one (Abd El-Aal, 2023).

$$R^2 = 1 - \frac{\sum (y_i - \hat{y})^2}{\sum (y_i - \bar{y})^2} \quad (12)$$

5.2. Algorithms' accuracy and performance

Cross-validation accuracy results for the utilized algorithms, processed in Python, are shown in Table 2.

Table 2. Machine learning algorithm's cross-validation accuracy in predicting unemployment rate, 1991–2022.

Model	MSE	RMSE	MAE	R^2
Gradient Boosting	1.32	1.14	0.95	57
Random Forest	1.99	1.4	1.2	50

Source: Compiled by the author.

Table 2 shows that the most accurate predictions may be found with gradient-boosting methods, with random forest coming in a close second.

Regarding the two algorithms' performance, the near-identical observed and predicted values in Table 3 and Figure 1 demonstrate the efficacy of the GB and RF algorithms in predicting, leading to effective economic strategies.

Table 3. Gradient-boosting algorithm and random forest prediction for Egypt's unemployment rate (%).

Year	Actual Unemployment Rate (%)	GB Algorithm's Prediction	RF Prediction	Algorithm's Prediction
1991	9.38	9.39	9.75	
1992	8.92	8.93	9.7	
1993	10.92	10.89	10.85	
1994	10.93	10.92	10.85	
1995	11.04	11	10.42	
1996	9	8.99	8.8	
1997	8.37	8.37	8.4	
1998	8.03	8.03	7.99	
1999	7.95	7.97	8.26	
2000	8.98	8.96	8.89	
2001	9.26	9.28	9.6	
2002	10.01	10.01	9.76	
2003	10.91	10.89	10.04	
2004	10.32	10.35	10.33	
2005	11.049	11.01	10.44	
2006	10.49	10.47	9.68	
2007	8.8	8.8	9.22	
2008	8.517	8.53	9.06	
2009	9.087	9.1	9.69	
2010	8.756	8.77	9.27	
2011	11.849	11.83	11.55	
2012	12.597	12.62	12.69	
2013	13.154	13.12	12.86	
2014	13.105	13.10	12.86	
2015	13.052	13.04	11.68	
2016	12.45	12.43	11.52	
2017	11.767	11.77	11.56	
2018	9.855	9.83	10.18	
2019	7.851	7.85	8.98	
2020	7.939	7.93	8.91	
2021	7.4	7.41	8.71	
2022	6.959	6.97	8.43	

Source: Compiled by the author.

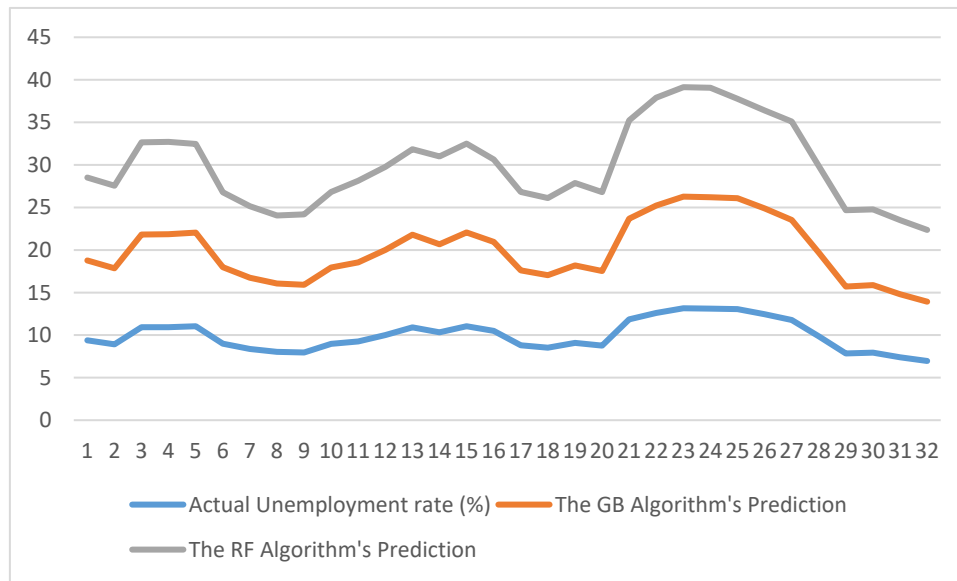


Figure 1. The GB and RF prediction performance.

5.3. Gradient-boosting and random forest algorithms are essential features

While these methods are most often used for making predictions, feature importance analysis can be used to determine which factors have the most bearing on a given model. Table 4 shows the results of this study.

Table 4. The GB and RF algorithms feature importance indicators.

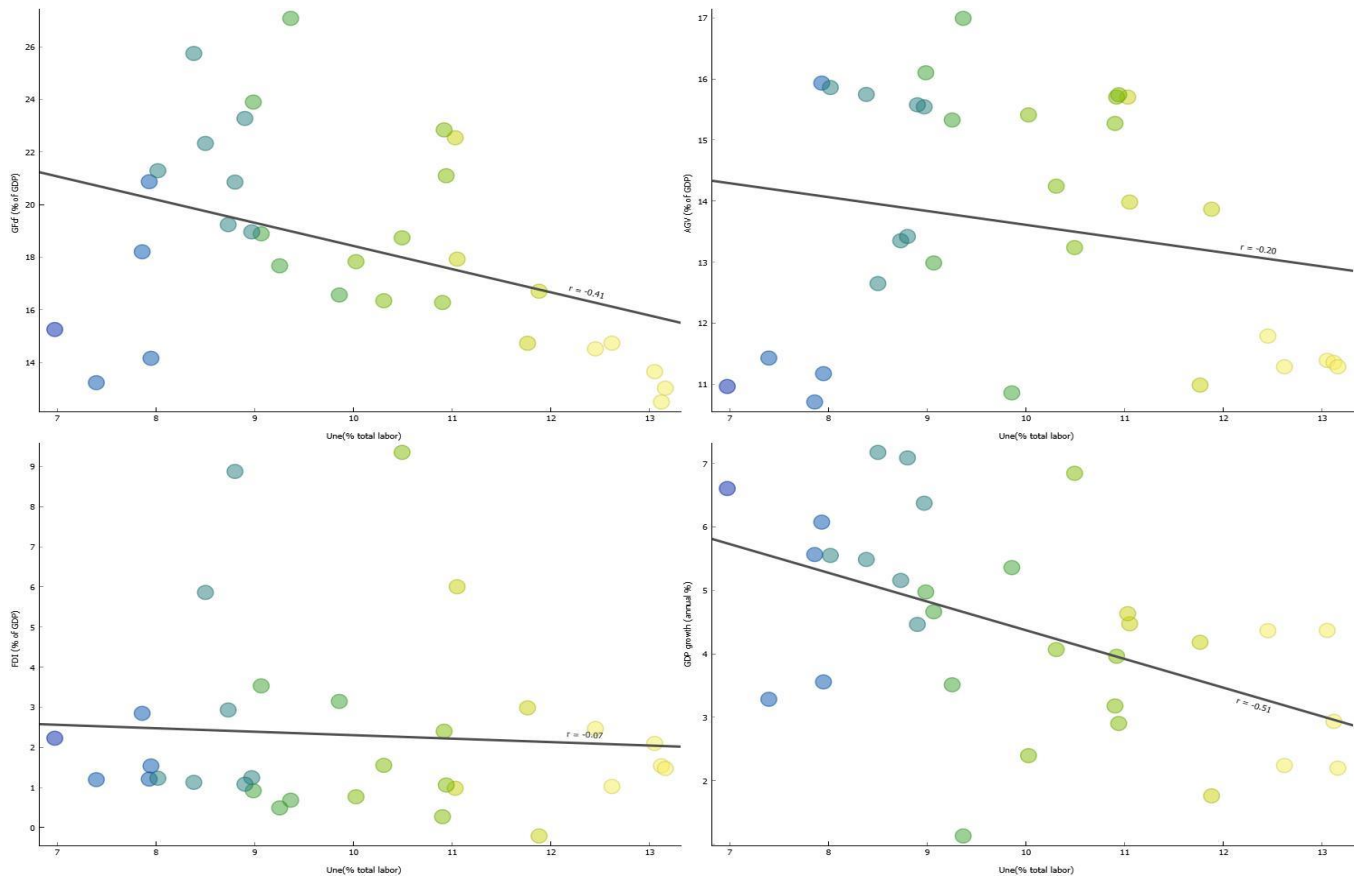
Variables	GB feature importance	RF feature importance
INV	37.3	30.2
GDPg	23.7	25.1
IM	11.7	2.7
LFp	9.5	4.5
SRV	6.7	20
HDI	4.4	3
AGV	2.3	0.5
GFcf	1.8	11.3
FDI	1.8	0.1
INF	0.5	1.3
EX	0.4	1.3

Source: Compiled by the author.

According to the GB algorithm, the unemployment rate in Egypt is primarily affected by the added value of the industrial sector by 37.3%, followed by the GDP growth of 23.7%. The volume of state imports is 11.7%, and labor force participation approaches it with a ratio of 9.5%.

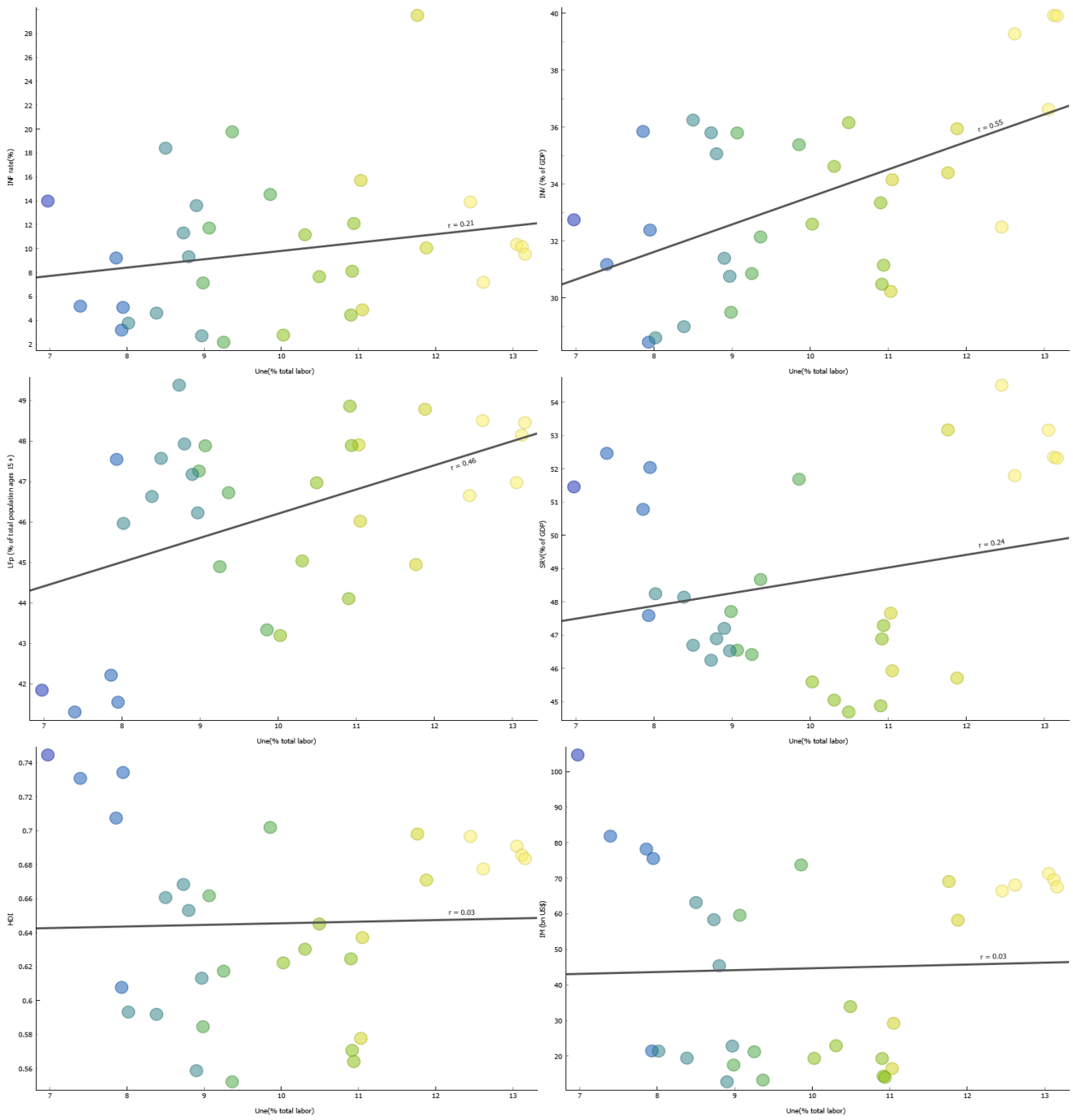
The RF algorithm's results are close to those of the GB algorithm, especially in that the main determinants of unemployment are the added value of the industrial sector and the economic growth rate in Egypt.

We can also determine the relationship between the dependent and independent variables through the following scatter plots.



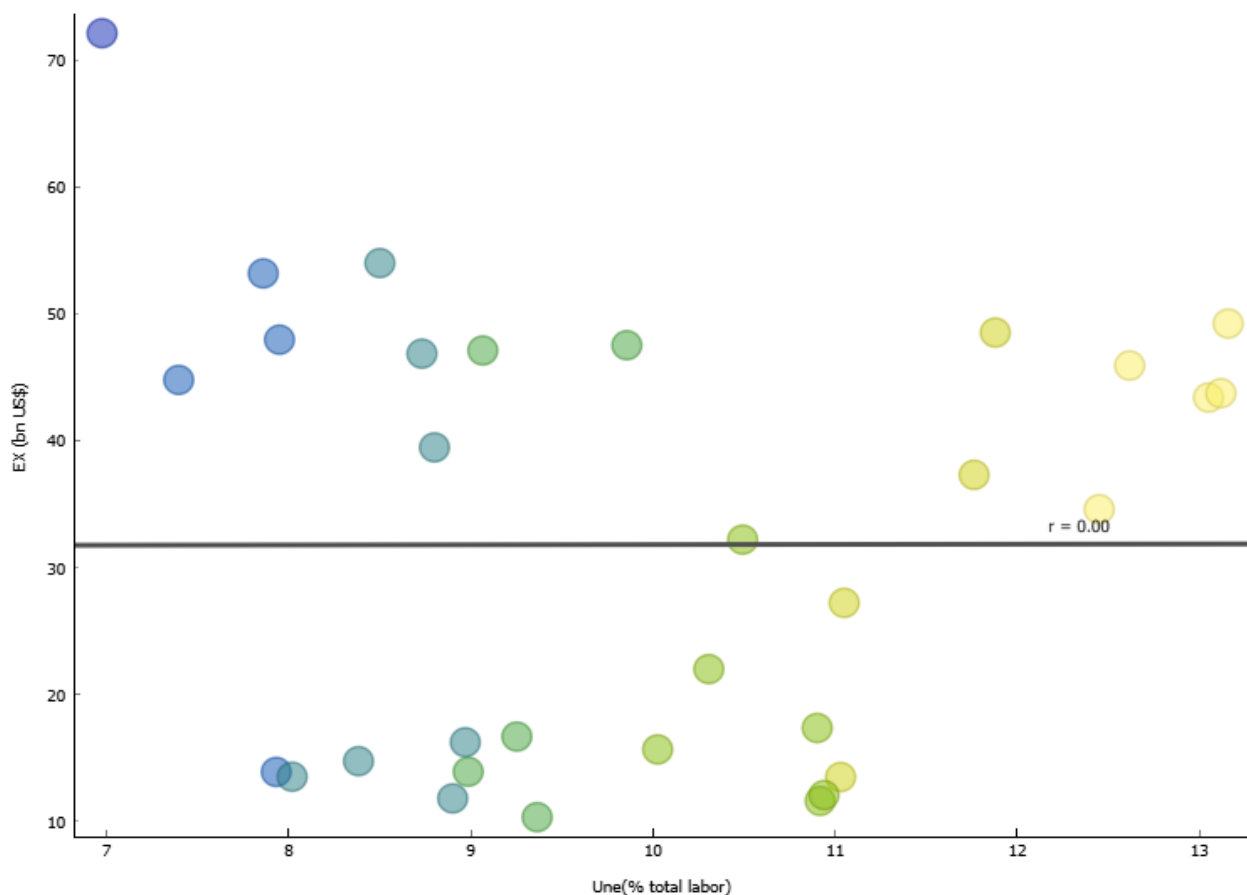
Source: Compiled by the author.

Figure 2. Scatter plots for the negative relationship between the unemployment rate and other variables.



Source: Compiled by the author.

Figure 3. Scatter plots for the positive relationship between the unemployment rate and other variables.



Source: Compiled by the author.

Figure 4. Scatter plot for the fixed relationship between the unemployment rate and exports.

According to the Pearson correlation and its translation in Figures 2, 3, and 4, the unemployment rate negatively relates to agriculture sector value added, FDI, GDP growth, and gross fixed capital formation. Moreover, it has an appositive relationship with inflation (which indicates that Egypt's economy is in many phases of stagflation), HDI (indicating that the unemployment rate increases by increasing in tertiary education), imports, labor force participation, industry sector value added (that is, it depends on technology-intensive industries), and services sector value added.

6. Conclusions

In conclusion, the research study “Determinants of Egypt's unemployment rate with machine learning algorithms” utilizes advanced machine learning techniques to elucidate the intricate relationship between various factors and Egypt's unemployment rate. Through meticulous data analysis and the application of cutting-edge algorithms, the study contributes to a comprehensive understanding of the multifaceted nature of unemployment dynamics within the Egyptian context.

The utilization of machine learning algorithms enables the identification of key determinants that significantly influence the unemployment rate, offering a nuanced perspective on their interplay. The predictive capabilities of the employed algorithms furnish valuable insights for policymakers, economists, and stakeholders, facilitating the development of targeted strategies to address

unemployment challenges in Egypt. The findings underscore the importance of integrating modern computational tools into economics and social studies. Machine learning algorithms, equipped to process vast datasets and discern complex patterns, prove invaluable in uncovering hidden relationships that traditional methods might overlook.

However, it is imperative to acknowledge that, while these algorithms provide predictive models, their interpretations must be contextualized within the broader socio-economic landscape of Egypt. The study establishes that the gradient-boosting (GB) algorithm outperforms the random forest (RF) algorithm. The GB algorithm reveals that the unemployment rate in Egypt is primarily influenced by the added value of the industrial sector, followed by GDP growth, country imports, and labor force participation. Additionally, it demonstrates a negative relationship between the unemployment rate and variables such as agriculture sector value added, foreign direct investment (FDI), GDP growth, and gross fixed capital formation. Conversely, a positive relationship is observed with variables including inflation, human development index (HDI), imports, labor force participation, industry sector value added, and services sector value added.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Conflict of interest

The author declare no conflict of interest.

References

- Imran M, Sial MH (2013) Employment, Trade Openness, and Capital Formation : Time Series Evidence from Pakistan. *J Glob Sci Issues* 1: 37–42.
- Sattar A, Bhalli MN (2013) Empirical Evidence from Pakistan. 51: 191–207.
- Abdul-khaliq S, Soufan T, Shihab RA (2014) The Relationship between Unemployment and Economic Growth Rate in Arab Country. *J Econ Sust Dev* 4: 62–67.
- Ogonna IC, Idenyi OS, Ifeyinwa AC, et al. (2016) The Implications of Rising Public Debt on Unemployment in Nigeria : An Auto Regressive Distributed Lag Approach. *Asian Res J Arts Soc Sci* 1: 1–15. <https://doi.org/10.9734/ARJASS/2016/26394>
- Bakare AS, Akoko PMBA, State O (2011) The Determinants of Urban Unemployment Crisis in Nigeria : An Econometric Analysis. *J Emerg Trends Econ Manage Sci* 2: 184–192. <https://hdl.handle.net/10520/EJC134174>
- Megbowon ET, Ngarava S, Mushunje A (2016) Foreign direct investment inflow, capital formation and employment in South Africa: A time series analysis. *Int J Econ Financ Stud* 8: 175–189.
- Qadar M, Muhammad N (2013) Unemployment Is An Outcome of Corruption. *Int J Adv Res Econ Commer* 2: 14.
- Azhar AAM, Ibrahim NA (2021) Factors Affecting Unemployment Rate In Malaysia. Available from: <https://oarep.usim.edu.my/handle/123456789/18552>.
- Siddiqa A (2021) Determinants of unemployment in selected developing countries: A panel data analysis. *J Econ Impact* 3: 19–26.

- Digvijay DB (2021) Factors affecting high unemployment in India. Available from: <https://mpira.ub.uni-muenchen.de/id/eprint/110621>.
- Gogas P, Papadimitriou T, Sofianos E (2022) Forecasting unemployment in the euro area with machine learning. *J Forecast* 41: 551–566. <https://doi.org/10.1002/for.2824>
- Dalmar MS, Ali AYS, Ali AA (2017) Factors affecting unemployment in Somalia. *J Econ Sustain Dev* 8: 200–210.
- Abugamea G (2018) Determinants of unemployment: Empirical evidence from Palestine. Available from: <https://mpira.ub.uni-muenchen.de/id/eprint/89424>.
- Salama A, Judit O (2019) Key factors ampact unemployment in the Arab world—central European. *J Labour Law Pers Manage* 2: 1. <https://doi.org/10.33382/cejllpm.2019.03.05>
- McCann P (2001) *Urban and Regional Economics*. New York: Oxford University Press.
- Turgut E (2021) Unemployment and factors affecting unemployment in developing countries. *Discuss Between* 7.
- Bowles S, Gintis H (1975) The Problem with Human Capital Theory Marxian Critique. *Am Econ Rev* 65: 74–82. <https://doi.org/10.2307/1818836>
- Friedman JH (2001) Greedy function approximation: a gradient boosting machine. *Ann stat* 1189–1232. Available from: <https://www.jstor.org/stable/2699986>.
- Hastie T, Tibshirani R, Friedman JH, et al. (2009) *The elements of statistical learning: data mining, inference, and prediction*. 2: 1–758. New York: Springer.
- Breiman L (2001) Random forests. *Mach learn* 45: 5–32. <https://doi.org/10.1023/A:1010933404324>
- Dietterich TG (2000) Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, 1–15. Springer Berlin Heidelberg.
- Tashman LJ (2000) Out-of-sample tests of forecasting accuracy: an analysis and review. *Int J Forecasting* 16: 437–450. [https://doi.org/10.1016/S0169-2070\(00\)00065-0](https://doi.org/10.1016/S0169-2070(00)00065-0)
- El-Aal A, Mohamed F (2024) The relationship between CO₂ emissions and macroeconomics indicators in low and high-income countries: using artificial intelligence. *Environ Dev Sustaina* 2024: 1–22. <https://doi.org/10.1007/s10668-024-04880-3>
- Probst P, Wright MN, Boulesteix AL (2019) Hyperparameters and tuning strategies for random forest. *Wires Data Min knowl* 9: e1301. <https://doi.org/10.1002/widm.1301>
- Abd El-Aal MF (2023) Analysis factors affecting Egyptian inflation based on machine learning algorithms. *Data Sci Financ Econ* 3: 285–304. <https://doi.org/10.3934/DSFE.2023017>
- Molinaro AM, Simon R, Pfeiffer RM (2005) Prediction error estimation: a comparison of resampling methods. *Bioinformatics* 21: 3301–3307.
- World Bank, world development indicators. Available from: <https://databank.worldbank.org/source/world-development-indicators>.



AIMS Press

©2024 the Author(s), licensee AIMS Press. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0>)