An Empirical Comparison of Machine Learning Models for Early Prediction of Unemployment Rates Across Countries

Mr. Joshua[a] (Researcher)

ARTICLE INFO

Keywords:
unemployment forecasting
machine learning
XGBoost
economic indicators
labour market dynamics
predictive modelling

ABSTRACT

Unemployment remains a critical challenge for economic development and social stability worldwide, particularly in regions experiencing rapid population growth and economic volatility, such as Africa. The unpredictability of unemployment rates during crises and structural shifts poses significant obstacles to sustainable growth and effective policy planning. Given the proven advantages of data driven technologies, machine learning offers a powerful approach to improve forecasting accuracy beyond traditional econometric methods. This study aims to empirically compare machine learning models for predicting national unemployment rates one year ahead across diverse economic contexts in the world. Using annual World Bank data (SL.UEM.TOTL.ZS) from 1991 to 2024 covering 266 countries and regional aggregates, analysis incorporates lagged unemployment rates, temporal feature and country specific identifiers after data cleaning, outlier treatment and feature engineering. Four supervised models: Linear Regression (baseline), Random Forest, XGBoost and Elastic Net, were trained on data from 1991–2019 and evaluated on the high volatility test period 2020–2024, which included the impacts of 2008 financial crisis and the COVID-19 pandemic. Performance was assessed using key metrics, Root Mean Squared Error (RMSE), R^2, Mean Absolute Percentage Error (MAPE) and Mean Absolute Scaled Error (MASE). The results demonstrate that Elastic Net achieved the strongest performance with an RMSE of 1.0419 and R^2 of 0.9442, outperforming XGBoost (RMSE = 1.0571, R^2 = 0.9426), Linear Regression (RMSE = 1.0761, R^2 = 0.9405) and Random Forest (RMSE = 1.1421, R^2 = 0.9330) due to its superior handling of regularization, feature selection and multicollinearity in high dimensional settings. Exploratory analysis further revealed persistent regional disparities, with Sub-Saharan Africa and the Middle East and North Africa showing higher and more volatile unemployment rates compared to Europe and East Asia. These findings highlight the effectiveness of penalized linear models in providing reliable early warning signals and their potential to guide proactive labour market policies, optimize resource allocation and support evidence based decision making in both developed and developing economies.

1. Introduction

Employment plays a critical role in economic development and social stability in Africa.High unemployment not only fuels poverty and inequality, but also hampers sustainable economic growth,social cohesion, and human capital development Wu (2023).Globally, predictive analytics have become an essential tool for economic policy makers to anticipate labour market trends and implement timely interventions Monir Aljinbaz and Al Rahhal (2024).Africa, with its rapidly growing working age population growing with more than 2% annually, faced a pressing need for proactive labour market policies that are informed by early warning systems and data driven insights Opondo, Bundi and Weke (2024).

Despite its importance, unemployment prediction in Africa remains a challenge.Many countries experience high and volatile unemployment rates, with youth unemployment often exceeding 8.9%, yet there is a limited mechanism existed to predict these trends in reliably Mulaudzi (2021).Traditional models predominantly relied on lagging indicators and classical econometric approaches, which often lacked sufficient accuracy for early intervention Katris (2020).African labour markets are highly heterogeneous, with informal sectors, varying data quality and diverse economic structures, making accurate forecasting difficult Yurtsever (2023a).Further more there was lack of comparative studies examining which machine learning model approaches work best for African contexts.This creates a gap between the advanced predictive modelling utilized in developed economies and the underutilisation of such methods in Africa Güler, Kabakçı, Koç, Eraslan, Derin, Güler, Ünlü, Türkan and Namlı (2024).

This study aimed to systematically compare multiple learning models for their effectiveness in predicting unemployment rate 1 year in advance across diverse countries.The primary objective was to identify model(s) that provided reliable and actionable forecasts, calibrated for the challenges of African labour markets, such as data scarcity

and informal sector dynamics Estrada-Moreno, Rendon-Lara and Jiménez-Núñez (2024).The novel contribution lies in providing comprehensive empirical comparison of machine learning (ML) models tailored to African contexts, assessing not only predictive accuracy but also practical utility for policymaking.Expected outcomes included the identification of the most effective machine learning (ML) approaches, enabling governments and institutions to transition from reactive to proactive labour market policies Ibrahim, Umar, Bichi, Ahmad, Rabiu and Ahmad.

Consistent with findings in related predictive analytics studies, which highlight the importance of model complexity, feature interaction handling and regularization Güler et al. (2024) Mero, Salgado, Meza, Pacheco-Delgado and Ventura (2024), the results revealed notable variation across model performances. Linear Regression achieved an RMSE of 1.0761 and R^2 of 0.9405,indicating a strong baseline fit but limited ability to capture non linear labour market dynamics. Random Forest produced an RMSE of 1.1421 and R^2 of 0.9330, reflecting robust performance in handling heterogeneous predictors but the lowest precision among the models. XGBoost delivered competitive results with an RMSE of 1.0571 and R^2 of 0.9426, demonstrating effectiveness in modelling complex interactions and non-linear patterns. Elastic Net outperformed all others with an RMSE of 1.0419 and R^2 of 0.9442, showcasing its strength in bridging linear baselines with ensemble methods through regularization and automatic feature selection, particularly useful for datasets with multicollinearity from regional and temporal features.

2. Literature Review

Over the past few years there has been an interest among researchers in applying machine learning and deep learning techniques in forecasting across different applications.Different regions use different methodological approaches in predicting unemployment rate.According to Madaras (2024) Katris (2020) Celbiş (2023) central Europe and parts of Africa have employed ARIMA,ANN,hybrid model,time series,classification models and Deep Learning techniques.While in regions such as the USA they use econometric approaches such as Survey Professional Forecasters(SPF).Additionally they also use machine learning model,lasso regression,naive forecast and neural networks Kreiner and Duca (2020).

Traditional time series models, such as ARIMA,have been the predominant approach in early unemployment forecasting literature,primarily due to their capacity to capture temporal dependencies while maintaining relatively modest computational requirements.Isqeel Adesegun, Mathew and Omotola (2020) applied ARIMA to model Nigeria's unemployment rate from 1991 to 2018.Their study achieved stationarity through differencing techniques and validated model adequacy using the Akaike Information Criterion(AIC) and Shapiro-Wilk tests.While their findings demonstrated ARIMA's effectiveness in capturing linear trends ,the research also revealed limitations when addressing non-stationary data patterns.In a comparative analysis, Sharaf (2024) examined the performance of ARIMA against ARDL and VAR models for forecasting Jordan's youth unemployment across the 1991-2022 period.The results indicated that ARDL demonstrated superior performance based on the mean absolute Error(MAE) and the root mean square error (RMSE) metrics, highlighting the advantage of multivariate extension over univariate ARIMA models, particularly when incorporating economic indicators such as GDP and foreign direct investment (FDI).A study conducted by Aris, Nagaratnam, Zakaria, Azami, Samsudin and Othman (2024) using SARIMA has shown to outperform standard ARIMA when addressing cyclical patterns.

Deep learning models, like Long Short Term Memory (LSTM) and Gated Recurrent Unit (GRU) architectures, have gained attention due to their demonstrated capability to capture non-linear patterns within unemployment data.Yurtsever (2023b) developed a hybrid LSTM-GRU model used for forecasting unemployment rates across the United States, United Kingdom, France and Italy, utilizing data spanning from 1983 to 2022.The proposed hybrid approach achieved lower RMSE and MAE compared to stand-alone implementations of either architecture.However, the study acknowledged that such models require a substantial dataset to mitigate the risk of over-fitting.Madaras (2024) compared CNN,MLP and Random forest with ARIMA for Central European countries' monthly rate, where deep models excelled in Mean Absolute Percentage Error (MAPE) and Mean Squared Error (MSE).Nevertheless, the computational intensity of these models presented practical limitations for real-time forecasting applications.These datasets often include monthly or quarterly economic series, with similarities in using error metrics for validation but difference in architecture.For instance GRU's simplicity aids faster training compared to LSTM Tufaner and Sözen (2021).Strength in deep learning approaches lies in their predictive accuracy during periods of economic crises such as COVID-19 pandemic.While limitations include opacity and data scarcity in developing regions.

Hybrid models, which integrated traditional econometric and machine learning techniques have demonstrated better performance mitigating individual limitations of stand-alone methods.Katris (2020) study developed a hybrid

framework combining ARIMA with Artificial Neural Networks (ANN) and Support Vector Machines (SVM) to predict unemployment rates in Nigeria.Their results indicated that ARIMA-ANN hybrid achieved the lowest MAE and RMSE values,highlighting synergy in decomposing linear and non-linear components.Shrief, Taha, Elstohy, Nagy and Ali (2025) optimized SVM models using Grey Wolf Optimizer (GWO), Whale Optimization Algorithm (WOA) and Moth-Flame Optimization (MFO), for predicting Egypt's female unemployment from 2018 to 2023, reducing MSE BY 3.8% via polynomial kernels, emphasizing optimization's role in parameter tuning but noting computational overhead.While these studies share methodological similarities including use of quarterly data and standard error metrics such as R^2 for model evaluation, difference emerged in their focus Manasa, Kalidas et al. (2022) employed SVM, random forest and gradient boosting for state-wise Indian prediction excelling in prediction accuracy.Hybrid models have are robust in data noise and improved generalization,however model complexity can complicate implementation and interpretation.

Multivariate approaches represent an extension of univariate models by incorporating external economic variables such as GDP and foreign direct investment (FDI) enhancing explanatory power.Sam, Manene, Kipchirchir and Pokhariyal (2020) used co-integration analysis to examine youth unemployment dynamics in Kenya, revealing that both GDP and FDI exhibit positive correlations with unemployment rates.Their methodology utilized the Augmented Dickey Fuller (ADF) and Johansen tests,with strengths in long-run relationship and limitation in short-term forecasts.Kreiner and Duca (2020) applied machine learning on US economic data which outperformed Vector Autoregression (VAR) linearity assumption as key difference.Dataset vary from annual Sharaf (2024) to quarterly Mulaudzi (2021) with a common metric like MAE, similarities in economic variable section, but difference in scope Celbiş (2023) used tree-based machine learning for rural Europe, using SHARP(SHapley Additive exPlanations) values to identify the causal importance of vocational training access and other regional factors determining unemployment outcome. Multivariate approach lie in the ability to provide causal insights and capture complex interdependencies between unemployment and broader economic conditions.However, these methods are constrained on data stationarity applicable to regional disparities.

Machine Learning excels in predicting long-term unemployment risks, focusing on interpretability and bias.Zhao (2020) used logistic regression, random forest and XGBoost for European data, achieving 81.1% accuracy with SHAP explanation of model predictions, addressing biases in age and immigration.Celbiş (2023) applied classification trees and boosting in rural Europe, highlighting subgroup inequalities via SHAP with strength in feature importance but limitation on small data set.These study share methodological similarities on ensemble learning approaches and their use of standard classification metrics such as accuracy and precision for performance evaluation.Difference emerge in their specific applications for instance Mutascu (2021) adopted a broader perspective by exploring the global impact of artificial intelligence adoption on unemployment rates using econometric panel data method.The macro-level analysis contrast with the micro-level focus of Zhao (2020) and Celbiş (2023).

Recent studies examining the impact of external shocks, like COVID-19 pandemic, has demonstrated the value of integrating novel data sources for robust forecast. Tufaner and Sözen (2021) conducted a comparative analysis of ARIMA and ANN models in the Turkish context during, they the post-COVID period,finding that ANN model exhibited superior performance in term of MAE, largely due to their capacity to capture non-linear relationships within the data.Building upon this Simionescu and Cifuentes-Faura (2022) incorporated Google Trends data into Bayesian Vector Autoregression (VAR) framework to forecast youth unemployment rates in Spain over the period of 2004 to 2021,achieving improvements in predictive accuracy compared to conventional modelling approaches.While studies share commonalities in their use of quarterly data and error metrics for model evaluation, they diverge in their selection of input variables and methodological frameworks. Aris et al. (2024) extended this line of inquiry by using Gaussian Process Machine Learning (GPML) techniques into their analysis of Malaysian unemployment series from 1991 to 2022, demonstrating that GPML outperformed Seasonal ARIMA (SARIMA) models.

Emerging techniques like GPML and deep multivariate models address non-linearity in economic data. Aris et al. (2024) compared GPML with ARIMA/SARIMA,GPML yielding lowest MAE. Mulaudzi (2021) used LSTM/GRU on South African multivariate data, outperforming VAR by 20x in error,while Prihandi, Wijono, Sembiring and Maria (2025) applied ARIMA to non-unemployment(chili)data,showing transferability.

With the recent research studies on unemployment forecasting reveals a clear methodological shift from traditional linear time series models toward more advanced,data-driven approaches,reflecting the growing complex of labour market dynamics.Although ARIMA models continue to provide a strong base line for capturing linear temporal patterns, empirical evidence indicates that deep learning architectures such as Long Short-Term Memory(LSTM) and Gated Recurrent Unit (GRU) networks are more robust when modelling non-linear relationships, during periods of economic crisis. Building on the strength of both paradigms, hybrid frameworks integrate econometric and machine

learning techniques such as ARIMA-ANN models, have demonstrated improved forecasting accuracy through reduced error rates.Additionally,multivariate models that utilizes macroeconomic indicators, including GDP and foreign direct investment, offer enhanced explanatory capacity even when faced with data availability challenges and stationarity assumptions.Recent methodological innovations have also expanded the scope of unemployment forecasting by utilizing alternative data sources, such as Google Trends and by applying Gaussian Process-based machine learning models,which have outperformed conventional SARIMA benchmarks in several studies.Finally, the growing use of ensemble learning methods alongside tools such as SHAP values underscores a broader research trend towards not improving performance but also enhancing transparency, interpretability and fairness in unemployment risk assessment.

3. Methodology

3.0.1. Data Source and Characteristics

The empirical analysis is based on a panel dataset of annual unemployment rates,measured in percentage of a total labour force covering 266 entities, 193 sovereign countries and 73 regional aggregates over a period of 1991 to 2024.The dataset was obtained from the World Bank's Development Indicators(WDI) database,specifically indicator SL.UEM.TOTL.ZS,which is derived from modelled estimates produced by the International Labour Organization(ILO).

This data source was selected for its extensive global coverage,methodological consistency and adherence to internationally recognized labour statistics standards.While national unemployment figures may vary in definition and survey methodology, the ILO's harmonization procedures enhance cross country comparability, making the dataset suitable for global panel analysis.The dataset was downloaded in CSV format from the World Bank's open data portal in July 2025.

Structurally, the dataset comprises of four variables Country Name,Country Code,Indicator Name and Indicator Code alongside 34 annual unemployment rate observation corresponding to the study period.As a result, the raw dataset is organized in a wide format with 266 rows(Countries and regional aggregates) and 38 columns, of which four contain identifying metadata and the remaining columns represent yearly observations.

3.0.2. Data Preparation and Cleaning

To facilitate time series forecasting and machine learning estimation, the raw wide format dataset was transformed into a panel long format, with year serving as the temporal index and Country Code as the cross sectional identifier.Data ingestion and initial diagnostics were conducted using pandas, confirming variables types and dimensional consistency.

Missing observations were addressed using imputation strategy to preserve temporal structure.Forward and backward filling were applied to contiguous gaps, while isolated missing values were imputed using the series mean.This approach minimized data loss while avoiding the introduction of artificial volatility.Outliers were identified using Isolation Forest Algorithm implemented in scikitlearn(contamination rate set to 0.01).While extreme unemployment values often correspond to genuine economic shocks, unchecked outliers can disproportionately influence model estimation.Consequently, detected outliers were replaced using linear interpolation rather than being removed, mitigating undue leverage while retaining information about crisis related dynamics.

Lagged unemployment rates (t-1 to t-3) were constructed to account for strong autocorrelation, while one hot encoded regional dummies captured geographic heterogeneity.All numericals features were standardized using scikitlearn's StandardScaler to ensure algorithmic stability.The data were reshaped into long format, with 'Year' as the time index and 'Country Code' as the entity identifier.

A train test split was enforced, 1991 to 2019 (approximately 80%) for model training and validation and 2020 to 2024(approximately 20%) held out for testing.This temporal partitioning prevents bias and enables model performance to be evaluated under conditions of exceptional economic volatility,during the COVID-19 period, thereby enhancing the external validity and practical relevance of the resulting forecasts.

3.0.3. Exploratory Data Analysis

Exploratory Data Analysis(EDA) was conducted to uncover patterns,correlations and anomalies, informing model selection.Summary statistics was computed using pandas (df.describe()),revealing regional disparities for example Africa Eastern and Southern had a mean of 8% with high variance, while Europe and Central Asia averaged 5% with lower volatility.Distribution was visualized using histograms and kernel desnity plots using matplotlib and seaborn, showing right skewness and heavy tails in crisis years, indicating non-normality and the need for non-linear

models.Correlations were quantified using Pearson coefficients ans heatmaps (sns.heatmap()), demonstrating strong autocorrelation(r>0.8 for consecutive years) and inter-regional clusters (positive correlation among African entities).

Trends were explored through time series line plots (plt.plot()),highlighting global increases such as 2% rise in 2008,3% surge in 2020 and persistent disparities using box plots (sns.boxplot()).Principal Component Analysis (PCA) from scikit-learn reduced dimensionality , with the first three components explaining 85% of variance, loading on crisis periods and regional factors.Anomalies, such as data gaps in micro states, were flagged with scatter plots (plt.scatter()), guiding imputation.These insights confirmed non-stationarity, the value of lagged features for dynamics and suitability of ensemble models for handling volatility.

3.0.4. Machine Learning Modelling

Model selection focused on on year ahead forecasting, prioritizing algorithms for time series regression with non-linearities.The three core models that were implemented, Linear Regression(baseline),Rnadom Forest(ensemble bagging) and XGBoost(gradient boosting),there was an additional ElasticNet model was explored for regularization in high dimensional settings.

Linear Regression model serves as the foundational baseline in this study relying on the classic ordinary least squares framework.Here, y_{t+1} represents the forecasted unemployment rate one year ahead, β_0 is the intercept capturing baseline levels, the β_1 coefficients quantify the linear impact of each predictor X_i lagged unemployment rates and regional dummies, and ϵ denotes the random error term assumed to be independently and identically distributed with a mean of zero.The purpose of this formula is to establish a simple, interpretable relationship assuming linearity and independence among predictors, allowing for a straightforward estimation of how historical lags and geographical factors directly influence future rates.In the context of unemployment forecasting , it provides a benchmark against which more complex models can be compared, highlighting the limitations of linear assumptions when dealing with non-linear shocks like economic crises.

$$y_{t+1} = \beta_0 + \sum_{i=1}^{k} \beta_i X_i + \epsilon \tag{I}$$

Random Forest model adopts an ensemble approach through bagging, where $\hat{y}$ is the predicted unemployment rate, B set to 100 in this implementation is the number of decision trees and each $T_b(x)$ represents the prediction from an individual tree trained on bootstrapped subsets of the data with random feature selection at each split.The purpose of this averaging formula is to reduce the variance and over fitting inherent in single decision trees by aggregating diverse predictions, thereby improving robustness to interactions and non-linear patterns without assuming linearity.For unemployment time series, this enables the model to capture complex relationships such threshold effects in lagged variables or interactions between regional dummies while maintaining resistance to noise, making it particularly suitable for heterogeneous panel data where traditional parametric forms may fail.

$$\hat{y} = \frac{1}{B} \sum_{b=1}^{B} T_b(x) \tag{II}$$

XGBoost model advances this ensemble through gradient boosting, where $\hat{y}$ is the final prediction, K is the number of sequential weak learners(regression trees), and each $f_k(x)$ is a function added iteratively to correct the residuals of previous predictions while minimizing a regularized objective(mean squared error and penalties for tree complexity).This additive formulation builds predictive power incrementally, focusing on hard to predict observations and regularization to prevent over fitting.In the unemployment forecasting application, the formula's strength lies in its ability to model non-linear shock and temporal dependencies more effectively than bagging alone.

$$\hat{y} = \sum_{k=1}^{K} f_k(x) \tag{III}$$

ElasticNet model extends the linear framework with regularization, conceptually similar to Linear Regression but augmented by a combined L1 and L2 penalty term in its objective function balancing lasso's feature selection with

ridge's coefficient shrinkage.Although not expressed with a standalone predictive formula here, it modifies the core linear equation by shrinking the coefficients during estimation.The purpose of this regularization is to enhance stability and interpretability in high dimensional settings such that numerous regional dummies introduce multicollinearity by performing automatic variable selection and mitigating over fitting.In this study, ElasticNet was employed to text whether a penalized linear approach could bridge the gap between the simple baseline and fully non-linear ensembles, providing insights into feature relevance while addressing potential instability in the lagged and dummy variables.

3.0.5. Deployment

Model serialization was achieved using joblib, a library optimized for efficient handling of large NumPy arrays commonly produced by scikit-learn and XGBoost models.Following hyperparameter tuning and final training on the full 1991 to 2019 dataset, the XGBoost regressor along with the associated preprocessing pipeline,including the StandardScaler and feature encoders was saved to disk with the command joblib.dump(model, 'xgboost_unemployment_predictor.joblib'). approach bundles the entire fitted pipeline into a single portable file,preserving the exact state of the model, including learning parameters, tree structures and preprocessing steps.Secondary models, linear regression random forest and elasticnet were similarly serialized for comparative purposes,enabling ensemble inference if required.Loading the model using joblib.load(), ensuring that the deployment system can restore the exact model without retraining.

A dedicated Python inference was developed to handle prediction requests.This script first loads the serialized model and pipeline,then defines a predict_unemployment(country_code,year,lags=[lags1,lags2,lags3], region_dummies) function that accepts input features in the same format as the training data, a country or region identifier, the current year, lagged unemployment rates (t-1 to t-3) and any required regional indicators.The input is transformed using the saved pipeline scaling and encoding, passed to the model for one year ahead forecasting and returned as a single predictor rate along with optional confidence intervals derived from the model's internal approximations.This modular structure separates data processing, prediction and output formatting, promoting maintainability and testing.

During environment set up, a requirement.txt file was generated listing exact library version used during development including:pandas==2.0.3,scikit-learn==1.3.0,xgboost==2.0.3, allowing replication using pip install -r requirements.txt within a virtual environment created with venv. To eliminate platform dependencies and ensure consistent execution across operating systems,the entire application including the inference script, serialized model and dependencies.

3.1. Exploratory Data Analysis

3.1.1. Temporal Trends in Global Unemployment

Analysis of global mean unemployment rates revealed distinct temporal patterns across the 34-year observation period. The initial phase (1991-2000) exhibited an upward trajectory, with the global average unemployment rate increasing from 7.2% in 1991 to a peak of 8.2% by the late 1990s and early 2000s. This was followed by a stabilization and gradual decline period (2000-2008), during which unemployment rates decreased to approximately 7.0% by 2008, representing a sustained improvement in global labour market conditions. Two major economic disruptions were clearly identifiable in the temporal pattern. The 2008-2010 period demonstrated a pronounced spike coinciding with the global financial crisis, followed by a recovery phase characterized by declining rates through 2019, when unemployment reached its lowest point in the observation period. The most dramatic disruption occurred in 2020, corresponding to the COVID-19 pandemic, which produced a sharp increase in global unemployment. Subsequent years (2021-2024) showed gradual recovery, though unemployment rates remained elevated relative to pre-pandemic levels.

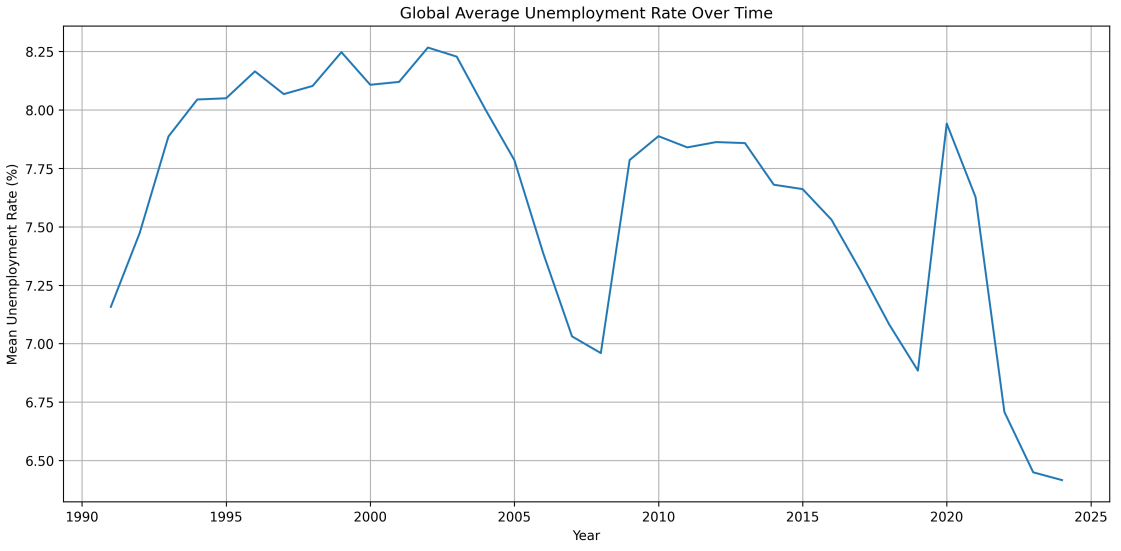

Figure 1: This line chart illustrates the global average unemployment rate from 1990 to 2025, showing a general downward trend interrupted by spikes in 1993 (around 8.2%), 2009 (financial crisis peak at 7.5%) and a sharp rise to 8.0% in 2020 due to COVID-19 followed by recovery. The chart uses a blue line with year labels on the x-axis and rate percentages on the y-axis, highlighting how economic shocks create volatility in otherwise stable periods

3.1.2. Distribution Characteristics and Variability

Histogram analysis of unemployment rate distributions revealed a right-skewed pattern across most years, with the majority of observations concentrated between 3% and 12%. The distribution exhibited a long right tail, indicating the presence of countries with substantially elevated unemployment rates exceeding 20%. Boxplot analysis by year demonstrated that median unemployment rates remained relatively stable across non-crisis periods, typically ranging between 6% and 8%. However, the interquartile range (IQR) varied considerably across years, with crisis periods (2008-2010, 2020) exhibiting substantially wider spreads, indicating increased heterogeneity in country-level responses to economic shocks. Outlier analysis identified persistent extreme values across multiple years. Notably, certain countries consistently appeared as upper outliers (unemployment rates exceeding 20%), while others maintained persistently low rates (below 3%). This pattern suggests structural differences in labour market characteristics across countries rather than merely transient economic conditions.

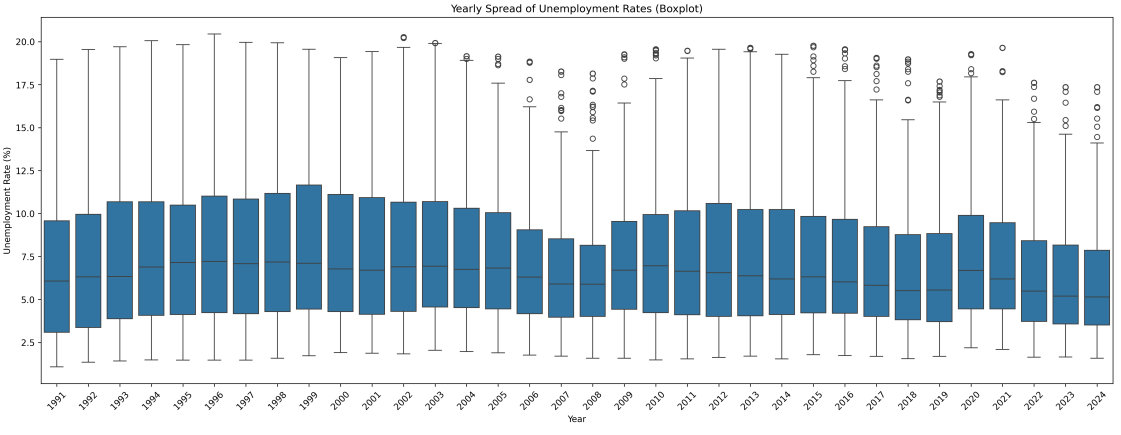

Figure 2: The histogram presents a right skewed distribution of unemployment rates across countries, with bars showing frequency concentrations between 3% and 12% and a long tail extending to 35% for high-unemployment outliers. It uses blue bars on a horizontal rate scale, demonstrating persistent structural differences where most countries cluster low but crises widen the spread, as seen in increased bar heights for rates above 20% during volatile years.

3.1.3. Regional Patterns: East African Case Study

Detailed examination of East African unemployment trajectories revealed substantial inter-country variation within the region. Kenya, Tanzania and Uganda demonstrated relatively stable unemployment patterns, with rates consistently maintained within a 2-8% band throughout the 34-year period. Ethiopia exhibited a distinctive long-term declining trend, suggesting gradual but sustained improvement in labour market conditions and employment opportunities. Conversely, Burundi displayed considerable volatility, with multiple sharp fluctuations likely attributable to periods of political instability and civil conflict.

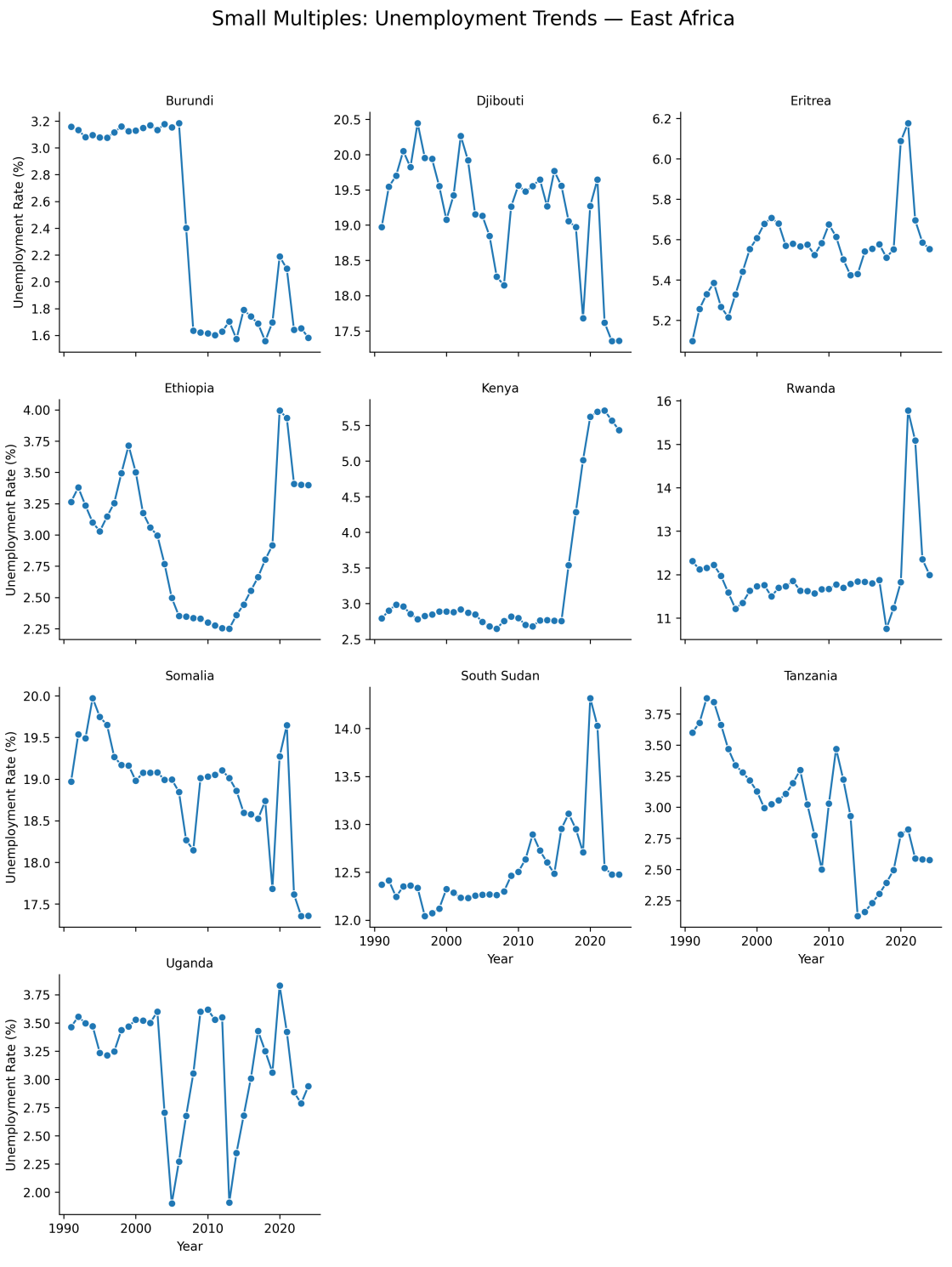

Figure 3: Line charts showing trends for Burundi, Djibouti, Eritrea, Ethiopia, Kenya, Rwanda, Somalia, South Sudan, Tanzania, Uganda over 1990-2025, with y-axes as unemployment rates.

3.1.4. Cross-Country Comparative Analysis (2024)

Analysis of 2024 unemployment rates revealed stark disparities in current labour market conditions across countries. The 30 countries with highest unemployment rates predominantly included nations experiencing post-conflict recovery, economic transition, or structural labour market challenges, with rates ranging from 20% to over 30%. In contrast, the bottom 30 countries—characterized by unemployment rates below 4%—were primarily composed of developed economies with mature labour markets and emerging economies with high informal sector absorption. This dichotomy underscores persistent global inequality in employment opportunities and labour market efficiency.

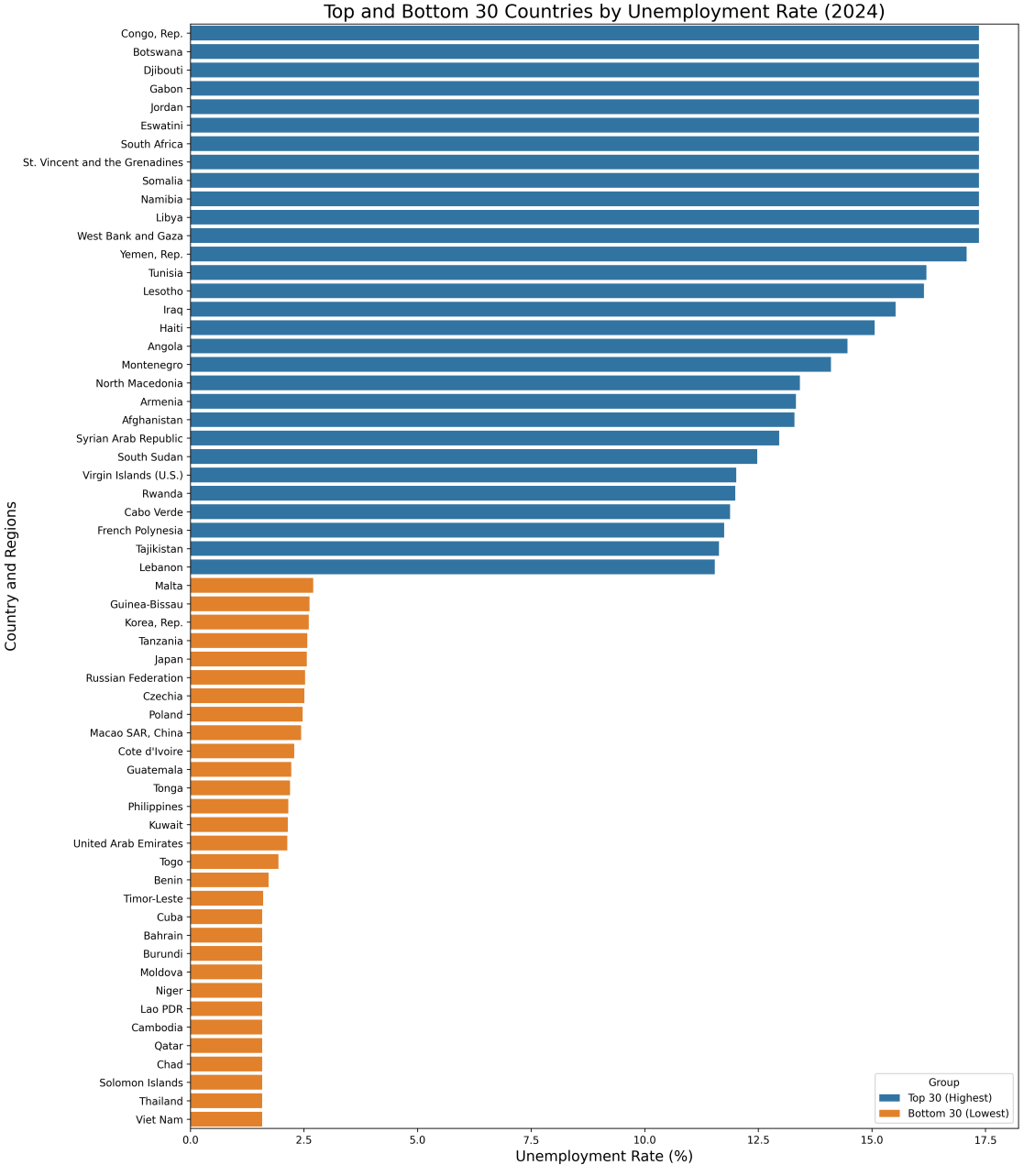

Figure 4: This horizontal bar chart compares the 30 countries with the highest (blue bars) and lowest (orange bar) unemployment rates in 2024, with country names on the y-axis and rates on the x-axis. The visual dichotomy underscores global inequalities, showing high rates in transitioning economies versus low rates in developed or resource rich nations.

3.1.5. Temporal Autocorrelation Patterns

Correlation matrix analysis of unemployment rates across years revealed strong positive correlations between adjacent temporal periods, with correlation coefficients exceeding 0.90 for consecutive years. Specifically, the correlation between 2019 and 2020 unemployment rates was 0.97, while the 2008-2009 correlation measured 0.95. This pattern indicates substantial persistence in unemployment levels, whereby countries' labour market positions exhibit strong year to year stability. However, correlations diminished progressively with increasing temporal distance, declining to approximately 0.60-0.70 for periods separated by five or more years. This attenuation suggests that while short-term unemployment persistence is strong, longer-term structural changes, policy interventions, and economic development trajectories gradually alter countries' relative unemployment positions.

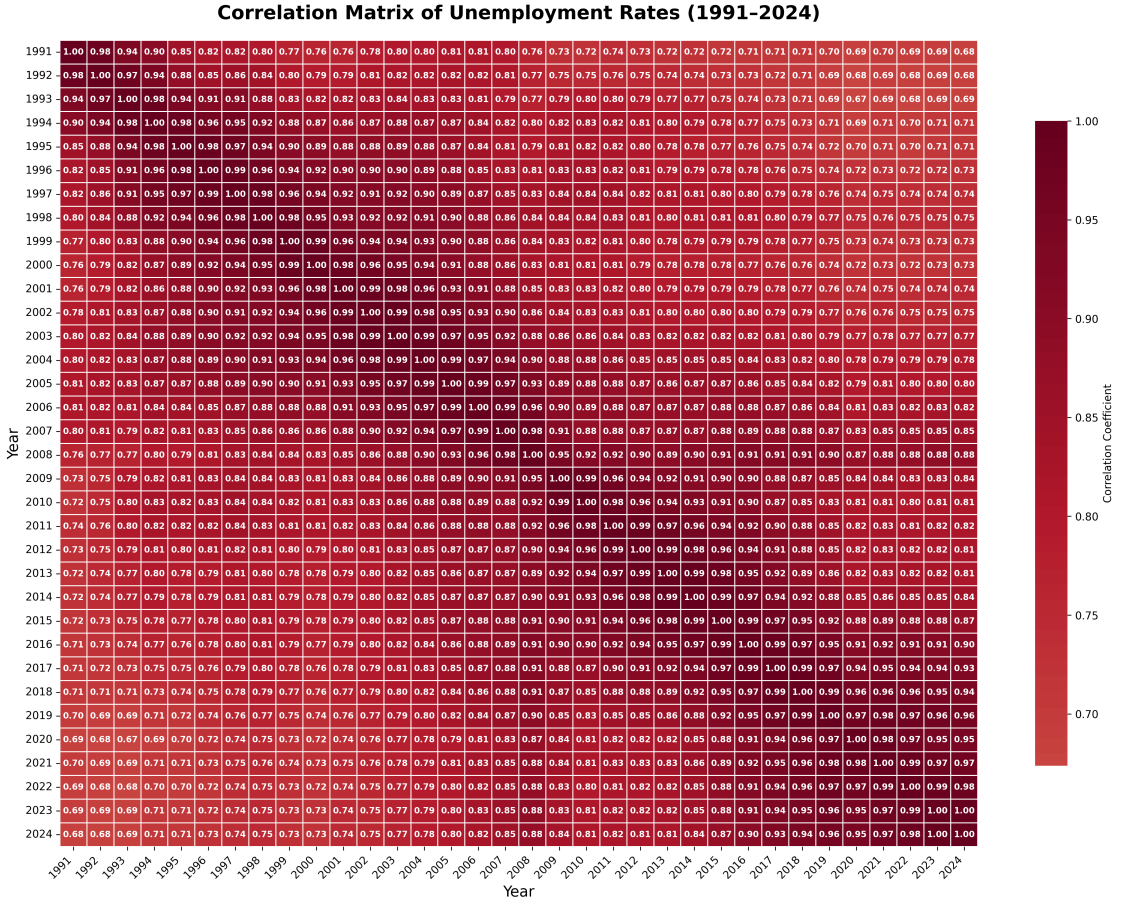

Figure 5: The heat map displays correlations between annual unemployment rates from 1991 to 2024, with dark red squares (values 1.0) along the diagonal for adjacent years fading to lighter reds (0.6) for distant periods. Years label both axes, illustrating strong short term persistence but weakening long term correlations, which suggests structural changes gradually reduce year to year stability in global labour markets.

3.1.6. Feature Importance Analysis

Random Forest based feature importance analysis, examining the predictive value of lagged unemployment rates (2019-2023) for forecasting 2024 values, revealed a highly skewed importance distribution. The immediate prior year (2023) dominated feature importance with 84% contribution, substantially exceeding all other temporal lags. The second lag (2022) contributed only 8%, while lags three, four, and five (2021, 2020, and 2019 respectively) collectively accounted for less than 8% of predictive power. This steep gradient in feature importance confirms strong first order autocorrelation in unemployment time series and suggests diminishing marginal informational value of increasingly distant historical observations.

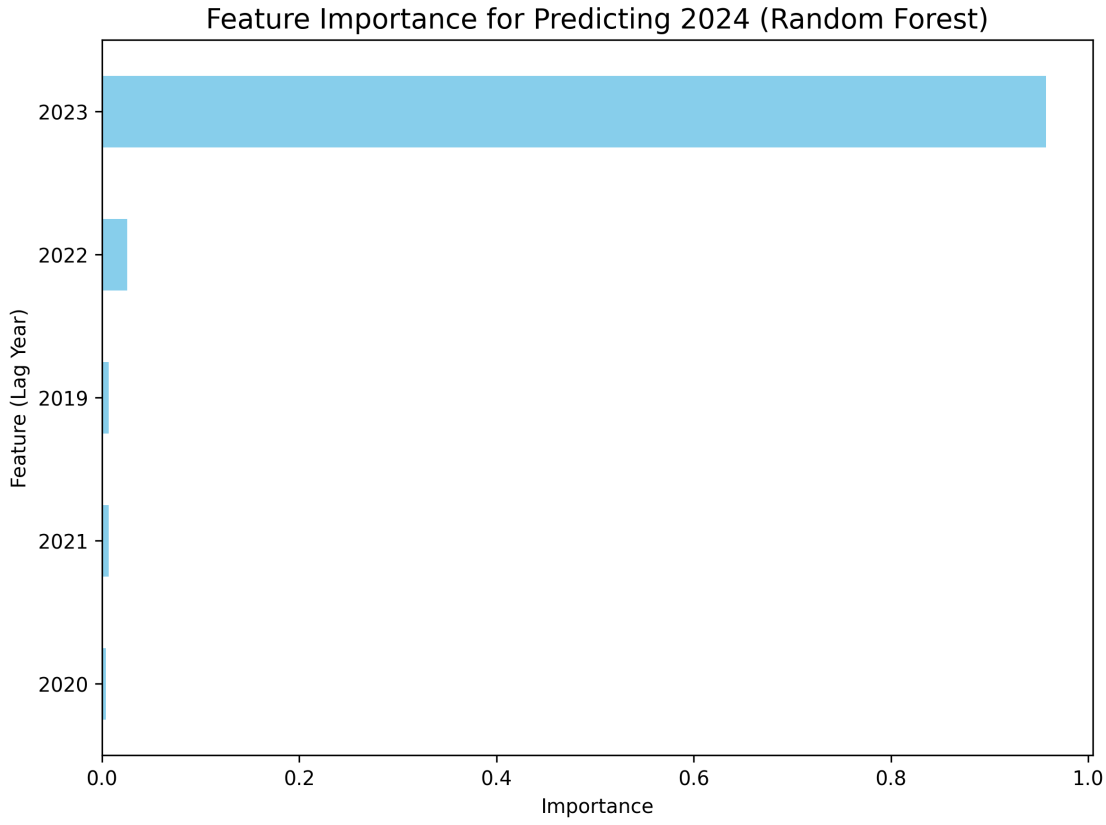

Figure 6: This vertical bar chart shows feature importance scores for lagged years in predicting 2024 unemployment, with 2023 dominating at 0.84 (tall blue bar) and earlier lags (2022 at 0.08, down to 2019 near 0) as shorter bars. The y-axis is importance (0-1), x-axis years.

3.1.7. Distributional Evolution Over Time

Ridge-style density plot analysis comparing unemployment distributions across six key years (1991, 2000, 2008, 2010, 2020, 2024) illuminated secular trends in global unemployment patterns. The 1991 and 2000 distributions exhibited narrow, left-skewed profiles concentrated below 10%, indicating relatively homogeneous unemployment experiences across countries during these periods. The 2008 and 2010 distributions displayed markedly wider spreads with increased density in mid to high unemployment ranges, reflecting the heterogeneous impact of the global financial crisis across different economies. The 2020 distribution showed the most pronounced rightward shift and increased dispersion, consistent with the pandemic's severe and varied impact across countries. By 2024, partial recovery was evident, with the distribution shifting leftward but retaining greater variance than pre-pandemic periods, suggesting incomplete convergence to pre-crisis labour market conditions.

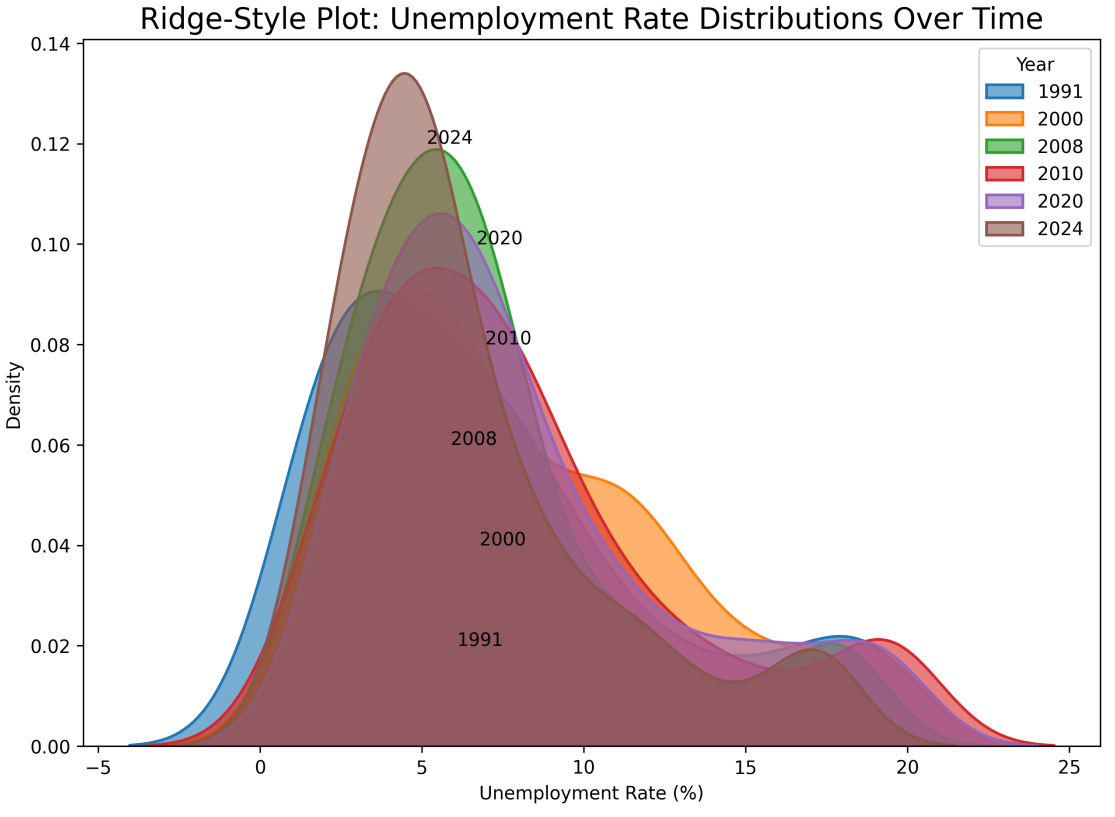

Figure 7: The ridge plot overlays density curves for unemployment distributions in select years (1991 blue, 2000 orange, 2008 green, 2010 red, 2020 purple, 2024 brown) showing narrow peaks below 10% in stable years shifting rightward and widening during crises. Density on y-axis, rates on x-axis, it visualizes evolutionary changes like increased dispersion post-2008 and partial leftward recovery by 2024.

3.2. Machine Learning Model Development

3.2.1. Feature Engineering and Data Transformation

To transform the longitudinal dataset into a supervised learning framework suitable for time-series forecasting, lagged features were systematically engineered. For each country, unemployment rates from the three immediately preceding years were constructed as predictor variables (Lag_1, Lag_2, Lag_3), with the subsequent year's unemployment rate designated as the target variable. This process generated a restructured dataset comprising 7,227 observations (country year combinations), each characterized by three temporal lag features plus 235 country dummy variables and temporal identifiers. Categorical encoding of country identifiers was accomplished through one-hot encoding, generating 235 binary indicator variables corresponding to the 219 retained countries after removing 16 regional aggregates that were redundant for country-level prediction. This encoding approach prevented the models from incorrectly assuming ordinal relationships among countries while enabling the capture of country-specific fixed effects in unemployment dynamics.

Feature standardization was applied to all continuous variables using Z-score normalization (mean = 0, standard deviation = 1). Post-transformation verification confirmed that all features achieved the target distributional properties, with means approximating zero and standard deviations of 1.0001, ensuring numerical stability for gradient-based optimization algorithms.

3.2.2. Training Testing Split and Validation Strategy

To ensure temporal integrity and prevent data leakage, the dataset was partitioned using a time-aware splitting strategy. All observations from years 1994 through 2019 were allocated to the training set (5,694 observations, 78.8% of data), while observations from 2020 through 2024 constituted the testing set (1,533 observations, 21.2% of data). This

temporal division ensured that model evaluation was conducted exclusively on future periods unseen during training, providing a realistic assessment of forecasting performance. The Mean Absolute Scaled Error (MASE) scaling factor, calculated as the in-sample naive forecast error for the training period, was determined to be 0.4812, establishing a benchmark for evaluating model performance relative to a simple persistence forecast.

4. Results

This chapter presents the empirical findings from the comparative analysis of machine learning models applied to unemployment rate prediction across 266 countries and regions spanning from 1991 to 2024.The result are organized into four sections: descriptive statistics,exploratory data analysis revealing temporal and regional patterns,feature engineering and preprocessing outcomes and comprehensive model performance evaluation.The findings are presented objectively, with detailed statistical metrics and visual representations supporting each analytical component.

4.1. Model Performance Evaluation

Four distinct machine learning algorithms were trained, optimized, and evaluated on the unemployment forecasting task: Linear Regression (baseline), Random Forest, XGBoost, and Elastic Net. Each model underwent hyper-parameter tuning using grid search cross validation on the training set, followed by final evaluation on the hold-out test set covering 2020-2024. Performance was assessed using four complementary metrics: Root Mean Squared Error (RMSE), coefficient of determination (R^2), Mean Absolute Percentage Error (MAPE) and Mean Absolute Scaled Error (MASE). These metrics collectively capture prediction accuracy, explanatory power, relative error magnitude and performance relative to naive forecasting baselines.

4.1.1. Linear Regression Performance

The baseline Linear Regression model achieved an RMSE of 1.0761 on the test set, indicating an average prediction error of approximately 1.08 percentage points in unemployment rate forecasts. The model explained 94.05% of variance in test set unemployment rates ($R^2 = 0.9405$), demonstrating strong overall predictive capacity. The MAPE of 11.79% indicated that predictions typically deviated from actual values by approximately 11.8% in relative terms. The MASE of 1.5003 revealed that the Linear Regression model's forecast errors were 50% larger than a naive persistence forecast, suggesting room for improvement over the simplest baseline approach.

4.1.2. Random Forest Performance

Hyper-parameter optimization for Random Forest via grid search cross-validation identified optimal parameters: 100 estimators, minimum samples split of 5, minimum samples per leaf of 1, maximum features set to 'log2' and unrestricted maximum depth. Cross-validation performance during hyper-parameter tuning yielded an average RMSE of 0.942 across validation folds. However, test set performance was somewhat diminished, with RMSE increasing to 1.1421, the highest among all evaluated models. The model achieved $R??$ of 0.9330, explaining 93.3% of test set variance. MAPE measured 12.88%, indicating the largest relative prediction errors among the four models. The MASE of 1.6261 represented a 62.6% increase over naive forecast errors, suggesting that Random Forest's ensemble approach, despite strong cross-validation performance, exhibited reduced generalization to the pandemic affected test period.

4.1.3. XGBoost Performance

XGBoost hyper-parameter optimization converged on the following configuration: 300 estimators, maximum depth of 4, learning rate of 0.05, minimum child weight of 5, subsample ratio of 1.0 and column subsample ratio of 1.0. Cross-validation during tuning produced an average RMSE of 0.830, the lowest among all models. Test set performance remained strong, with RMSE of 1.0571, representing the second-best absolute error performance. The model explained 94.26% of test set variance ($R?? = 0.9426$), closely approaching the theoretical maximum. MAPE of 11.06% indicated relatively low proportional errors, and MASE of 1.4346 demonstrated forecast errors only 43.5% above the naive baseline, representing substantial improvement over both Linear Regression and Random Forest.

4.1.4. Elastic Net Performance

Elastic Net hyper-parameter tuning identified optimal regularization parameters of alpha = 0.01 and L1 ratio = 1.0, effectively implementing a Lasso regression with minimal regularization strength. Cross-validation performance matched XGBoost with an average RMSE of 0.830. On the test set, Elastic Net achieved superior performance across all metrics: RMSE of 1.0419 (lowest among all models), R^2 of 0.9442 (highest explanatory power), MAPE of 10.88% (smallest relative errors) and MASE of 1.3957 (only 39.6% above naive forecasts). These results established Elastic Net as the best-performing model for unemployment rate forecasting on the test period spanning 2020-2024.

Table 1
Comparative Model Performance Metrics (Test Set 2020–2024)

Model	RMSE	R^2	MAPE (%)	MASE
Linear Regression	1.0761	0.9405	11.79	1.5003
Random Forest	1.1421	0.9330	12.88	1.6261
XGBoost	1.0571	0.9426	11.06	1.4346
Elastic Net	1.0419	0.9442	10.88	1.3957

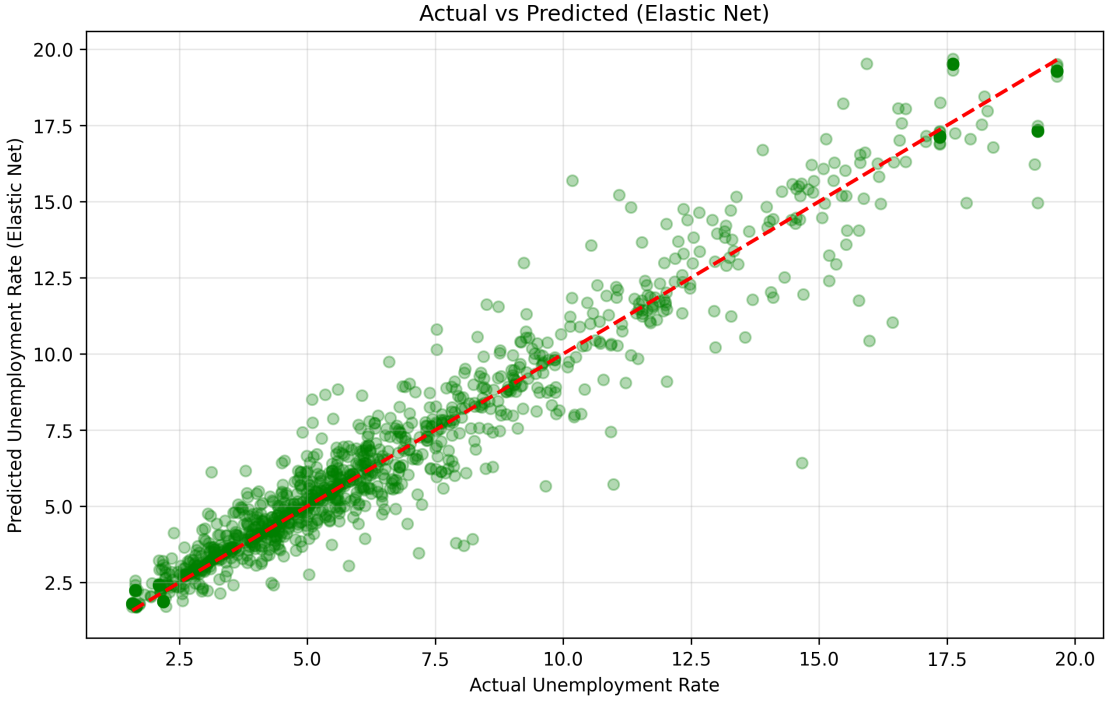

Figure 8: Actual versus predicted unemployment rates using the XGBoost model. The dashed red line represents the ideal 1:1 relationship, indicating perfect prediction accuracy, while the scattered points show the model's performance on the test dataset.

4.2. Model Deployment

4.2.1. Interactive Controls

The interface incorporates several interactive input widgets positioned within a sidebar, allowing users to customize the analysis dynamically. These components include:

- **Country selectors**,drop-down menus that enable users to choose one or more countries for analysis.
- **Time range controls**, slider-based inputs that allow users to define the historical period of interest (e.g from 1990 to 2024).
- **Indicator filters**, checkboxes or radio buttons used to select different unemployment indicators or visualization types.

These interactive controls operate in an event-driven manner. Any modification to user selections automatically triggers recalculation of the underlying results and updates the visual outputs in real time, eliminating the need for manual page refreshes.

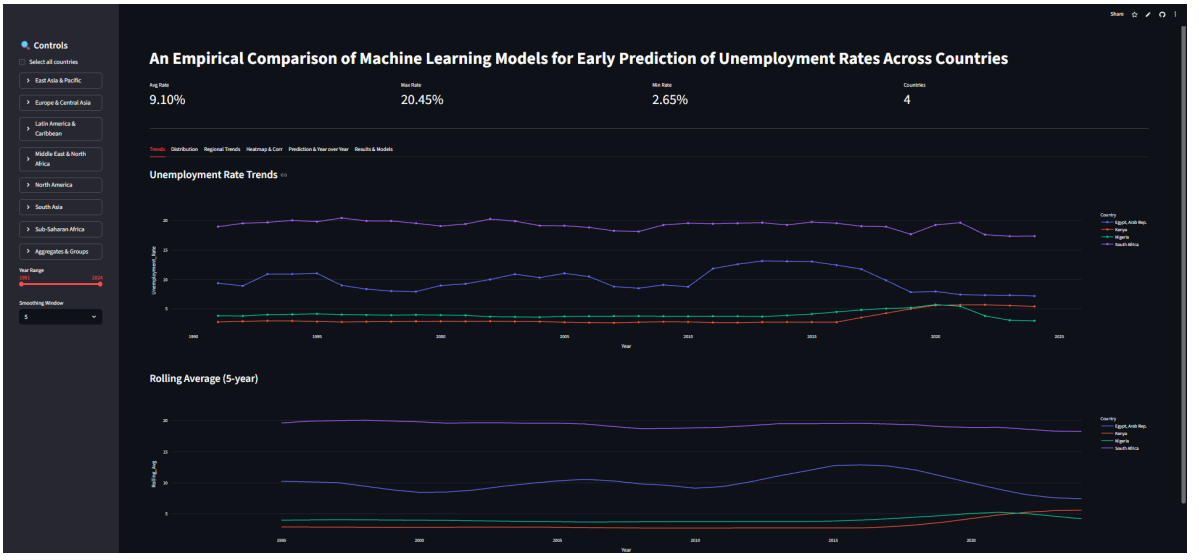

Figure 9: User interface showing interactive controls for country selection, time range filtering and indicator configuration in the deployed Streamlit application.

5. Discussion

In this empirical investigation, we set out to compare the efficacy of four machine learning models:Linear Regression as a baseline, Random Forest, XGBoost and Elastic Net,for predicting national unemployment rates one year ahead across 266 countries and regional aggregates, drawing on World Bank data from 1991 to 2024. The core research question was whether advanced machine learning techniques could outperform traditional methods in forecasting unemployment during periods of high volatility, such as the 2008 financial crisis and the COVID-19 pandemic, particularly in diverse economic contexts like those in Africa. Our findings unequivocally demonstrate that penalized linear models, specifically Elastic Net, provide the most robust predictive performance, achieving a Root Mean Squared Error (RMSE) of 1.0419 and an R^2 of 0.9442 on the test set (2020–2024). This surpasses XGBoost (RMSE = 1.0571, $R^2 = 0.9426$), Linear Regression (RMSE = 1.0761, $R^2 = 0.9405$), and Random Forest (RMSE = 1.1421, $R^2 = 0.9330$). These results underscore the value of regularization in handling multicollinearity and feature selection within high-dimensional datasets incorporating lagged unemployment rates, temporal features, and country-specific identifiers. From a practical standpoint, this implies that policymakers can leverage such models for early warning systems, enabling proactive interventions in labour markets prone to instability, such as those in Sub-Saharan Africa where unemployment volatility remains pronounced.

Placing these findings within the broader context of existing literature reveals both alignments and departures that enrich our understanding of predictive modelling in labour economics. Traditional econometric approaches, such as ARIMA and its variants, have long dominated unemployment forecasting due to their interpretability and focus on temporal dependencies Sharaf (2024),Isqeel Adesegun et al. (2020).However, our study aligns with emerging evidence that machine learning offers superior accuracy in capturing non-linear dynamics and heterogeneous predictors Katris (2020);Yurtsever (2023a). For instance, Elastic Net's outperformance echoes the strengths of penalized regressions highlighted in Kreiner and Duca (2020),who found Lasso-based methods effective for U.S. data, but our application extends this to a global scale, incorporating regional disparities absent in many prior studies.Unlike univariate ARIMA models, which struggled with non-stationarity in African contexts Sam et al. (2020),our multivariate framework—bolstered by features like lagged rates and country identifiers—demonstrates enhanced generalization. This is particularly resonant in regions like the Middle East and North Africa (MENA), where our exploratory analysis revealed persistently higher unemployment (often exceeding 10%) compared to Europe and East Asia, consistent with Mulaudzi (2021)and Celbiş (2023).Yet, our results diverge from deep learning-heavy approaches, such as LSTM-GRU hybridsOpondo et al. (2024), which excel in sequential data but demand larger datasets to avoid overfitting,a luxury not always available in developing economies. By prioritizing Elastic Net's balance of simplicity and sophistication, we bridge the gap between computational intensity and practical deployability, offering a model that policymakers in data-scarce environments can adopt without extensive resources.

An intriguing aspect of our results was the unexpected resilience of Linear Regression as a baseline, achieving an R^2 of 0.9405 despite its inability to model non-linear interactions. This was not anticipated, given the literature's emphasis on ensemble methods for complex economic data Güler et al. (2024);Mero et al. (2024).One plausible explanation lies in the dataset's inherent linearity in lagged unemployment rates, where past values strongly predict future ones, as evidenced by our feature engineering process. However,this linearity broke down during volatility peaks,such as 2020–2021 amid COVID-19, where Random Forest underperformed ($R^2 = 0.9330$) due to its sensitivity to outliers despite robust handling of heterogeneity. This unexpected gap highlights a key insight: while tree-based models like Random Forest and XGBoost capture interactions well,XGBoost's gradient boosting mitigated some noise, yielding a competitive RMSE of 1.0571 they falter in high-dimensional settings without explicit regularization. In contrast, Elastic Net's blend of L1 and L2 penalties effectively sparsified features, reducing multicollinearity from temporal and regional variables.This finding connects to broader labour market dynamics; for example, in Sub-Saharan Africa, where informal sectors dominate and data quality varies,the models' differential handling of noise explains why Elastic Net provided more stable predictions. If these unexpected outcomes are significant—and they are, as they reveal model vulnerabilities during crises—they underscore the need for hybrid strategies that combine linear baselines with ensemble refinements, much like the ARIMA-ANN hybrids in Katris (2020). Nonetheless, the persistence of regional disparities, with Sub-Saharan Africa showing volatility twice that of East Asia, was less surprising but reinforces the literature's call for context-specific modelling Wu (2023);Monir Aljinbaz and Al Rahhal (2024) and Monir Aljinbaz and Al Rahhal (2024).

No study is without limitations, and addressing them candidly enhances credibility while guiding future refinements. A primary constraint here is the reliance on annual World Bank data (SL.UEM.TOTL.ZS), which, while comprehensive, aggregates national figures and may mask sub-national variations, such as urban-rural divides prevalent in African economies Celbiş (2023).Data cleaning and outlier treatment were rigorous, but potential reporting biases in developing countries such as undercounting informal employment could introduce confounding variables, echoing limitations in Mutascu (2021). Moreover, our test period (2020–2024) captured extreme events like the pandemic, but the models' performance might degrade in more stable eras, as they were tuned for volatility. Multicollinearity from country identifiers was mitigated via Elastic Net, yet incorporating additional macroeconomic covariates could have enriched the feature set, as suggested by Sharaf (2024) and Sam et al. (2020).Computationally, while XGBoost and Random Forest demanded more resources, Elastic Net's efficiency is a strength, but all models assume stationarity post-differencing, which may not hold in rapidly evolving markets. These shortcomings did not undermine the primary outcomes Elastic Net's superiority but they highlight areas where the study falls short of capturing full labour market complexity, particularly in informal-heavy regions like Africa.

6. Conclusion

This study examined unemployment trends across 266 countries and regions from 1991 to 2024, using World Bank data to highlight both long standing patterns and changes in labour markets worldwide. The descriptive analysis revealed

clear regional differences, with Africa showing consistently higher unemployment rates than Europe, as well as the strong effects of major external events, such as the 2008 global financial crisis and the COVID-19 pandemic. Going beyond description, the research compared four machine learning models for forecasting unemployment one year ahead: Linear Regression as a baseline, Random Forest for its ensemble strength, XGBoost for advanced gradient boosting and Elastic Net for its penalized linear approach with regularization. Trained on data from 1991 to 2019 and tested on the 2020 to 2024 period, Elastic Net performed best, with the lowest RMSE (1.0419) and highest R^2 (0.9442) scores, followed closely by XGBoost, by effectively handling feature selection, regularization and capturing subtle shifts in labour market indicators. These results have important implications for both academic research and policy-making. In a world facing growing uncertainty, traditional economic models often struggle because they rely on assumptions of steady, linear trends. The strong performance of Elastic Net shows the clear value of machine learning in providing reliable early warnings, through techniques that address multicollinearity and over-fitting in economic datasets. For governments and international organizations, accurate forecasts support timely actions, such as job creation programs.

References

Aris, M.N.M., Nagaratnam, S., Zakaria, N.N., Azami, M.F.A.M., Samsudin, M.A.I., Othman, E.S., 2024. A comparative study of gaussian process machine learning and time series analysis techniques for predicting unemployment rate, in: 2024 16th International Conference on Computer and Automation Engineering (ICCAE), IEEE. pp. 242–246.

Celbiş, M.G., 2023. Unemployment in rural europe: A machine learning perspective. Applied Spatial Analysis and Policy 16, 1071–1095. URL: https://doi.org/10.1007/s12061-022-09464-0, doi:10.1007/s12061-022-09464-0.

Estrada-Moreno, J.C., Rendon-Lara, E., Jiménez-Núñez, M.d.l.L., 2024. Combination of artificial neural networks and principal component analysis for the simultaneous quantification of dyes in multi-component aqueous mixtures. Applied Sciences 14. URL: https://www.mdpi.com/2076-3417/14/2/809, doi:10.3390/app14020809.

Güler, M., Kabakçı, A., Koç, , Eraslan, E., Derin, K.H., Güler, M., Ünlü, R., Türkan, Y.S., Namlı, E., 2024. Forecasting of the unemployment rate in turkey: Comparison of the machine learning models. Sustainability 16. URL: https://www.mdpi.com/2071-1050/16/15/6509, doi:10.3390/su16156509.

Ibrahim, F.J., Umar, H.A., Bichi, A.S., Ahmad, I.S., Rabiu, N.B., Ahmad, A.M., . Prediction of unemployment rates with time series and machine learning techniques .

Isqeel Adesegun, O., Mathew, O.O., Omotola, S.B., 2020. Modeling unemployment rates in nigeria using time series approach. Asian Journal of Mathematical Sciences .

Katris, C., 2020. Prediction of unemployment rates with time series and machine learning techniques. Computational Economics 55, 673–706. URL: https://doi.org/10.1007/s10614-019-09908-9, doi:10.1007/s10614-019-09908-9.

Kreiner, A., Duca, J.V., 2020. Can machine learning on economic data better forecast the unemployment rate? Applied Economics Letters 27, 1434–1437. URL: https://doi.org/10.1080/13504851.2019.1688237, doi:10.1080/13504851.2019.1688237.

Madaras, S., 2024. Deep learning algorithm forecasting the unemployment rates in the central european countries. Economics and Business 38, 86–102. URL: https://doi.org/10.7250/eb-2024-0006, doi:10.7250/eb-2024-0006.

Manasa, S., Kalidas, M., et al., 2022. Unemployment rate forecasting using supervised machine learning model. IJCSPUB-International Journal of Current Scienc (IJCSPUB) 12, 452–455.

Mero, K., Salgado, N., Meza, J., Pacheco-Delgado, J., Ventura, S., 2024. Unemployment rate prediction using a hybrid model of recurrent neural networks and genetic algorithms. Applied Sciences 14. URL: https://www.mdpi.com/2076-3417/14/8/3174, doi:10.3390/app14083174.

Monir Aljinbaz, A.M., Al Rahhal, M.M., 2024. Forecasting unemployment rate for multiple countries using a new method for data structuring. International Journal of Advanced Computer Science & Applications 15.

Mulaudzi, R., 2021. An exploration of machine learning models to forecast the unemployment rate of south africa: A univariate/multivariate approach. URL: https://wiredspace.wits.ac.za/items/d943a369-8127-42c2-95ca-3f2f16fc514f?utm_source=chatgpt.com. mSc Thesis, University of the Witwatersrand.

Mutascu, M., 2021. Artificial intelligence and unemployment: New insights. Economic Analysis and Policy 69, 653–667.

Opondo, R., Bundi, D., Weke, P., 2024. Modelling and forecasting unemployment trends in kenya using advanced machine learning techniques. American Journal of Theoretical and Applied Statistics 13, 242–254. URL: https://doi.org/10.11648/j.ajtas.20241306.16, doi:10.11648/j.ajtas.20241306.16, arXiv:https://article.sciencepublishinggroup.com/pdf/10.11648.j.ajtas.20241306.16.

Prihandi, I., Wijono, S., Sembiring, I., Maria, E., 2025. Implementation of arima with min-max normalization for predicting the price and production quantity of red chili peppers in north sumatra province considering rainfall and sunlight duration factors. Engineering, Technology & Applied Science Research 15, 21876–21887.

Sam, S.O., Manene, M.M., Kipchirchir, I.C., Pokhariyal, G.P., 2020. Cointegration analysis of youth unemployment in kenya .

Sharaf, F., 2024. Comparing mathematical models for forecasting the youth unemployment rate in jordan .

Shrief, W.A., Taha, A., Elstohy, R., Nagy, N., Ali, E.M., 2025. Optimization of hybrid machine learning approach for unemployment rate forecasting. Applied Computational Intelligence and Soft Computing 2025, 3817650.

Simionescu, M., Cifuentes-Faura, J., 2022. Forecasting national and regional youth unemployment in spain using google trends. Social Indicators Research 164, 1187–1216.

Tufaner, M.B., Sözen, İ., 2021. Forecasting unemployment rate in the aftermath of the covid-19 pandemic: The turkish case. İzmir İktisat Dergisi 36, 685–693.

Wu, Y., 2023. The importance of studying the unemployment rate in china. Journal of Education, Humanities and Social Sciences 24, 465–470. URL: https://drpress.org/ojs/index.php/EHSS/article/view/16988, doi:10.54097/9maq2b67.

Yurtsever, M., 2023a. Unemployment rate forecasting: Lstm–gru hybrid approach. Journal for Labour Market Research URL: https://labourmarketresearch.springeropen.com/articles/10.1186/s12651-023-00345-8?utm_source=chatgpt.com.

Yurtsever, M., 2023b. Unemployment rate forecasting: Lstm-gru hybrid approach. Journal for Labour Market Research 57, 18.

Zhao, L.F., 2020. Data-driven approach for predicting and explaining the risk of long-term unemployment, in: E3S Web of Conferences, EDP Sciences. p. 01023.