Contents

Project Abstract: Forecasting Youth Unemployment Rates

This project focuses on forecasting youth unemployment rates using historical data from the World Bank. The problem addressed is the prediction of future youth unemployment rates for different countries based on past trends.

The data used is the "Unemployment, youth total (% of total labor force)" indicator from the World Bank, specifically the dataset "API_SL.UEM.1524.ZS_DS2_en_csv". The data includes unemployment rates for various countries across several years.

The methods employed involve several steps:

1. Data Loading and Cleaning: The data is loaded from a CSV file and cleaned by dropping irrelevant columns and handling missing values.

2. Exploratory Data Analysis (EDA): Visualizations such as line plots and heatmaps are used to explore unemployment trends over time for selected countries and to understand the correlation between unemployment rates across different years.

3. Data Preparation for Modeling: The data is prepared for regression modeling by selecting relevant historical years as features and a target year for prediction. Features are standardized using `StandardScaler`.

4. Model Training and Evaluation: Several regression models, including RandomForestRegressor, Linear Regression, Ridge, and ElasticNet, are trained on the prepared data. Their performance is evaluated using metrics like Mean

Squared Error (MSE) and R-squared. Cross-validation and learning curves are used to assess model stability and identify potential overfitting.

5. Hyperparameter Tuning: GridSearchCV is applied to the RandomForestRegressor, Ridge, and ElasticNet models to find the optimal hyperparameters for improved performance.

6. Prediction: The best-performing model (ElasticNet based on testing and cross-validation results) is used to make predictions for future unemployment rates, specifically demonstrated by forecasting the 2019 unemployment rate for South Africa.

Major findings include:

- Visualizations revealed diverse unemployment trends across different countries and a strong positive correlation between unemployment rates in consecutive years.
- The ElasticNet model, after hyperparameter tuning, demonstrated promising performance with a low Mean Squared Error and a high R-squared score on the test data, suggesting it is a suitable model for this forecasting task.
- A prediction for South Africa's youth unemployment rate in 2019 was made using the trained ElasticNet model.

Overall, this project provides a comprehensive approach to forecasting youth unemployment rates using readily available historical data and standard machine learning techniques. The results suggest that regression models, particularly regularized linear models like ElasticNet, can effectively capture the underlying patterns in the data for forecasting purposes.

Introduction

Youth unemployment is a significant global challenge with far-reaching social and economic consequences. High rates of youth unemployment can lead to social unrest, increased poverty, and underutilization of human capital, impacting both individual well-being and national development. In South Africa, the problem is particularly severe according to Statistics South Africa, the youth unemployment rate stood at approximately 46.1% in the first quarter of 2025, one of the highest in the world (Statistics South Africa, 2025). This persistent challenge reflects deeper structural issues such as skills mismatches, slow economic growth, and limited job creation in key sectors.

Understanding and forecasting youth unemployment trends are crucial for policymakers, economists, and social organizations to develop effective strategies and interventions.

Accurate predictions can support the design of evidence-based policies aimed at reducing unemployment and improving access to economic opportunities for young people.

This project aims to explore historical youth unemployment data and build a predictive model to forecast future unemployment rates. The problem we address is the need for accurate and reliable forecasts of youth unemployment to inform policy decisions and resource allocation.

The objectives of this project are to:

1. Load, clean, and prepare a dataset containing historical youth unemployment rates for various countries.
2. Conduct exploratory data analysis to visualize trends and identify patterns in youth unemployment over time and across different regions.
3. Develop and evaluate several regression models for forecasting youth unemployment rates.
4. Utilize hyperparameter tuning techniques to optimize the performance of the selected models.
5. Provide a forecast of youth unemployment for South Africa as a demonstration of the model's capability.

By achieving these objectives, this project seeks to contribute to a better understanding of youth unemployment dynamics and provide a data-driven tool for anticipating future trends, particularly in South Africa, where tackling unemployment remains a national priority.

Methodology

The methodology adopted for this project involves a structured approach to forecasting youth unemployment rates, encompassing data handling, exploratory analysis, model development, and evaluation.

1. Data Collection: The primary data source for this project is the "Unemployment, youth total (% of total labor force)" indicator from the World Bank's World Development Indicators. The data was accessed as a CSV file ("API_SL.UEM.1524.ZS_DS2_en_csv").

2. Data Preprocessing and Cleaning:
    - The downloaded CSV file was loaded into a pandas DataFrame.
    - Initial inspection revealed several columns with missing headers or entirely missing values, which were subsequently dropped.
    - Column names were cleaned by removing leading/trailing spaces and addressing 'Unnamed' columns that arose due to data structure.

- For specific analyses and modeling, data was filtered for countries of interest and reshaped from a wide format (years as columns) to a long format (a 'Year' column and an 'Unemployment Rate' column) using the `melt` function.
- Missing values in the numerical year columns were handled by filling them with 0 for the purpose of correlation analysis and model training. This approach was chosen based on the nature of the data and the selected models, but other imputation methods could also be considered.
- The 'Year' column, when converted from column headers, was cleaned to ensure it contained only numeric values and was converted to an integer type.

3. Exploratory Data Analysis (EDA):
    - Basic descriptive statistics of the DataFrame were generated to understand the data's distribution and identify potential issues.
    - Line plots were created to visualize the trend of youth unemployment over time for selected countries, allowing for the observation of individual country trajectories.
    - Bar plots were used to compare unemployment rates across different countries for a specific year.
    - A correlation heatmap was generated to visualize the relationships between unemployment rates across different years, providing insights into temporal dependencies in the data.

4. Model Building:
    - The data was prepared for supervised learning by defining features (unemployment rates from a range of historical years) and the target variable (unemployment rate in a future year). Specifically, years 2012-2017 were used to predict the unemployment rate in 2018.
    - The feature data was standardized using `StandardScaler` to ensure that each feature contributed equally to the models, which is particularly important for models sensitive to the scale of input features (like linear models).
    - The dataset was split into training and testing sets to evaluate model performance on unseen data.
    - Several regression models were implemented:
        - RandomForestRegressor: An ensemble method that uses multiple decision trees to make predictions.
        - Linear Regression: A basic linear model that finds the best-fitting straight line through the data.

        - Ridge Regression: A linear regression model with L2 regularization to prevent overfitting.
        - ElasticNet Regression: A linear regression model that combines L1 and L2 regularization.

5. Model Evaluation and Hyperparameter Tuning:
    - The performance of each model was evaluated using standard regression metrics:
        - Mean Squared Error (MSE): Measures the average squared difference between actual and predicted values.
        - R-squared (R^2) Score: Indicates the proportion of the variance in the target variable that is predictable from the features.
    - Overfitting was assessed by comparing training and testing set performance.
    - Cross-validation (specifically 5-fold cross-validation) was employed to obtain a more robust estimate of model performance and assess stability.
    - Learning curves were plotted to visualize how the training and cross-validation scores change with the number of training examples, helping to identify potential issues like high bias or high variance.
    - GridSearchCV was used to systematically search for the optimal hyperparameters for the RandomForestRegressor, Ridge, and ElasticNet models, aiming to improve their performance on unseen data. The best parameters were selected based on the cross-validation MSE.

6. Prediction:
    - The best-performing model (ElasticNet, based on the evaluation metrics) was used to make a forecast for the youth unemployment rate in 2019 for a specific country (South Africa) as a demonstration of the project's practical application. The input data for the prediction was prepared in the same way as the training data (using the preceding years as features and scaling with the fitted scaler).

Discussion of Results and Findings

This section interprets and analyzes the results obtained from the exploratory data analysis, model training, and evaluation steps, highlighting key insights and challenges encountered during the project.

Exploratory Data Analysis Insights:

- The initial data inspection and visualizations (line plots for individual countries) revealed significant variability in youth unemployment rates across different nations and over time. Some countries show relatively stable rates, while others exhibit considerable fluctuations.
- The bar plot for a specific year demonstrated the disparity in youth unemployment levels between countries, emphasizing the global nature of the challenge but also the diverse national contexts.
- The correlation heatmap provided valuable insights into the temporal dependencies of youth unemployment. The strong positive correlations observed between unemployment rates in consecutive years suggest that past unemployment trends are indeed strong indicators of future rates, validating the approach of using historical data for forecasting. The decreasing correlation as the time gap between years increases is also an expected and important finding, indicating that the influence of older data diminishes.

Model Performance Comparison:

The project evaluated several regression models for forecasting youth unemployment rates, including RandomForestRegressor, Linear Regression, Ridge, and ElasticNet. The performance was assessed using Mean Squared Error (MSE) and R-squared (R2) score on the test set and through cross-validation.

Model	Training MSE	Training R^2	Test MSE	Test R^2	Mean CV MSE
RandomForestRegressor	1.93	0.986	4.28	0.963	13.63
Linear Regression	7.65	0.945	1.78	0.985	8.11
Ridge Regression (Tuned)	-	-	1.77	0.985	8.08
ElasticNet Regression (Tuned)	-	-	1.71	0.985	8.06

- The Linear Regression model performed surprisingly well on the test set, achieving a low MSE and a high R-squared score. However, its significantly higher training MSE compared to the test MSE suggests potential issues with bias on the training data, although the cross-validation MSE is reasonable.
- The RandomForestRegressor showed excellent performance on the training data (very low MSE, high R-squared), indicating it learned the training patterns very well. However, its performance on the test set and the significantly higher cross-validation MSE compared to the training MSE suggest some degree of overfitting to the training data. While still a strong model, it might not generalize as well to completely unseen data as a more regularized model.

- Regularized Linear Models (Ridge and ElasticNet): Both Ridge and ElasticNet, particularly after hyperparameter tuning, demonstrated strong and consistent performance. Their test MSE and R-squared scores were comparable to or better than the basic Linear Regression, and their cross-validation MSEs were lower than that of the RandomForestRegressor. This indicates that regularization helped to improve their generalization ability and prevent overfitting.
- The ElasticNet Regression model with the best hyperparameters (`alpha': 0.01, 'l1_ratio': 1`, which effectively made it a Lasso model) achieved the lowest test MSE and the highest test R-squared score among all evaluated models. Its cross-validation MSE was also the lowest, suggesting good stability and generalization. This indicates that for this dataset and problem, the Lasso type of regularization (L1 penalty) was effective in selecting relevant features and improving the model's performance.

Interpretation of ElasticNet Coefficients:

The coefficients of the best ElasticNet model provide insights into the influence of each historical year's unemployment rate on the prediction for the target year (2018 in the training phase).

Conclusion

This project successfully explored and analyzed historical youth unemployment data and developed predictive models for forecasting unemployment rates. We started by loading and cleaning the dataset, followed by exploratory data analysis to understand the trends and correlations within the data.

The key objectives of the project were met:

- We successfully loaded, cleaned, and prepared the dataset for analysis and modeling.
- Through visualizations, we gained insights into country-specific unemployment trends and the temporal relationships between unemployment rates in different years.
- We developed and evaluated several regression models, including RandomForestRegressor, Linear Regression, Ridge, and ElasticNet.
- Hyperparameter tuning was applied to improve the performance of the models, with GridSearchCV identifying optimal parameters for the tuned models.
- Finally, we demonstrated the forecasting capability of the best-performing model (ElasticNet) by predicting the youth unemployment rate for South Africa in 2019.

The analysis of model performance indicated that regularized linear models, particularly ElasticNet, performed well on this dataset, showing good generalization ability and lower susceptibility to overfitting compared to the un-tuned RandomForestRegressor. The interpretation of the ElasticNet coefficients provided some insight into which historical years had a stronger influence on the predictions.

Future Work:

Several avenues could be explored to extend this project:

- Incorporate more features: Including other relevant economic indicators (e.g., GDP growth, inflation, education levels, labor force participation rates) could potentially improve the accuracy of the forecasts.
- Explore different time series models: Investigate dedicated time series forecasting models like ARIMA, SARIMA, or more advanced techniques like LSTMs, which are designed to handle sequential data and might capture temporal dependencies more effectively.
- Forecast for multiple future years: Extend the forecasting horizon to predict unemployment rates further into the future, although this would require careful consideration of the increasing uncertainty with longer horizons.
- Regional or specific country analysis: Conduct more in-depth analysis and modeling focusing on specific regions or countries of interest, potentially using more granular or region-specific data if available.
- Model interpretability: Further explore techniques to enhance the interpretability of the models, especially complex ones like RandomForest, to gain deeper insights into the factors driving youth unemployment.
- Consider external factors: Account for potential impacts of unforeseen events (e.g., economic crises, pandemics, policy changes) that could significantly influence unemployment rates and are not captured in historical data alone.

References

Statistics South Africa (2025) *Quarterly Labour Force Survey (QLFS): Quarter 1 2025*. Pretoria: Statistics South Africa. Available at: https://www.statssa.gov.za/publications/P0211/Media%20Release%20QLFS%20Q1%202025.pdf (Accessed: 18 October 2025).