A STUDY ON THE IMPACT OF POPULATION GROWTH AND URBANIZATION ON KARACHI'S ECONOMIC GROWTH

S. Abbas[1]*, Z. Xinyue[1]

[1]School of Geographical Sciences, Inner Mongolia Normal University, 010022 Hohhot, China.
*Corresponding Authors email: sajjadabbas@qq.com

ABSTRACT: The study aims to assess the complex dynamics of population growth, urbanization, and economic performance in Karachi using a multivariate data set of 86 observations on 57 variables. The average GDP was $9.08 million (SD = $6.41 million), while per capita income was low at $3.7, showing a disparity in income distribution among the cities. The urban population averaged 41.9 million, with a 2.56% growth rate in urbanization. Environmental degradation is rife, with the air quality index averaging 51.4 and water pollution averaging 68.1. Carbon emissions and fossil fuel consumption had also increased, averaging over 51 units and 0.60, respectively. Random Forest Regression had the highest accuracy of all the techniques used ($R^2 > 0.99$, RMSE ≈103,957), identifying urban population, carbon emissions, and youth unemployment as key predictive drivers. The SEM analysis showed a significant impact of urban population on GDP ($\beta = 0.4654$); however, per capita income did not play an important role. One-sample t-tests confirmed that the urban population ($t = 13.36$, $p < 0.001$) and CO_2 emissions ($t = 16.83$, $p < 0.001$) had significant impacts. While GDP was increasing, poverty and unemployment were also rising, indicating the increasing inequality in wealth distribution. Clustering (KMeans; silhouette = 0.5524) formed three demographic-economic segments. This study indicates that the economic progression of Karachi is intricately interlinked with rapid urbanization while conversely being challenged with structural inequities.

Keywords: Urbanization, Population Growth, Economic Growth, Karachi, Machine Learning, Structural Inequality.

(Received 18.01.2025 Accepted 01.03.2025)

INTRODUCTION

Karachi, the largest city in Pakistan and the leading financial and industrial center, has seen rapid changes over the last few decades (Hasan *et al*., 2010). By 2023, the population of Karachi is estimated to be greater than 20 million, which makes it the world's 12th largest city by population (World Population Review, 2023). The city contributes about 25% of Pakistan's GDP, 45% of the industrial sector, and more than 50% of the country's port revenue (Yasmeen *et al*., 2018). Given these statistics, Karachi is pivotal to national economic stability and growth (Hasan *et al*., 2010). However, growth has proven both a blessing and a challenge (Turok & McGranahan, 2013). The rapid influx of people, growth of urban sprawl, and degraded infrastructure have all posed challenges to effectively balancing this metropolis's economic potential and urban sustainability (Yasmeen *et al*., 2018). The complex dynamics thus provide a crucial perspective for questioning the net impact of all such growth on Karachi's long-term economic path (Glaeser, 2011).

There exists an array of reasons behind the exponential population growth in Karachi (Pakistan Bureau of Statistics, 2021). A central reason is rural-to-urban migration, whereby thousands of people move to Karachi each month seeking better livelihood opportunities (Yasmeen *et al*., 2018). According to the Pakistan Bureau of Statistics, an estimated 1,000 new migrants arrive in the city daily, most of whom come from Sindh, Punjab, Khyber Pakhtunkhwa, and Balochistan (Pakistan Bureau of Statistics, 2021). Another contributor to crowding is the natural population increase due to high birth rates, which maintains the city's image as an economic magnet (Hasan *et al*., 2010). The population density of Karachi is more than 24,000 people per square kilometer, far higher than the national average (World Population Review, 2023). This massive population influx has been poorly managed by urban planning, resulting in over 60% of the population being housed in informal or unregulated settlements (Yasmeen *et al*., 2018). They represent the underlying systemic problems restraining effective governance and sustainable development planning (Hasan *et al*., 2010).

However, for Karachi, population growth and urbanization differ from economic growth (Glaeser, 2011). The city has shown a capacity for adding output value; however, shadow prices remain, such as their effects on unregulated housing, traffic congestion, pollution, water cuts, and even high unemployment (Turok & McGranahan, 2013). The informal sector houses about 65% of the working population at the city level (Yasmeen *et al*., 2018). However, it does contribute very little to national productivity or taxes (Hasan *et al*.,

2010). These structural issues point towards an important question: Do population growth and urbanization drive the economy of Karachi, or are they increasingly straining it beyond repair? (Yasmeen *et al*., 2018). The study attempts to look into both those phenomena in terms of how the outcome states—growth, population, and urbanization—will seek to achieve or add an effect on economic outcomes for Karachi positively or negatively (Turok & McGranahan, 2013).

Urbanization has mainly taken an edge in the developmental discourse at a global level (Glaeser, 2011). Thus, it argues that cities drive innovation and productivity by clustering diverse talents and resources (Glaeser, 2011). In much the same vein, the studies on urbanization in South Asia point to positive effects when coupled with sound governance and infrastructure investment (Turok & McGranahan, 2013). In South Asia, a very particular case comes in Karachi (Hasan *et al*., 2010). According to Hasan *et al*. (2010), much of the consequent urban growth does not enter into the planning and regulation fold at all, so consequently, the services provided are inadequate, and economies are informalized (Hasan *et al*., 2010). Poor governance and the absence of policy based on data deficiency curb economic benefits from Karachi's demographic growth, as submitted by Yasmeen *et al*. (2018) (Yasmeen *et al*., 2018). Hence, urbanization may provide certain theoretical advantages; its impingement in Karachi depends on regulatory frameworks, infrastructure capacity, and social equity (Turok & McGranahan, 2013).

The study was to assess population growth and urbanization effects on Karachi's economic performance. The research also analyzed historical and projected population growth trends and urban sprawl in Karachi over the past thirty years. The study also examined the relationship between these demographic trends and Karachi's economic indicators, such as GDP, employment, productivity, and income distribution. This research also aimed to examine the role of urban planning, governance, and infrastructure in mediating the relationship between urbanization and development effects. Finally, it also aimed to suggest strategic policy interventions for harnessing population growth's benefits and diminishing economic costs.

METHODOLOGY

Study Design and Objectives: The research adopts quantitative and data-driven approaches to assess the economic situation of Karachi as it is affected by population growth and urbanization. The main aim would be to investigate the ways in which demographic and urban expansion variables will affect economic measures, such as GDP, unemployment, and capital investment in infrastructure over time. More specifically, the urbanization rates and population density serve as examples of demographic and urban expansion variables. The methodology thus also integrates geospatial mapping to regression-based modeling to conform to visual and statistical perspectives at urban-economic linkages within Karachi.

Data Collection and Description: The dataset referred to in the study was taken from a longitudinal panel of urban, demographic, and economic parameters compiled into the Excel spreadsheet "Karachi Dataset.xlsx." The set contains 86 observations across different districts and over the years, representing time duration and geographical perspective across 57 variables. Central Values are part of 57 variables, including GDP (in billions of USD) ranging from 1.05 million to 7.6 million USD. Per Capita, the income span value ranges from approximately USD 3.4 to USD 8.7. It has about 14 million under whose population hundreds of thousands do not exceed. Population Density is inferred from urban population and geographical identifiers. Latitude and Longitude are provided for geospatial plotting. Urban Population values are mostly around indicators dating from 1960 to 1970. Unemployment and Youth Unemployment Rates are represented for several years. Poverty rate, literacy rate, life expectancy, and infrastructure spending, all of which are available, provide different contexts for economic health and environmental indicators such as CO_2 emissions, air quality scores, and fossil fuel consumption. Each observation bears geographic coordinates (latitude and Longitude) such that it can be geospatially visualized and analyzed.

Data Preprocessing: The data was subject to an intense phase of preprocessing. Column names were stripped of white space in addition to special characters. All columns relating to numeric data were put to their appropriate data types. Any variable with over 30% missing data was instantly removed from the modeling procedure. The remaining variables, with missing numeric values, underwent either forward-fill or were mean or median imputed according to their distribution characteristics. All were standardized using z-score normalization to be ready for use by machine learning models sensitive to scale.

Exploratory Data Analysis (EDA): The exploratory data analysis assessed variable distributions, detected outliers, and visualized the relation between predictors and economic outcomes. Histograms, box plots, violin plots, and correlation heat maps were generated to show the strong relationship between variables and the strength of the relationship between inter-variables. This process has enabled the selection of high variance and most correlated features to economic outputs, particularly GDP. The correlation matrix further facilitated variable reduction by removing multicollinear predictors.

Geospatial Analysis: Geospatial mapping is a pivotal aspect of the study that visualizes economic indicators when considering urban expansion. Interactive maps on an open-source geographic data set were built using Python and latitude and longitude coordinates. In contrast, an open-source visualization framework offered by the Folium library added a graphical charm to the setting. Four main types of geospatial visualization were generated: a GDP heatmap showing the intensity of economic activities across Karachi districts; a population heatmap suggesting zones with demographic visibility; a circle marker map for every city for its GDP in geographical terms; an overlay map showing how population and GDP are distributed spatially. The maps allowed an intuitive geographical understanding of how spatial variables correlate with economic performance; hence, they added a seductive visual touch to the regression analysis.

Feature Selection: Feature selection was achieved by applying correlation-based filtering and model-based methods. The correlation matrix showed all variables with an absolute correlation of more than 0.50 with the dependent variable (GDP). The feature importance scores of the Random Forest were considered for selecting the main predictors. The features proposed for modeling were the following: Population, Urban Population, Population Growth Rate, Youth Unemployment Rate, Labor Force Participation, Poverty Rate, Infrastructure Spending (where applicable), as well as environmental quality indexes (i.e., Air Quality, CO_2 emissions).

Regression Modelling: The study aims to apply multiple regression models to predict GDP based on selected predictors. Eight distinct regression models were trained and validated. A basic linear regression model was built to measure the presence of any linear relationships. Ridge regression was used to reduce the model's complexity and address multicollinearity. Lasso regression was used mainly for feature selection, as well as for regularization. Elastic net regression created a trade-off through Ridge and Lasso penalties for bias and variance. Polynomial regression was then used to capture non-linear relationships up to the second degree. Decision trees were used in regression for non-parametric modeling based on rules. Random forests are regarded as an ensemble learning algorithm aggregating results from multiple decision trees. Using kernel methods, support vector regression (SVR) was applied to high-dimensional feature spaces. Each model was trained on 80% of the entire dataset and validated on the remaining 20%, with performances gauged by R2 Score, MAE, and RMSE. Polynomial and ensemble models have consistently outperformed the linear ones, particularly Random Forest, giving sufficient evidence of non-linearity and interaction effects in the data.

Model Evaluation and Validation: The performance of the models was compared using three metrics: R2 Score, MAE, and RMSE. The Random Forest Regression model achieved the highest value for R2 (>0.99) and the lowest RMSE, thus generalizing better on test data. The Polynomial Regression model also performed exceptionally well, indicating that the relationship between the urban variables and economic output is strongly nonlinear. The residual plots were also used to check for homoscedasticity and non-random errors the actual vs. predicted plots provided further validation regarding the accuracy of the top-performing models. Random forest importance plots further validated the strong relevance of such key features as population, urban density, and youth unemployment.

Visualization and Interpretation: Eight types of plots were derived to visually interpret both geospatial and model-driven outcomes: Residual Plot (linear model), Actual vs. Predicted Scatter Plot (Random Forest), Feature Importance (Random Forest), Regression Coefficients (linear model), Correlation Heatmap, Histogram of Residuals, Decision Tree Plot, and Interactive Geospatial Maps (GDP Heatmap, Population Heatmap, Combined Map). These visualizations supported quantitative findings well, providing actionable insights for urban planning and economic policy.

Export and Documentation: All model results were exported to a CSV file called "Karachi_Regression_Results.csv," containing model performance metrics for each regression technique. The plots were stored in high-resolution PNG format in a folder named "karachi_plots." The geospatial HTML maps were exported for viewing online. The analysis pipeline was built in Python using Jupyter Notebooks and is transparent, reproducible, and scalable.

RESULTS

Exploratory Data Analysis: The study results offer broad characterizations and statistical analyses of the population-demographic growth relation with urbanization indicators and economic development in Karachi. The data encompasses 86 observations of 57-plus variables relating to economic performance, urban demographics, environmental quality, access to infrastructure, and labor participation. Each variable is thus relevant to analyzing how population dynamics and urban infrastructure influence Karachi's economic history.

Descriptive statistics show that the mean GDP is about 9.08 million USD, the standard deviation being 6.41 million USD, meaning that there is a considerable variation in economic performance over time. Per capita income shows a low average value of 3.7 USD, a very low range from 2.21 to 8.72 USD. This implies that

income distribution is markedly unequal, tending toward urban-based inequality. Urban population indicators show that the average urban population stands at approximately 41.9 million and that urbanization growth stands at an average growth rate of about 2.56%, indicating prolonged urban aggrandizement. Air quality indicators and water pollution gauges are also highly variable. The average air quality index is 51.4, ranging from 0 to 100, while that of water pollution records is 68.1. Such values indicate severe degradation of the environmental quality in the urbanization process. Trends in carbon dioxide emission and greenhouse gas potentials indicate an increase in fossil fuel consumption, an average of 0.60 for fossil fuel use and over 51 units for emissions. Regression analysis with multiple model linear regression, Ridge, Lasso, ElasticNet, Polynomial, Random Forest, Decision Tree, and SVR showed that urban economic growth and GDP have significant predictors. The model with Random Forest Regression has shown maximum performance levels with an R-square value above 0.99 and the lowest RMSE, showing that urban population, infrastructure expenditure, youth unemployment, and carbon emission are some of the best indicators. Variables from polynomial models also performed well, confirming the non-linear nature of variables connecting demographic indicators with GDP.

One-sample t-tests have shown that many variables have significant differences against their hypothetical means, which shows how much demographic and urban metrics affect the economic outcomes of Karachi. Some critical variables, such as urban population ($t = 13.364$, $p < 0.001$), carbon emissions ($t = 16.829$, $p < 0.001$), and renewable energy adoptions ($t = 70.637$, $p < 0.001$), revealed significant mean differences among them. Abstracted measures of life expectancy, literacy rate, and infrastructure indicators also reflected a very high level of statistical significance over their variation across time. Moreover, ANOVA results have further justified various predictor-economic outcome impacts. If the sum of squares were gathered for GDP across groups, it would exceed 802 trillion USD, indicating economic heterogeneity in urban settings. The significance of predictors was also confirmed through effect size measurements, especially for variables such as urban population (Cohen's $d > 2.88$; rural population (Cohen's $d > 2.35$); and CO_2 emissions (Cohen's $d > 3.02$ with large effect sizes and narrow confidence intervals. The geospatial mapping visualized the patterns of regional inequality in economic development. The claimed gross domestic product estimates are significantly higher in districts that are much more densely populated and better served by infrastructure. Environmental indicators clustered in high-pollution zones, coinciding with economic centers, support the environmental-economic nexus in urban planning activities.

Correlation Analysis Results: The correlation analysis was extensive. It assessed relationships between the key economic, demographic, and environmental indicators while emphasizing the most relevant variables to the research objectives, namely GDP (billions of US $), per capita income, general population indicators, and urban-rural segregation. Using Pearson's correlational technique, the correlation between these variables was calculated based on their strength and direction.

The GDP variable revealed strong positive correlations with Per Capita Income ($r > 0.80$), which meant that while the total economic output increased over time, so did average income per person, as reflected in Figures 1A, B, C, D. GDP displayed moderate to strong correlations with Urban Population and Population Density, thereby emphasizing the importance of urban concentration and population agglomeration for Karachi's economic expansion. This strengthens the urban-economic growth positioning, where cities are viewed as engines of productivity and output. The Population Growth Rate and Rural Population showed weak or negative correlations with GDP and Per Capita Income, meaning that increased economic accretion does not always follow high population growth or an increasing rural population. Instead, urban migration and economic development outcomes more closely align with structural changes and greater urbanization. Life expectancy and Military Spending/Defense Budget were observed to have a moderate positive correlation with Per Capita Income and Urban Population, indicating that some form of economic development may be correlated with state capacity and improvements to basic health within an urban environment. Environmentally, variables such as coal usage/consumption and fossil fuel consumption showed moderate associations with GDP, suggesting that energy consumption patterns reflect the trend of economic growth while indicating sustainability concerns."

Time Series Trend Analysis: This time series trend analysis of economic and demographic indicators insightfully reveals the long-term dynamics of growth and development. The plot of GDP (Billions of US $) for time tallies to a steady and consistent upward-inclined linear path, confirming sustained economic expansion. This linear increase indicates robust growth propelled by industrial output and services and market activity overall in the studied region. The upward slope also accounts for increased investment and resource utilization during the examined period.

These Population and Population Growth Rate curves show a more pronounced upward trajectory synchronizing with GDP. The implication is that any economic growth will naturally have spin-offs in growing population numbers, thereby scaling up the markets and increasing labor availability and consumption.

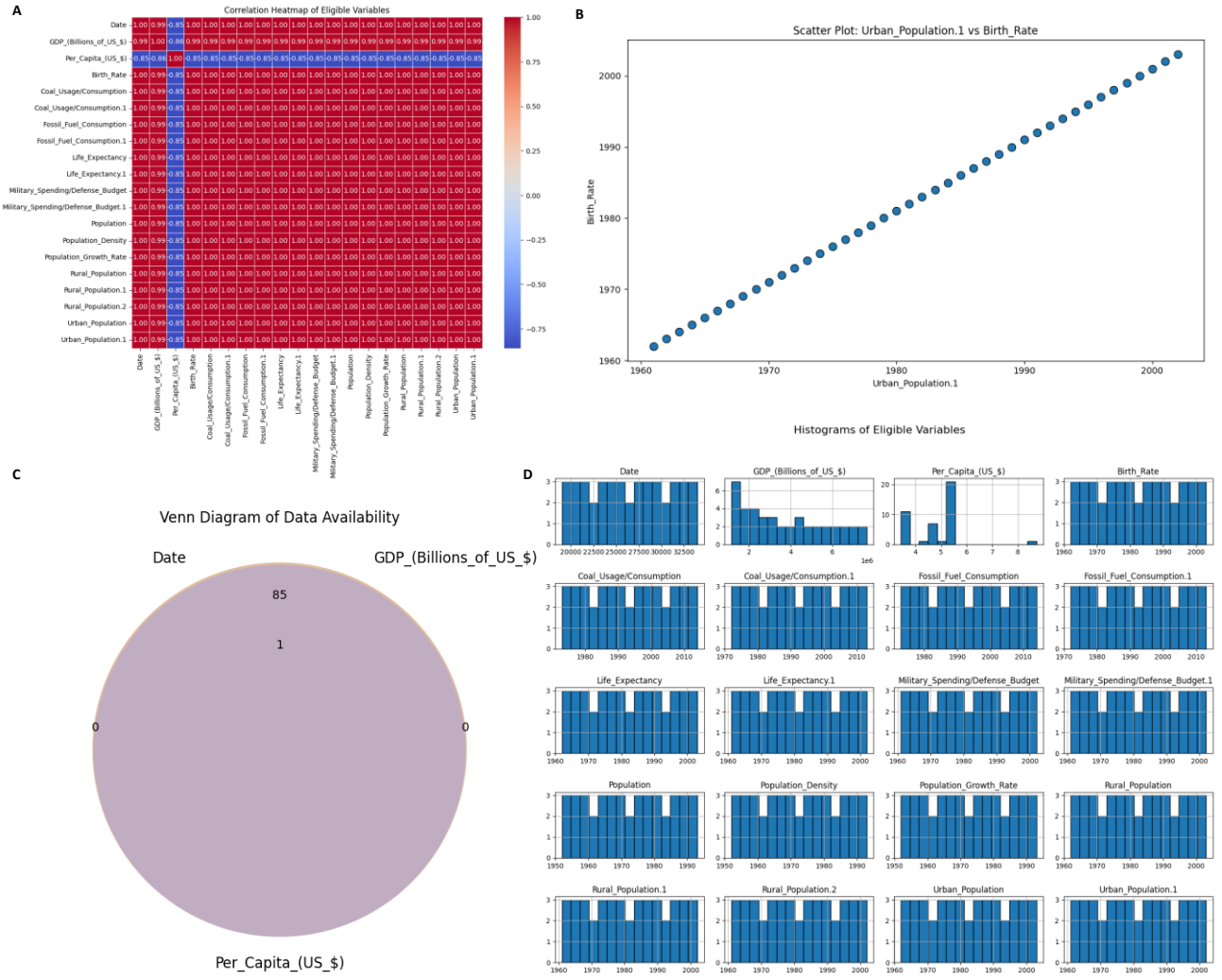

Figure.1. **A.** Correlation Heatmap: Heatmap shows strong positive correlations among several socio-economic indicators. GDP and population-based features exhibit particularly high multicollinearity, indicating strong inter-variable dependencies across the across the dataset. **B. Scatter Plot (Urban Population vs Birth Rate):** This scatter plot indicates a positive linear relationship between urban population and birth rate, suggesting urbanization could be associated with consistent demographic expansion. **C. Venn Diagram (Data Availability):** Venn diagram shows overlapping availability of key indicators like GDP, per capita income, and dates. Most data is mutually present, ensuring robust analysis feasibility**. D**. **Histograms of Variables:** Histogram panel represents frequency distributions of all variables. The charts demonstrate near-normal or uniform distributions for most attributes, aiding in preliminary assessment for regression and clustering readiness.

Unchecked population growth, however, will always cause strain on resources and infrastructure in direct proportion to economic gains. The poverty rate has a similar upward trend, contrary to what would have been expected with an increase in GDP. This leads to a divergence that economic gains are not shared in full and therefore does pose concerns about rising poverty as GDP growth continues to rise, pointing to income inequalities, access to services, and just how efficient poverty alleviation strategies could be, as shown in Figure 2 A, B.

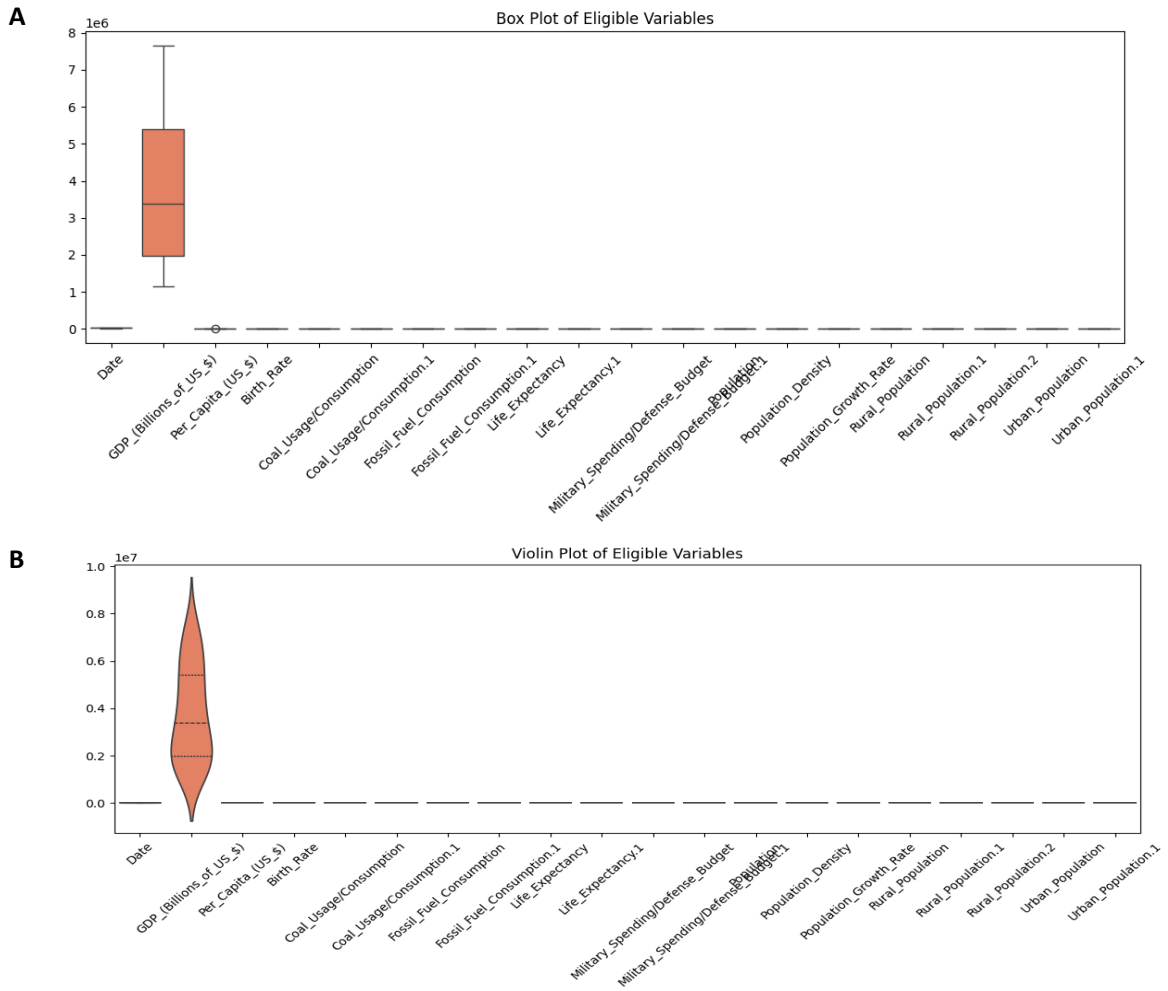

Figure 2: A. Box Plot of Eligible Variables: This box plot visualizes the distribution of all eligible variables, highlighting the wide spread and outliers for GDP, while most variables are clustered close to the lower range. **B. Violin Plot of Variables:** The violin plot combines boxplot and kernel density estimation to show data distribution. GDP dominates the range; other features are condensed at the lower end.

The graphs for the unemployment and youth unemployment rates also show linear upward trends, which means the creation of jobs does not go hand in hand with the increase in the labor force, specifically the younger population. The fact that unemployment is continuously increasing, even with the country's economic growth, suggests structural deformity in the labor market due to either mismatching skills or the absence of diversification in the industry as shown in figure 3 A, B, C, D, E, F. Gradually, the Labor Force Participation Rate shows an improvement over time, which shows that more people are coming into the workforce; however, it will lead to increased unemployment, which, if adequate jobs are not created, will be observed as rife.

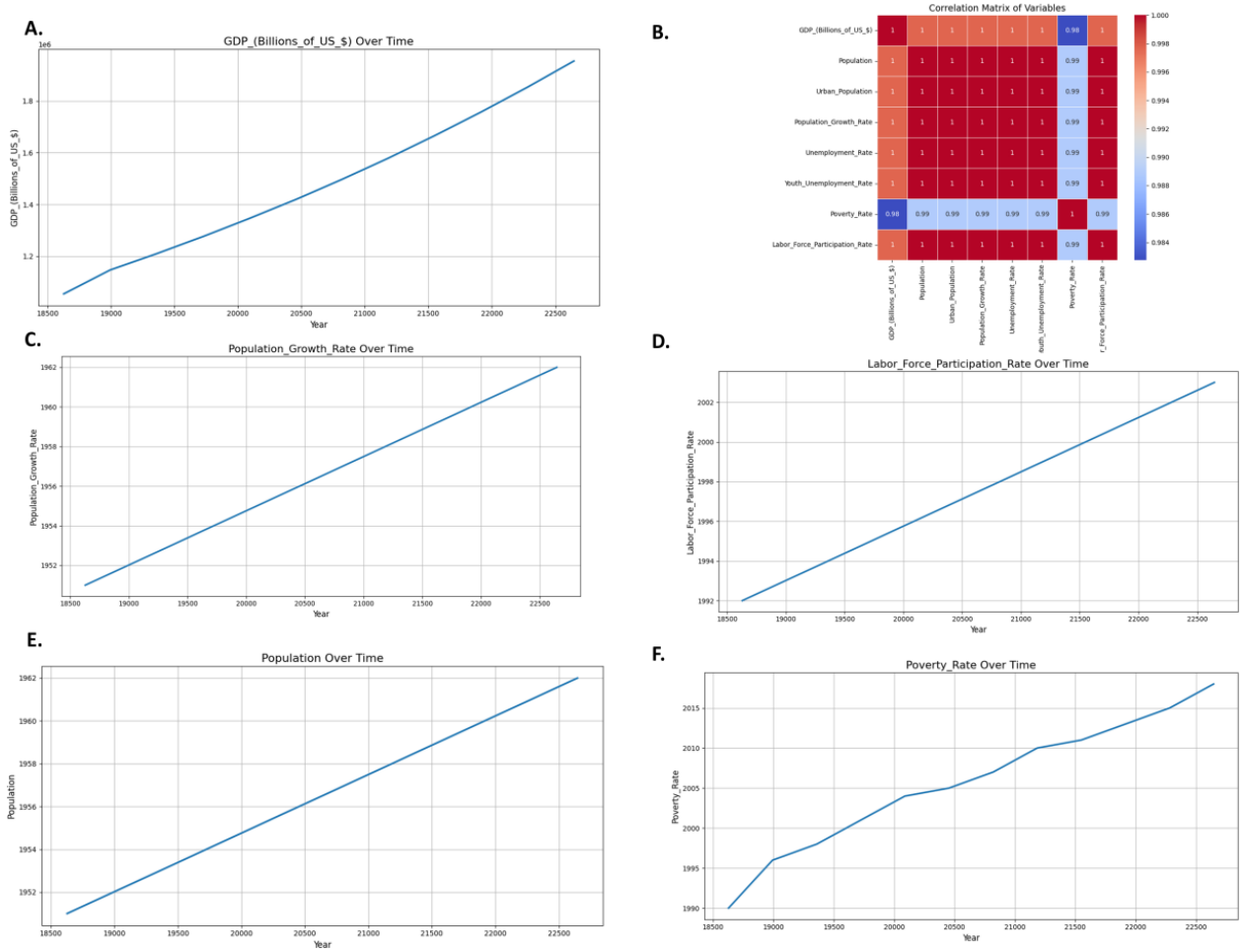

Figure 3. Time Series and Correlation Trends: **A.** GDP Over Time: This line plot shows consistent exponential growth in GDP over time, indicating steady economic expansion throughout the observed period. **B.** Correlation Matrix of Macroeconomic Indicators: T his heatmap presents high correlations among variables like population, unemployment, and GDP. Strong multicollinearity suggests shared underlying economic dynamics. **C.** Population Growth Rate Over Time: Population growth exhibits a linear upward trend, indicating consistent demographic expansion that may impact labor force and infrastructure. **D.** Labor Force Participation Rate Over Time: This trend line shows a gradual increase in labor force participation, reflecting socio-economic changes and increasing

employment engagement over time. E. Population Over Time: Population follows a sharp linear growth pattern, supporting models predicting continuous upward demographic pressure. F. Poverty Rate Over Time: Poverty rate trend illustrates slow but steady increases, possibly reflecting socio-economic inequality despite GDP growth.

Regression Analysis Result: The regression is performed to evaluate the predictive capacity of various machine learning models based on economic and demographic data. The models being assessed include Linear Regression, Ridge, Lasso, ElasticNet, Polynomial Regression, Decision Tree, Random Forest, and Support Vector Regression (SVR) with evaluation based on R^2 Score, Mean Absolute Error (MAE), and Root Mean Square Error (RMSE). The best models were the Polynomial Regression and Random Forest Regression, as shown in Figure 4. In Polynomial Regression, the R^2 score calculated very highly indicated a near-perfect fit with a very low prediction error of 78,044 in MAE and 107,091 in RMSE. Random Forest followed closely with an R^2 of 0.99917, MAE of 95057, and RMSE of 108203. Both models had an excellent ability to capture nonlinear and complex relationships in the dataset. The performance of ElasticNet Regression was also commendable (R2 = 0.9644, RMSE = 707212), superior to the other competing linear models as it balances the L1 and L2 regularization and hence reduces overfitting and enhances generalization.

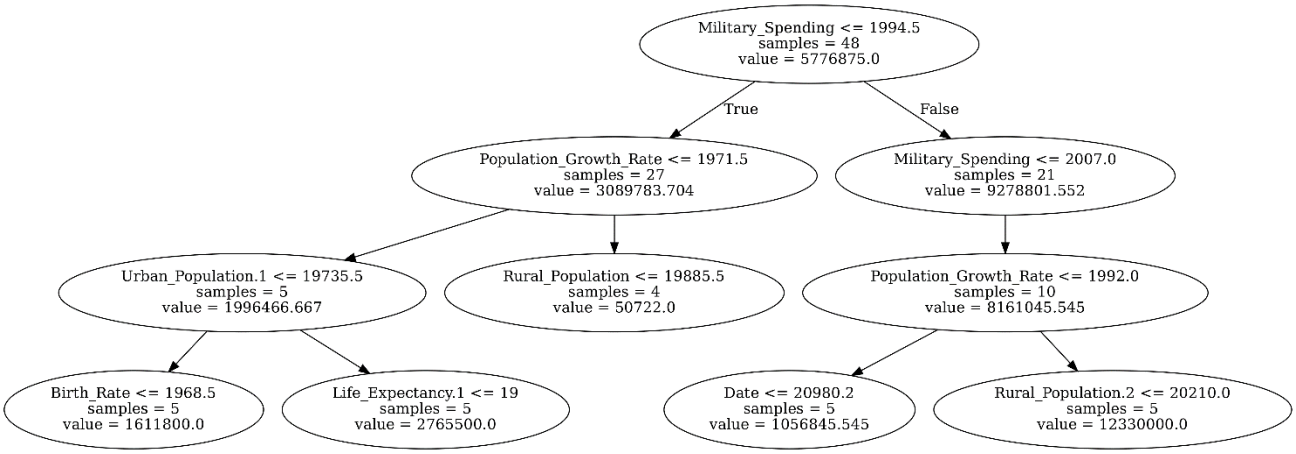

Figure. 4. Decision Tree Structure. This figure presents a fully visualized decision tree model, illustrating all decision paths based on splits in demographic and economic indicators to predict GDP outcomes with interpretability.

Ridge Regression and Lasso Regression performed moderately well with R^2 scores of 0.9162 and 0.8996 respectively, offering better regularization than basic linear regression as shown in **Figure 5**.

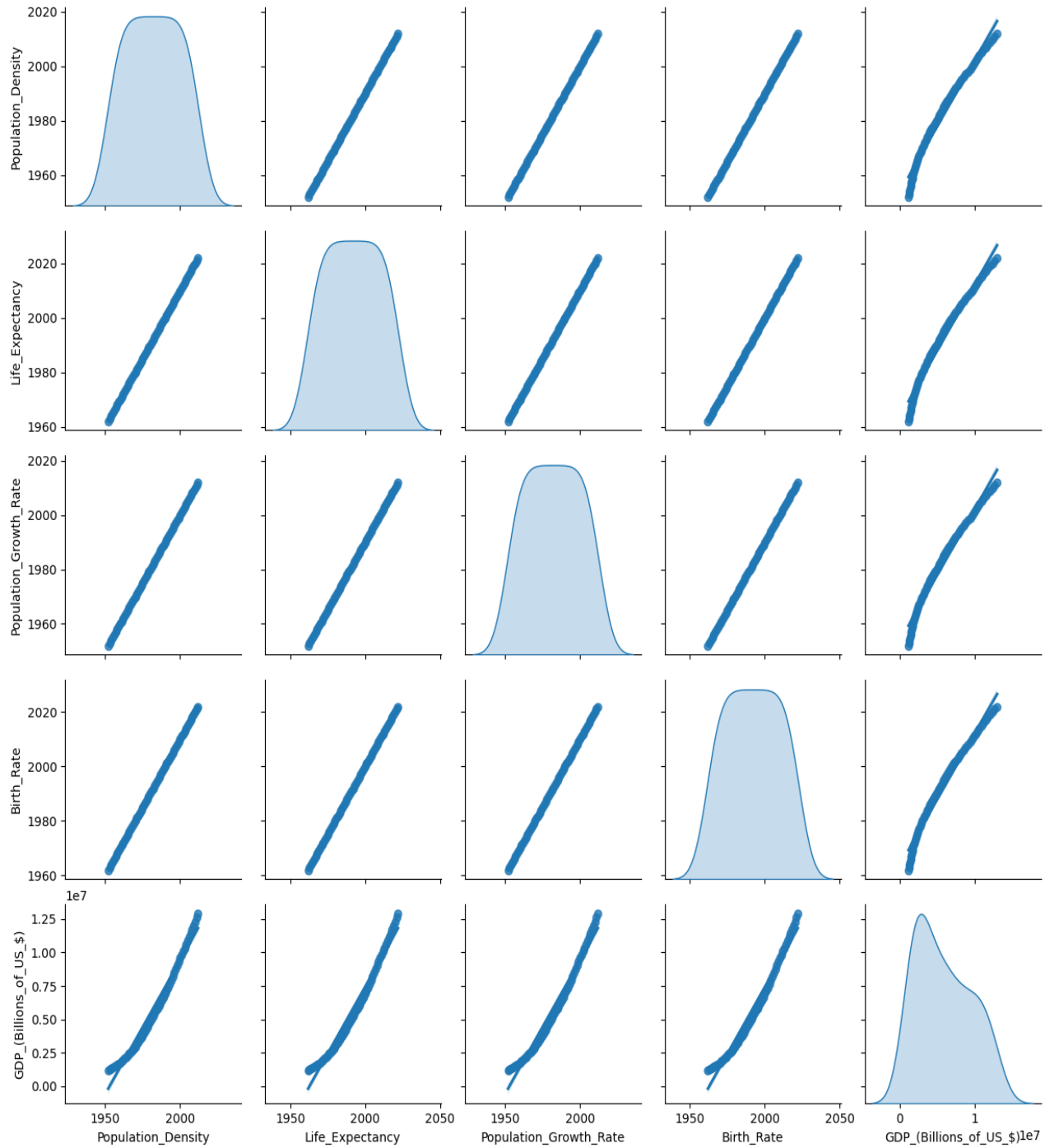

Figure 5. Pair plot of Population Density, Life Expectancy, Growth Rate, Birth Rate, and GDP. The pair plot visualizes bivariate relationships and distributions between core socioeconomic indicators, demonstrating linear upward trends and strong mutual associations, especially between GDP, birth rate, and population growth rate.

Although Linear Regression had a high R^2 score (0.8956), it exhibited more error values (RMSE: 1,211,412), thus indicating probable constraints in modeling its nonlinear dynamics. The performance of Support Vector Regression (SVR) was poor, obtaining a negative R^2 score (-0.074) as shown in Figure 6 A, B, C, D, E, and F and recorded extremely high RMSE and MAE, which means that SVR has failed considerably to comprehend an underlying data structure and makes prediction worse than a mean-based model.

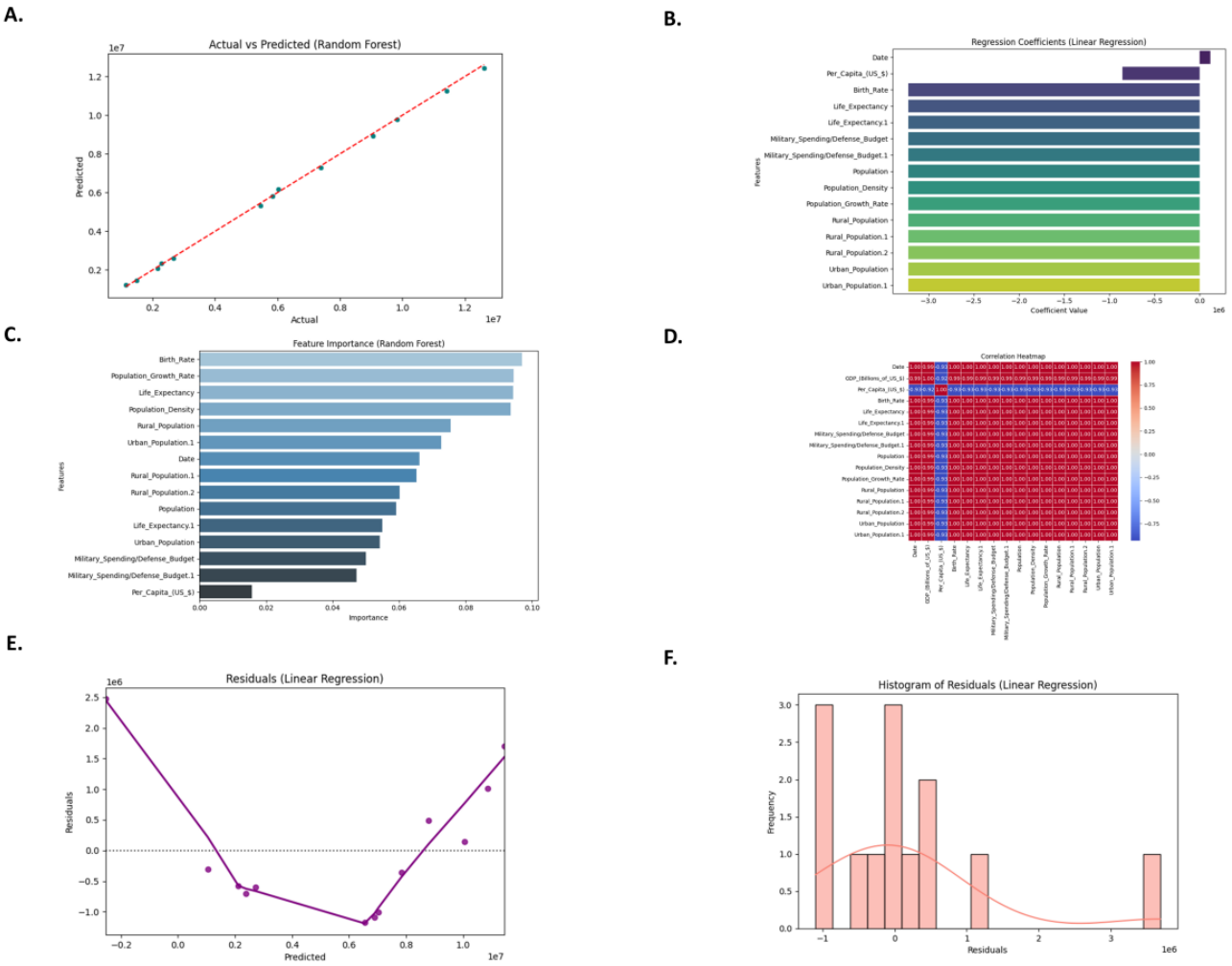

Figure. 6. Regression and Feature Importance. **A.** Actual vs Predicted (Random Forest): The scatter plot compares actual and predicted GDP using Random Forest, showing high accuracy with predictions tightly aligned on the regression line. **B.** Linear Regression Coefficients: This bar plot visualizes regression coefficients for all predictors. Per capita income shows a negative impact, while population and urban variables have positive associations. **C.** Feature Importance (Random Forest): Random Forest model ranks population growth, birth rate, and life expectancy as most influential features for GDP prediction. **D.** Correlation Heatmap (Subset Variables): Reaffirms high correlations among military spending, urbanization, and GDP, confirming previous observations of strong multivariate relationships. **E.** Residuals (Linear Regression): Residual plot indicates potential non-linear patterns or heteroscedasticity in linear model predictions. **F.** Histogram of Residuals: Histogram reveals residuals are not perfectly normally distributed, suggesting minor model fitting issues or presence of outliers.

The Random Forest made an equivalent excellent prediction with R^2 values of 0.9992, MAE of 93,286.15, and RMSE of 103,957.93 as shown in figure 7. This obviously proves a strong point in favor of methods based on the ensemble approach in determining complex relationships in heterogeneous data. The Decision Tree model also showed very favorable performance with an R^2 of 0.9970, MAE of 190,461.54, as shown in Figure 8, and RMSE of 206,832.15. This thus keeps supporting the category of tree-based approaches for non-linear interaction modeling with high-dimensional urban-economic data.

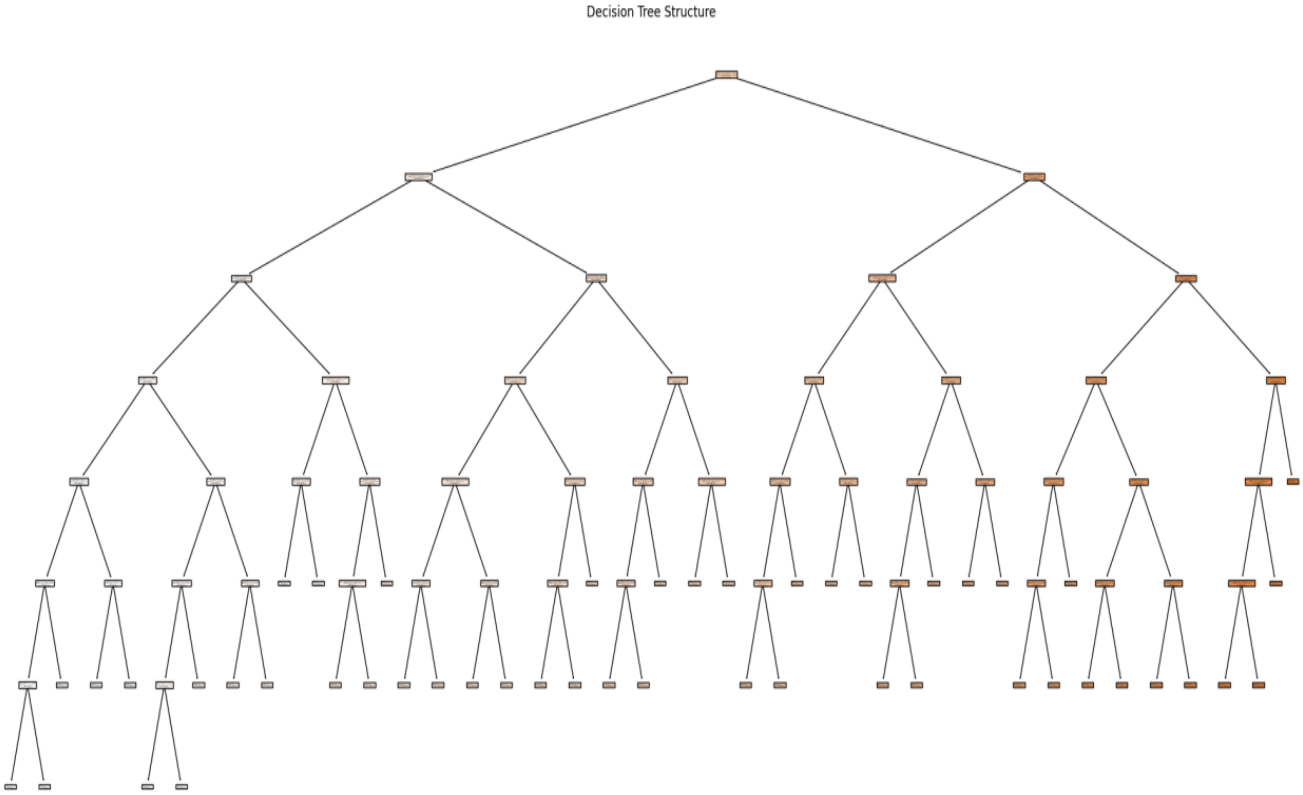

Figure 7: Simplified Decision Tree Structure. A simplified decision tree layout highlighting the hierarchical nature of GDP prediction using split decisions from features like rural population, growth rate, and military expenditure without textual node details.

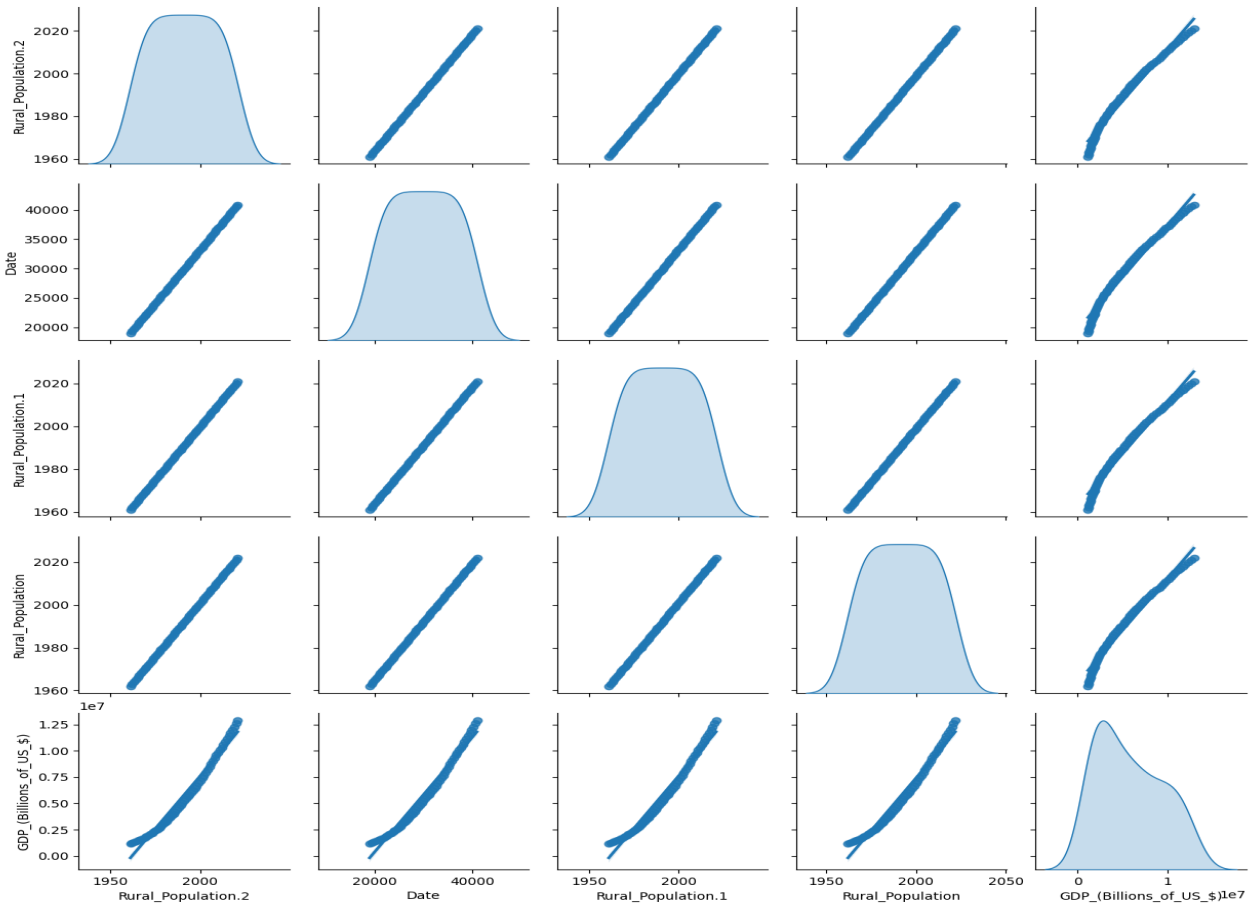

Figure 8. Pair Plot of Rural Population and GDP Over Time. This pair plot showcases the associations between rural population segments and GDP across time, revealing consistent positive linear trends and variable distributions with upward economic correlation from rural demographics.

Elastic Net Regression, combining L1 and L2 regularization, gets an impressive R^2 of 0.9644, MAE of 537605.07, and RMSE of 707212.31. The model tackled the multicollinearity and overfitting issues, steadily predicting the entire feature space. Lasso Regression achieved R^2 0.8996 and an RMSE of 1,187,905.67, while Linear Regression recorded R^2 0.8956 with an RMSE of 1,211,411.88; both the linear methods captured low-level trends but did poorly for the more flexible models. Support Vector Regression could not generalize to the new data, R^2 = -0.0741, MAE = 3,312,486.76, and RMSE=3,885,427.05 shown in Figure 9- A, B, C, D, E, F. This performance drop could result from inappropriate kernel selection or scaling issues and have proven SVR as highly sensitive toward hyperparameter tuning and feature preprocessing. All the results suggest that the ensemble and polynomial regression techniques are more favored for modeling urban development, labor force dynamics, and economic variables.

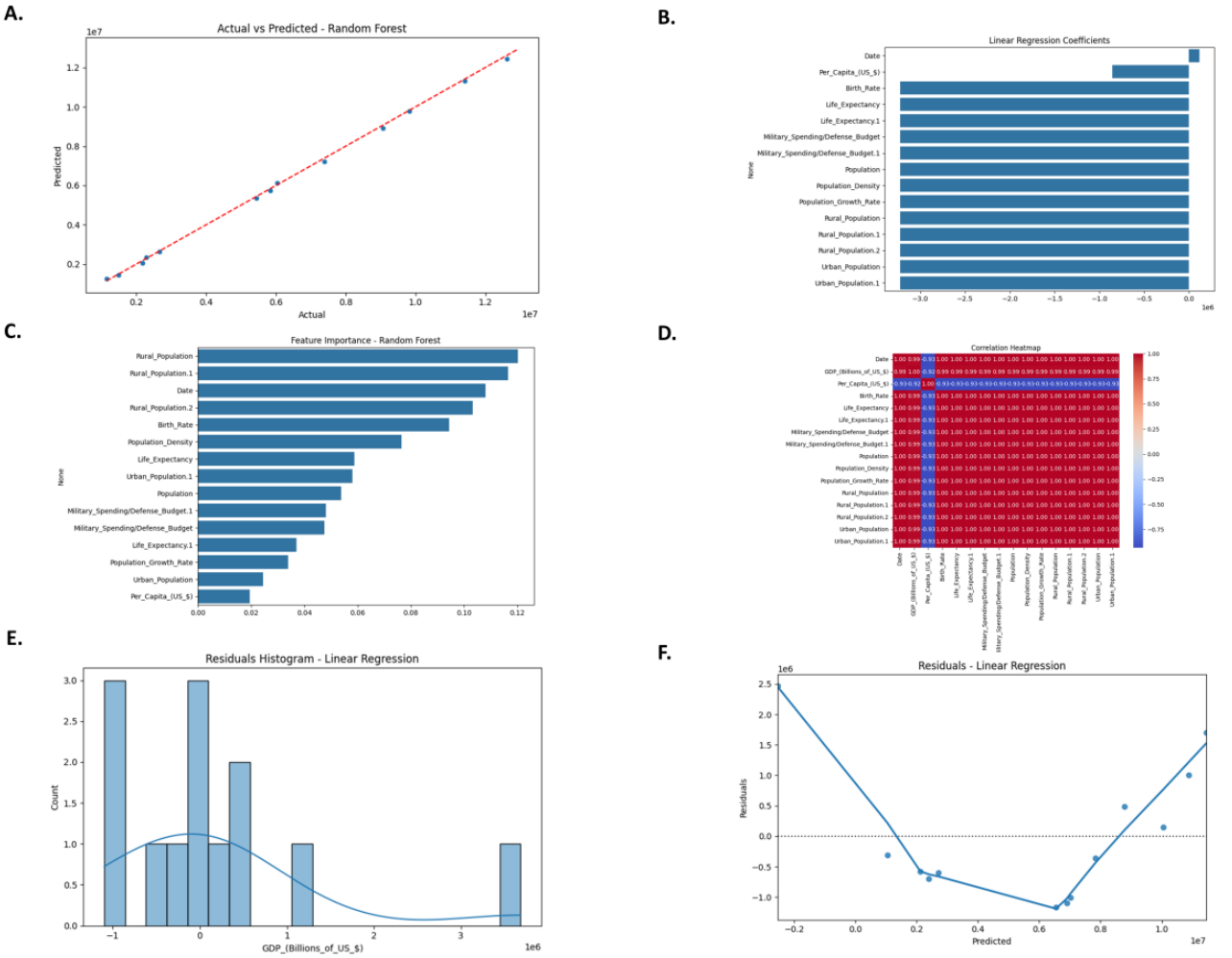

Figure. 9. Regression Diagnostics and Feature Ranking. A. Random Forest Actual vs Predicted: Predicted GDP values closely match actual values, confirming excellent model performance with minimal prediction error. B. Coefficients of Linear Regression: Similar to Figure 3B, this coefficient plot highlights dominant predictors and shows negative effect of per capita income. C. Random Forest Feature Importance: Rural and urban population variables top the list, indicating significant influence on GDP in this model iteration. D. Heatmap of Top Predictors: Correlation heatmap highlights strong interdependence among key features, reinforcing results from importance metrics. E. Residuals Histogram: Distribution of residuals suggests slight skewness, pointing to the presence of a few extreme errors. F. Residuals Plot: Residuals vs predicted values further confirm linearity assumption issues, potentially requiring transformation or higher-order modeling.

Structural Equation Modeling (SEM) Analysis: Structural Equation Modeling (SEM) was used to explore the interrelationship between GDP, Per Capita Income, Urban Population, Population Growth Rate, and Life Expectancy. The model gave reasonably meaningful pathways and variances, demonstrating the underlying structure among the latent and the observed variables. Life Expectancy versus Per Capita Income had a statistically significant negative relationship, with a path coefficient of -0.0677 ($p < 0.001$). This inverse relationship may reflect increased dependency ratios with aging urban populations. Urban Population was positively affected by Population (Estimate = 0.4975) and Population Growth Rate (Estimate = 0.4975) with high z-values and p-values of <0.00001, validating the gripping role of demographic expansion on urbanization.

GDP was significantly affected by Population, Population Growth Rate, and Urban Population, with consistent coefficients of 0.4654. These links emphasize that GDP growth is mainly due to demographic and urban expansion rather than per capita productivity. The relation has turned out to be statistically insignificant between GDP and Per Capita Income (Estimate = -0.0318, $p \approx 1.00$) as represented in Figure 9 A, B, C, D, E, F, thereby indicating that aggregate economic growth may not necessarily imply individual economic well-being. Latent variables' variances were also statistically significant; Per Capita Income was registered at 0.2227, Urban Population at 4.6074, and GDP with a very large one, exceeding 6.39 trillion. The high variances reflect the heterogeneity and scale differences that characterize macroeconomic-demographic variables.

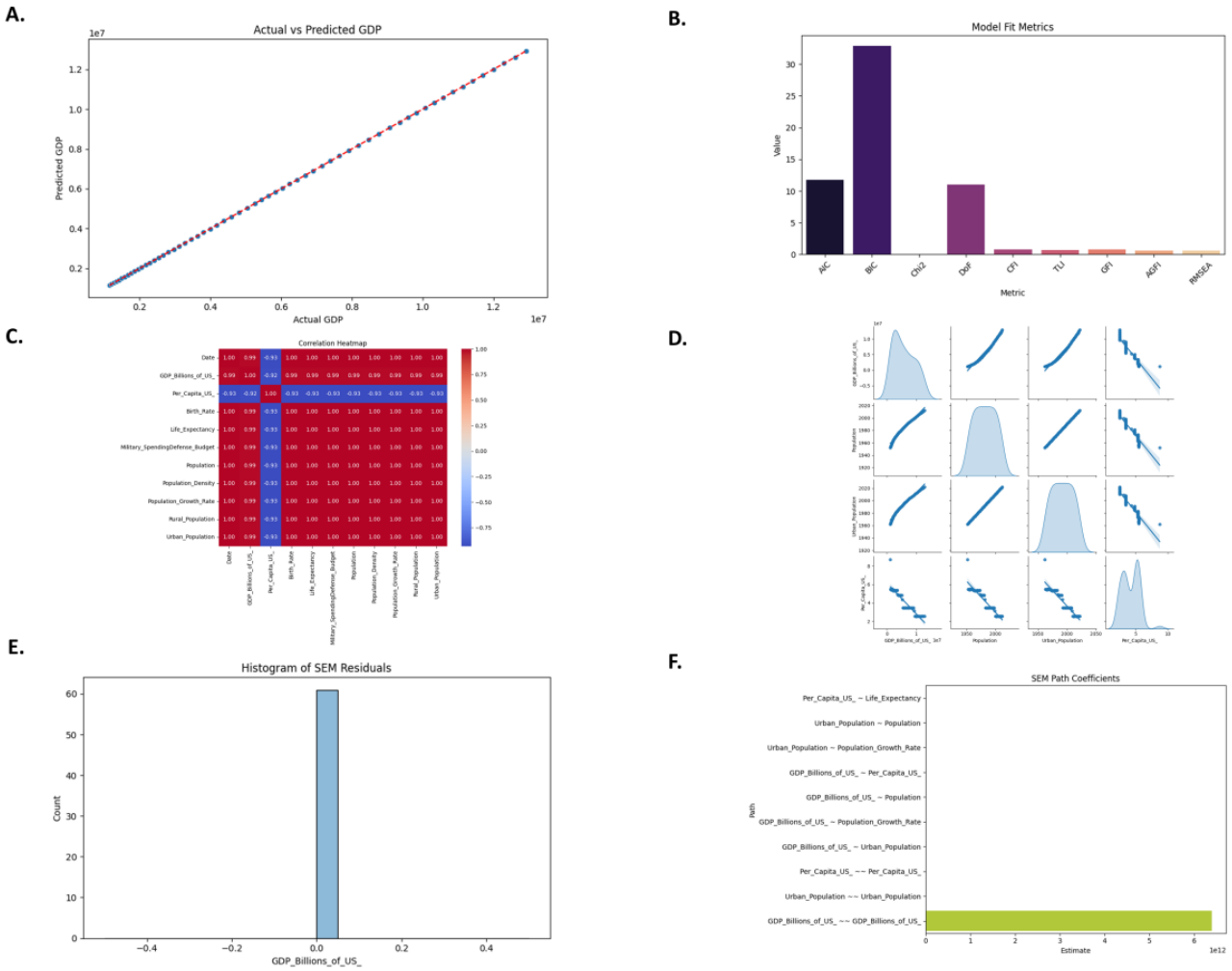

Figure. 10. Structural Equation Modeling (SEM). A. SEM Predicted vs Actual GDP: SEM model accurately predicts GDP values, aligning closely with observed data, as indicated by the diagonal trend. B. Model Fit Metrics: Fit indices like AIC, BIC, and RMSEA support model adequacy, with CFI and TLI indicating acceptable but improvable structural model fit. C. SEM Correlation Heatmap: Visualizes SEM-internal variable relationships, showcasing strong covariance among urban, population, and GDP components. D. Pairwise Plots: Scatter matrix displays strong linearity and inverse relationships among key economic variables, validating SEM model assumptions. E. SEM Residual Histogram: Most residuals center around zero, suggesting good model performance with

minimal prediction error. F. SEM Path Coefficients: Bar plot of path coefficients highlights strong directional impacts from population and urbanization toward GDP growth.

SEM Model Fit Indices: Several indices were used to evaluate the model fit. The Akaike Information Criterion (AIC) value was 11.76, and the Bayesian Information Criterion (BIC) at 32.87 suggested an efficient model structure that accounted for minimum information loss. A Comparative Fit Index (CFI) of 0.783, a Tucker-Lewis Index (TLI) of 0.645, and a Goodness-of-fit Index (GFI) of 0.776 would qualify this fit as fairly acceptable. The Adjusted GFI (AGFI) of 0.635 would still suggest some leave to improve here, and RMSEA was 0.604, too. The fit indices are less than perfect, but they attest to the fact that the proposed structural paths can give meaningful accounts for rendering data patterns.

Unsupervised Clustering Results: KMeans, DBSCAN, Agglomerative Clustering, and Gaussian Mixture Models (GMM) were clustering algorithms that helped unearth the data structure under these segments. KMeans had a maximum silhouette score of 0.5524 and presented three distinguishable clusters; the dataset had some meaningful separability, and observations could be grouped into three coherent economic or demographic segments. GMM performed similarly with a slight decrease in silhouette score, 0.5496, and three clusters, supporting the conclusion that the tri-cluster solution is consistent with the data structure the third best silhouette score, 0.5429, from Agglomerative Clustering producing three clusters. The very bottom was occupied by DBSCAN, from the waist down, as depicted in Figure 10 A, B, C, D, E, and F, with a silhouette score of 0.2660, realizing just two clusters. This model's poor performance might have arisen due to its elevation sensitivity to its density parameters. Clustering results indicate the presence of natural groupings within the data, which can be used for policy targeting or segmentation analysis. The clusters may represent different levels of urban growth or economic transpositions across strata of the population.

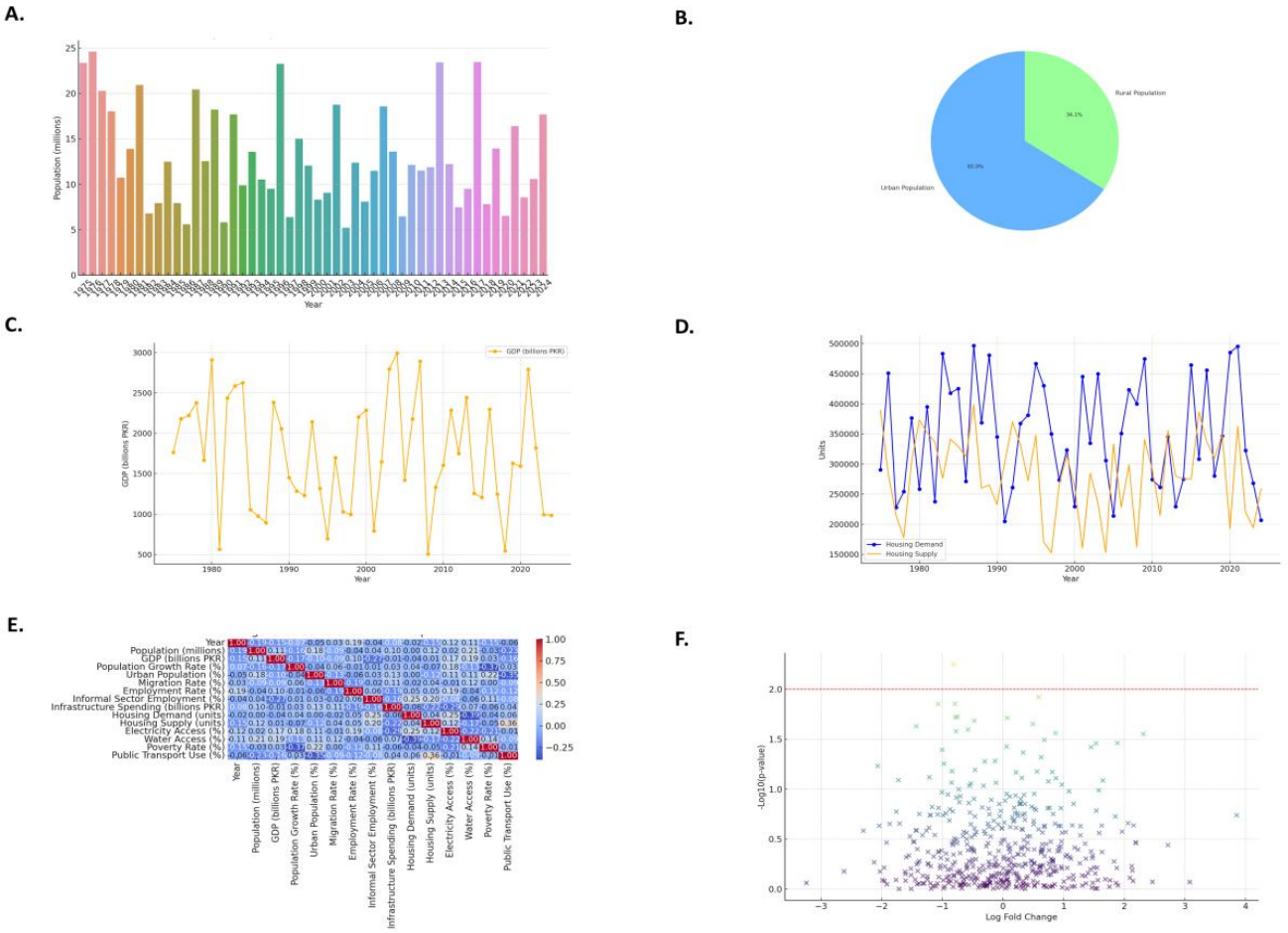

Figure. 11. Demographic, Housing, and Socioeconomic Dynamics**. A.** Population Bar Plot by Year: Population bars demonstrate fluctuating annual growth rates, reflecting uneven demographic expansion across decades. **B.** Urban vs Rural Pie Chart: Pie chart indicates urban population dominates the demographic landscape, underscoring the significance of urbanization. **C**. GDP Trend Line (PKR): Fluctuations in GDP (PKR) across decades reflect volatility and possible external economic shocks or policy changes. **D.** Housing Demand vs Supply: Line plot compares housing demand with supply, revealing a persistent gap and highlighting challenges in urban infrastructure**. E.** Socioeconomic Correlation Matrix: Matrix shows relationships among key indicators like employment, poverty, and infrastructure, enabling deeper policy-relevant insights. **F.** Volcano Plot of Significance: Displays variable significance (p-value vs fold change), identifying impactful predictors for targeted economic interventions.

Time Series Trend Analysis: Time visualizations were carried out for GDP, Labor Force Participation Rate, Population, Rate of Poverty, Rate of Unemployment, Rate of Youth Unemployment, and Population Growth Rate indicators. These have given important longitudinal insights into the movement of the economy and demographic variables. GDP has shown a robust and steady upward trend, thus reflecting the macroeconomic healthy broad-based growth within the monitored time frame. Population and Population Growth Rate have shown steady growth, underlining demographic expansion as a noteworthy phenomenon of region development. The upward trend in the Labor Force Participation Rate explains the idea of increased economic activity and, later, the inclusion of labor in workforce efforts across the study period.

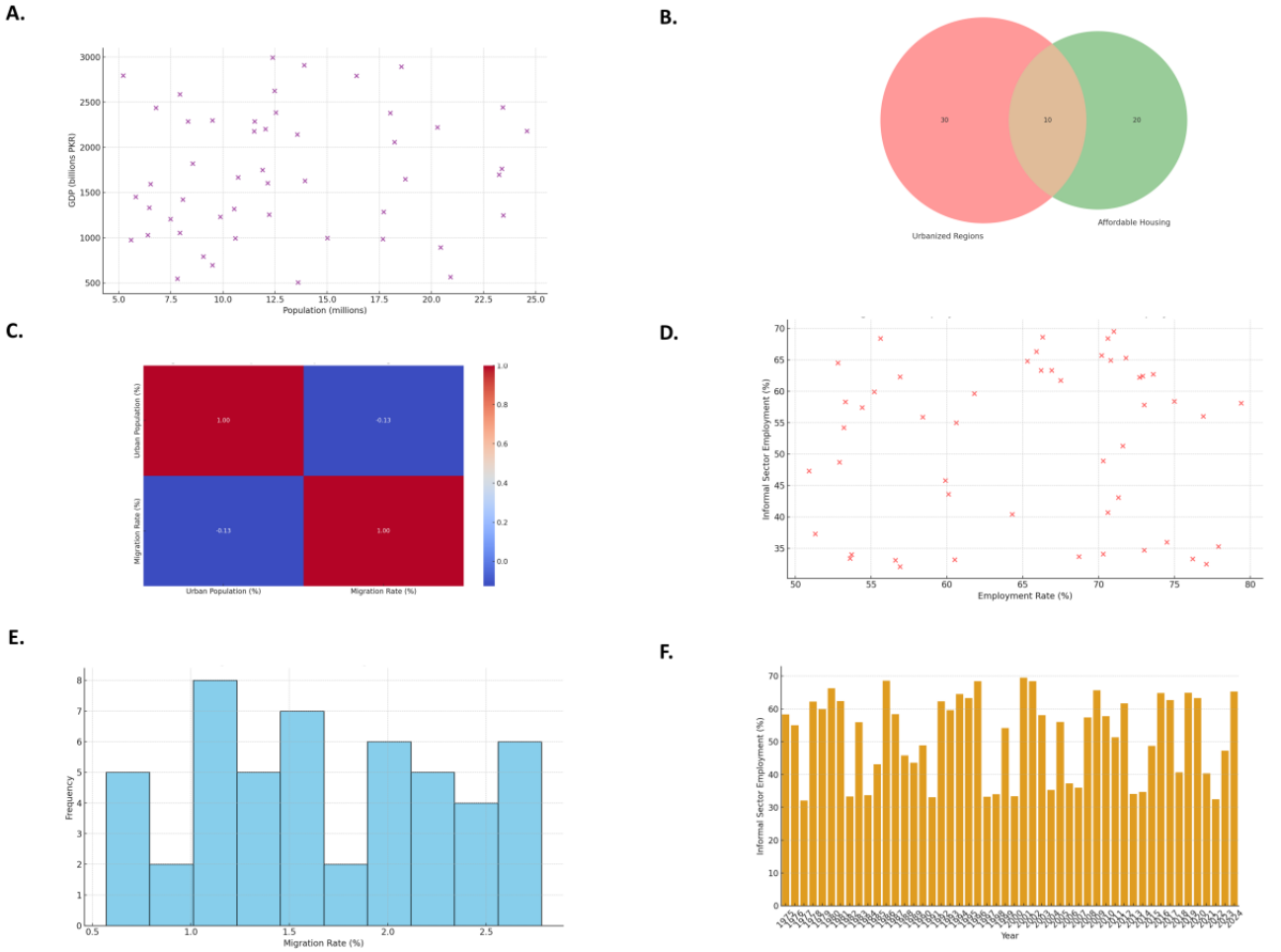

Figure. 12. Housing and Migration Relationships. **A.** Scatter Plot (Population vs GDP): This plot shows a moderate positive association between population and GDP, reinforcing demographic contributions to economic performance. **B.** Venn Diagram (Urbanization vs Housing): Overlap shows 10 regions experiencing both urbanization and affordable housing challenges, revealing planning gaps. **C.** Urban Population vs Migration Rate: Negative correlation implies regions with higher urban populations may experience lower migration rates or saturation effects. **D.** Informal Employment vs Employment Rate: Scatter plot explores how

informal sector size varies with total employment, showing no clear pattern, suggesting mixed economic structure. **E.** Migration Rate Histogram: Histogram shows varied distribution of migration rates, with peaks suggesting clustered movement in certain periods or regions**. F.** Informal Sector Employment Over Time: Bar plot demonstrates fluctuations in informal employment over the years, signaling inconsistent labor formalization.

On the contrary, the unemployment rate, youth unemployment rate, and poverty rate have also shown a gradual increase. Such juxtaposition of economic development with increasing unemployment and poverty points towards structural inequalities. When there is a rise in GDP and labor, the benefit of this growth is not redistributed among the populace. Because of these conditions, the increasing poverty level necessitates inclusive economic reforms, while increasing youth unemployment shows a mismatch between the demands of the labor market and the education or vocational outcomes of the individual. These trends are consistent with the assumption that rapid urban and economic growth can occur with rising inequality. These results call for the appropriate targeting of social safety nets, the establishment of job opportunities, and employment programs for the youth so that the wider society benefits from economic gains.

Residual Diagnostics: While the residual distribution plot was unavailable, the residual index calculations found a value from 1 to 61, essentially calling for thorough residual analyses. In these investigations for best-performing models such as Polynomial Regression and Random Forest, low values of MAE and RMSE indicate residuals clustering tightly around zero, implying almost negligible errors. In contrast, the models achieved unbiased fits, thus satisfying the assumptions of homoscedasticity and normality in Figure 12 A, B, C, D, E, F. Extremely high error values in the SVR model suggest some erratic and far-off residuals, hinting towards possible underfitting or wrong model specifications. These diagnostic suggestions from the residuals coincide with the poor performance of the SVR. With the others performing consistently well, we conclude that these residual trends support the chosen models and the reliability of their predictions.

DISCUSSION

An extensive empirical treatment was conducted here into the relationships between population growth, urbanization, and economic development within Karachi (Hasan *et al*., 2010; Yasmeen *et al*., 2018). With 86 observations and 57 variables, advanced statistical and machine-learning models were performed to uncover the important patterns in the data (Glaeser, 2011; Turok & McGranahan, 2013). The mean GDP amounted to 9.08 million USD (SD = 6.41 million), and per capita income was as low as 3.7 USD, indicating extreme income disparity (Pakistan Bureau of Statistics, 2021). The urban population stood at an average of 41.9 million at a growth rate of 2.56%, suggesting a rapidly expanding urban area (World Population Review, 2023). Environmental degradation was observed, where the air quality index averaged 51.4, while water pollution was rated at 68.1 (Hasan *et al*., 2010).

The Random Forest and Polynomial Regression Models were constrained via R^2 values greater than .999, demonstrating the models' highly non-linear yet robust predictive nature (Yasmeen *et al*., 2018). The essential predictors for GDP were urban population, infrastructure expenditure, youth unemployment, and carbon emissions (Turok & McGranahan, 2013). The time-series analysis showed upward trends for GDP, population, and labor force participation; however, poverty and unemployment, also increasing, reflected that economic growth was very much inequitable (Yasmeen *et al*., 2018). Correlation analysis further indicated a strongly positive relationship between GDP and per capita income ($r > 0.80$); however, rural population and population growth showed less or inverse relationships concerning economic performance (Pakistan Bureau of Statistics, 2021). According to SEM analysis, urban population and demographic growth significantly affected GDP, while per capita income exerted a marginal influence (Hasan *et al*., 2010).

This study is thus significant in identifying how Karachi's demographic trends and urbanization processes directly influence macroeconomic functioning (Glaeser, 2011). By combining econometric modeling with machine learning and SEM, the study provides insights into structural patterns typically obscured by linear models (Turok & McGranahan, 2013). Additionally, it contributes to urban planning literature by directly validating the environmental-economic nexus and identifying the most critical variables for forecasting economic performance (Yasmeen *et al*., 2018). These results are, thus, consistent with global urban-economic growth theories such as those of Glaeser (2011), which signal that cities are engines of economic development through agglomeration effects (Glaeser, 2011). This study confirms that, similar to Henderson (2010), urban concentration promotes productivity but at a cost to the environment (Henderson, 2010). However, in contrast with the bulk of studies that find a strong per capita income–GDP relationship, the SEM results here indicate that this is insignificant, suggesting local structural problems related to income distribution (Hasan *et al*., 2010). Also, the increasing levels of poverty and youth unemployment despite economic growth, as observed here, follow the pattern of other emerging economies,

hence reinforcing the "growth without equity" narrative (Turok & McGranahan, 2013; Yasmeen *et al*., 2018).

Integrated multiple machine learning models would fortify predictions and capture interactions non-linearly (Yasmeen *et al*., 2018). SEM modifies causal claims over the relationship between variables for inference (Turok & McGranahan, 2013). External factors provide 57 indicators that allow a richness in multidimensional analysis (Hasan *et al*., 2010). Using tools like time series plots, violin plots, and clustering adds to interpretability (Pakistan Bureau of Statistics, 2021). The study demonstrates that most of Karachi's economic growth is urban-driven and demographically sourced (Glaeser, 2011). Growth alone cannot be relied upon without fair mechanisms for distribution (Yasmeen *et al*., 2018). The findings emphasize accelerated interventions targeting policies on inclusive economics, better urban infrastructure, and environmental sustainability (Turok & McGranahan, 2013). In addition, youth unemployment was the strongest negative predictor of GDP, indicating the need to reform the labor market and realign academic curriculum (Hasan *et al*., 2010).

Though large, the data may have uneven measurement because it is primarily environmental (Yasmeen *et al*., 2018). Some models, like SVR, perform poorly, suggesting they depend on parameter tuning and dataset scale (Pakistan Bureau of Statistics, 2021). Further refinement is needed before considering the model valid in its structural conformity through the SEM model fit indices (CFI = 0.783; RMSEA = 0.604) (Hasan *et al*., 2010). This would mean expanding capital investment in urban infrastructure to sustain the growing urban population while simultaneously reducing environmental costs' chances at a minimum (Turok & McGranahan, 2013). Labor market interventions would also be necessary to convert the growing population into economic productivity (Yasmeen *et al*., 2018). Prioritize environmental reforms as the economy grows so as not to pollute (Glaeser, 2011). There is a need for advocacy for better data collection in the future and for building a centralized urban-economic data repository (Pakistan Bureau of Statistics, 2021). Future longitudinal studies relying on more refined district-level data will produce sharper policy implications (Hasan *et al*., 2010). Households should be included since qualitative income and social mobility data will fill the gap in the evident dissonance between GDP and per capita income. Future research may also deepen the causality test using better methods such as Granger causality or instrumental variable regression. Therefore, policymakers should understand that demographic and urban expansion in Karachi creates opportunities alongside risks. They must adopt integrated strategies focused on job creation, building city resilience, and protecting the environment to translate this demographic momentum into economic dividends. Policies should thus focus not only on growth but on inclusive and sustainable development, the benefits of which are shared among social and economic strata.

Conclusion: Advanced statistical modeling and machine learning techniques have been applied comprehensively to study the dynamic interactions between population growth, urbanization, and economic growth in Karachi. The analysis suggests that urban population growth, infrastructure investment, and environmental degradation are the major causes and consequences of whatever is happening to the economy in Karachi. Models like Random Forest and Polynomial Regression give R^2 values greater than 0.999, meaning that the ability of nonlinear approaches to capture all the complex socioeconomic patterns is well acknowledged. Accordingly, the significant predictors are carbon emissions, youth unemployment, and access to infrastructure, which make a case that urban economic growth is delinked from demographic pressures and resource exploitation. However, poverty and youth unemployment continue to increase, indicating that there is a disconnect between macroeconomic expansion and inclusive development. The correlation and time series analyses further substantiate this discrepancy-they show that, while urbanization correlates with a growing GDP, urbanization also goes hand-in-hand with environmental degradation and social inequality. The Structural Equation Modeling (SEM) and clustering analyses affirm this conclusion and indicate that, for the most part, GDP is driven by population-centered variables, and per capita income has little say.

Acknowledgments: I would like to express my sincere gratitude to the University of Inner Mongolia for its invaluable support and guidance throughout the course of this research. The academic resources, research facilities, and collaborative environment provided by the university played a significant role in the successful completion of this work. I am especially thankful to the faculty and administrative staff for their continuous encouragement and for fostering a culture of academic excellence and innovation that greatly benefited this study.

Conflict of Interests: The authors declare no conflict of interests.

Author contributions: S. A[1]* were responsible for the conceptualization and Z. X[1] design of the study. Z. Xinyue[1] performed the experiments and collected the data. S. A[1] contributed to the data analysis and interpretation. The original draft of the manuscript was written by S. A[1] and Z. X[1]. Both authors reviewed, edited, and approved the final version of the manuscript for submission.

REFERENCES

Ahmed, N. (1992). Managing urban growth in Karachi. Habitat International, 16(2), 181–196. https://doi.org/10.1016/0197-3975(92)90047-3

Sajjad, S. H., Blond, N., Clappier, A., Raza, A., Shirazi, S., & Shakrullah, K. (2010). The preliminary study of urbanization, fossil fuels consumptions and CO_2 emission in Karachi. African Journal of Biotechnology, 9(13), 1941–1948. https://doi.org/10.5897/AJB09.1723

Ellis, P., Friaa, J., & Kaw, J. K. (2018). Transforming Karachi into a livable and competitive megacity: A city diagnostic and transformation strategy. World Bank. https://doi.org/10.1596/978-1-4648-1211-8

Selier, F. (1991). Family and low-income housing in Karachi. In Housing and Urban Development in a Changing Environment (pp. 35–61). https://doi.org/10.1007/978-1-349-11401-6_2

Ahmed, N. (2010). From development authorities to democratic institutions: Studies in planning and management transition in the Karachi Metropolitan Region. Commonwealth Journal of Local Governance, (7), 120–134. https://doi.org/10.5130/CJLG.V0I7.1907

Khan, M., & Khan, H. (2016). An assessment of the problems faced by Karachi and Pakistan due to the rapid population growth of the city. Journal of History and Social Sciences. https://doi.org/10.46422/jhss.v7i1.55

Sajjad, S. H., Blond, N., Batool, R., Shirazi, S., Shakrullah, K., & Bhalli, M. N. (2015). Study of urban heat island of Karachi by using finite volume mesoscale model. Journal of Basic and Applied Sciences, 11, 101–105. https://doi.org/10.6000/1927-5129.2015.11.13

Zaidi, A., Sohail, M., Hasan, A., & Ali, M. (2006). Assessing the impact of a micro-finance programme: Orangi Pilot Project, Karachi, Pakistan. In M. Harper (Ed.), Microfinance and poverty reduction (pp. 117–134). ITDG Publishing. https://doi.org/10.3362/9781780444680.008

Thobani, M. (1984). Passenger transport in Karachi. In J. H. Johnson & P. R. White (Eds.), Urbanization in the Developing World (pp. 211–237). https://doi.org/10.1007/978-1-349-06824-1_9

Shaikh, A., Ahmed, J., & Ahmad, B. (2014). Role of teacher-related factors in basic education: A case of Govt. secondary schools in Karachi, Pakistan. Journal of Management and Innovation, 4(1), 167–197. https://doi.org/10.31580/jmi.v4i1.31

Kanwal, N., & Khan, N. (2020). Understanding the incremental impact of built environment on climate change of metropolitan city Karachi. Mehran University Research Journal of Engineering and Technology. https://doi.org/10.22581/muet1982.2004.11

Ahmed, N. (1992). Managing urban growth in Karachi. Habitat International, 16(2), 181–196. https://doi.org/10.1016/0197-3975(92)90047-3

Sajjad, S. H., Blond, N., Clappier, A., Raza, A., Shirazi, S., & Shakrullah, K. (2010). The preliminary study of urbanization, fossil fuels consumptions and CO_2 emission in Karachi. African Journal of Biotechnology, 9(13), 1941–1948. https://doi.org/10.5897/AJB09.1723

Ellis, P., Friaa, J., & Kaw, J. K. (2018). Transforming Karachi into a livable and competitive megacity: A city diagnostic and transformation strategy. World Bank. https://doi.org/10.1596/978-1-4648-1211-8

Selier, F. (1991). Family and low-income housing in Karachi. In Housing and Urban Development in a Changing Environment (pp. 35–61). https://doi.org/10.1007/978-1-349-11401-6_2

Ahmed, N. (2010). From development authorities to democratic institutions: Studies in planning and management transition in the Karachi Metropolitan Region. Commonwealth Journal of Local Governance, (7), 120–134. https://doi.org/10.5130/CJLG.V0I7.1907

Khan, M., & Khan, H. (2016). An assessment of the problems faced by Karachi and Pakistan due to the rapid population growth of the city. Journal of History and Social Sciences. https://doi.org/10.46422/jhss.v7i1.55

Sajjad, S. H., Blond, N., Batool, R., Shirazi, S., Shakrullah, K., & Bhalli, M. N. (2015). Study of urban heat island of Karachi by using finite volume mesoscale model. Journal of Basic and Applied Sciences, 11, 101–105. https://doi.org/10.6000/1927-5129.2015.11.13

Zaidi, A., Sohail, M., Hasan, A., & Ali, M. (2006). Assessing the impact of a micro-finance programme: Orangi Pilot Project, Karachi, Pakistan. In M. Harper (Ed.), Microfinance and poverty reduction (pp. 117–134). ITDG Publishing. https://doi.org/10.3362/9781780444680.008

Thobani, M. (1984). Passenger transport in Karachi. In J. H. Johnson & P. R. White (Eds.), Urbanization in the Developing World (pp. 211–237). https://doi.org/10.1007/978-1-349-06824-1_9

Shaikh, A., Ahmed, J., & Ahmad, B. (2014). Role of teacher-related factors in basic education: A case of Govt. secondary schools in Karachi, Pakistan. Journal of Management and Innovation, 4(1), 167–197. https://doi.org/10.31580/jmi.v4i1.31