**Proposal**

**Influence on GDP, can we generalize it?**

**Artimes Rashidi Torghi**

**Introduction**

Low-income countries face huge economic challenges and financing needs. They rely on international institutions, including the IMF's (The International Monetary Fund) Poverty Reduction and Growth Trust for financial support, stronger countries contribute to the funding of this support so most of the countries involve and related to each other economics and improving other countries economically, it is concern for developed countries.

GDP is important because it gives information about the size of the economy and how an economy is performing. 46 nations are least developed in 2021 (worldpopulationreview.com) and The International Monetary Fund (IMF) classifies 152 countries as developing nations (finance.yahoo.com). It is crucial detecting the reasons of being developed or how different economics indicator can influence GDP. So, in this research, I will find which economic indicator are more influence GDP and try to generalize it for developing countries. For this purpose, some analyses are needed to track growth and poverty reduction (Bird, G. (2004). Growth, Poverty and the IMF). In this research, I explore different economic indicator from OECD’s country which These countries accounted for 42% of Global GDP and 27% of global GDP growth in the past 10 years (2012-2022) with population of 1.4 billion. to find out their effects and relationship with GDP, I use different variables such as Population and labor related data for stronger prediction, which Ivan Kitov (Kitov, I. O. (2008). GDP growth rate and population) showed Real GDP growth rate in developed countries is a combination of two terms: the first term represents the reciprocal of the duration of income growth with work experience and the second term accounts for the influence of changes in the population's age distribution on economic growth. Despite the Onur Sunal and Özge Sezgin's (Sunal, O., & Alp, O. S. (2016). Effect of different price indices on linkage between real GDP growth and real minimum wage growth in Turkey) research finds no causality between real minimum wage growth rates, and real GDP growth in both directions in Turkey, I use wages related variables to capture the relationship between these variables and GDP, so I can expand the results to other countries. Unfortunately, there is a relationship between economics’ data availability and developed countries. More developed countries tend to have more accurate data. By building a model based on these developed countries, I will use these data to expand my result and using it for undeveloped countries. I also use taxation policy variables to improve my model (Rachid Bahloula's study (BAHLOULA, R. (2022). Estimating Optimal Level of Taxation for Growth Maximization in MOROCCO) suggests that there may be an optimal level of tax revenue as a percentage of GDP that maximizes economic growth.)

Understanding the connection between gender-related factors and GDP growth is critical for achieving gender equity and sustainable economic development. R. A. Ahmed and Ashfaq Ali Shafin's (Ahmed, M. R., & Shafin, A. A. (2020). Statistical and machine learning analysis of impact of population and gender effect in GDP of Bangladesh) research explores the correlation among different gender factors and GDP growth. This aspect of the research highlights the broader socio-economic implications of labor market and social welfare policies. I can broader my results with gender employment and population differences among these counties, I also include immigration population variable for each country per year, there might be a bias for unauthorized immigrants and their employment. For further research, it is insightful to take an account for labor supply and tendency to work for Mixed-Status immigration (in Immigration Enforcement and Labor Supply paper by Joaquin Alfredo-Angel Rubalcabaa Jose R. Bucheli, Camila Morales (MORALES, N. Immigration Enforcement and Labor Supply: Hispanic Youth in Mixed-Status Families) shows the behaviors of mixed-status households in response to immigration enforcement actions) because it can affect labor supply and it would be helpful for policy implication for policy makers when the relationship between them affects GDP specially when most of the developed countries affected by immigration, both in labor’s supply and immigrants expenses. By using this variable, based on the model it might be helpful for some developing countries receives immigrants.

Economic forecasting is a very important aspect that policymakers in the financial and corporate organization rely on, because helps them to determine future events that might affect the economy and the citizens at large. The expansion of the latest forecasting patterns was important to address these relationships. Hence, this research examines the effect of these variables on GDP. To achieve this, I use different machine learning methods as result obtained from the study shows that machine learning is a better model to use in economic forecasting for quick and reliable data to avert future events (Paruchuri, H. (2021). Conceptualization of machine learning in economic forecasting)

In the next section, I show different machine learning methods to capture desirable results and how can distinguish between different results and how it helps to evaluate the research. As usual in machine learning methods, I am partitioning the data to test and train set so the trained model can evaluate the unseen data in the test set.

**Empirical Method**

In this research, I employ a comprehensive empirical method that combines insights from existing studies with my research design to investigate the impact of labor market policies, taxation policies, and social welfare programs on Gross Domestic Product (GDP) in OECD countries from 2003 to 2019. To achieve this, I draw on the findings and methodologies of prior research, incorporating machine learning techniques such as K-Nearest Neighbors (KNN), Support Vector Regression (SVR), and XGBoost, which have shown promise in capturing complex relationships within data.

**Data Collection and Description:**

I begin by gathering a rich dataset that spans various economic and demographic variables, including GDP, population data, labor market statistics, taxation policies, and social welfare programs. This dataset has information from OECD countries over a period of 17 years, allowing for a thorough analysis of the dynamics between these variables and GDP. The variables to be considered include:

GDP, Population, Long term interest rate on government bonds, Labor force, Labor force by females, Unemployment rate, Migration population each year, Contribution to world trade volume, Total direct taxes, Social security benefit by government, Real minimum wages, Short term interest rate, Hours worked per employee, Labor productivity of the total economy, Wage rate, Unit labor cost in total economy, Working age population, Healthcare quality indicators, Healthcare expenditure, Inflation rate, Trade balance

These variables are crucial for forecasting GDP because they collectively represent key economic, demographic, and policy related factors that impact a country's economic performance.

**Data Partitioning:**

To ensure robust model training and evaluation, I divided the dataset into training and testing sets. The training set, typically comprising 70-80% of the data, is used for model development and fine-tuning. The testing set, containing the remaining 20-30%, serves as an independent dataset for assessing model performance. In my approach, I choose to exclude the last two years from the training set, using them as a test set to evaluate model forecasting accuracy.

I used K-Nearest Neighbors (KNN) as one of my primary modeling methods. Followed by SVR and XGboost, to build a model on training set and then based on the model’s results, I evaluate these conducted models and test them on the testing set.

For conduct better performed model, to avoid underfitting and overfitting, I use Cross validation in training set to setup hyperparameters.

**KNN:**

Inspired by the research that find the machine learning K-Nearest Neighbors(KNN) model captures the self-predictive ability of the U.S. GDP and performs better than traditional time series analysis(GDP Forecasting: Machine Learning, Linear or Autoregression?, Giovanni Maccarrone, Giacomo Morelli, Sara Spadaccini), I applied KNN to find the similarity of each variable among different countries.

KNN doesn't make specific assumptions about the underlying data distribution, which makes it a non-parametric algorithm. It relies on the idea that data points that are close to each other in feature space are likely to have similar target values and it’s works by finding the K nearest data points to a given input point in feature space and then making predictions based on the average of the target values for regression tasks among these K neighbors.

For KNN, need to select the number of nearest neighbors, K. A small K might lead to overfitting (highly influenced by outliers or noise), while a large K might lead to underfitting (average value). The choice of K depends on the data and should be determined through cross-validation technique. Because KNN measured distance, Standardizing or normalizing the continuous variables is crucial because KNN is sensitive to the scale of the features.

the predicted GDP values are computed as the average or weighted average of GDP values based on the k nearest neighboring data points using both Manhattan and Euclidean distances. First, I calculate the distance between x and all data points in the training set (both Manhattan and Euclidean distance) then, will find optimal K nearest neighbor for our training set. In this case, the predicted GDP will be the Distance-Weighted Mean: y = Σᵢ (wᵢ \* yᵢ) for the k-nearest neighbors, where wᵢ is a weight of each neighbor. A common choice is wᵢ = 1 / dᵢ, where dᵢ is the distance from x to the neighbor i.

**Support Vector Regression (SVR):**

Inspired by Shijie Ye, Guangfu Zhu, and Zhi Xiao (Ye, S., Zhu, G., & Xiao, Z. (2012). Long term load forecasting and recommendations for China based on support vector regression), which successfully applied Support Vector Regression (SVR) to establish a nonlinear relationship between load and GDP, I use SVR into my methodology. SVR is particularly useful in capturing complex, non-linear patterns and is thus well-suited to explore the intricate interactions between various economic variables and GDP. I use SVR to model the relationship between the economic factors and GDP, aiming to enhance forecasting accuracy. Unlike some other regression methods, SVR doesn't make strong assumptions about the underlying data distribution. Instead, it relies on finding a hyperplane that best fits the data while allowing for a certain margin of error. The goal of SVR is to find the hyperplane that minimizes the error within this epsilon-tube. SVR is based on a subset of the training data points, known as support vectors. It assumes that only a small number of data points significantly influence the construction of the regression model, which allows SVR to handle outliers robustly. model tuning and cross-validation are necessary to achieve the best results for a specific problem.

SVR tend to minimize ( fit the best line(hyperplane).

Then, find hyperplane that has maximum number of points and unlike regression models that try to minimize error, SVR try to fit the best line within the threshold value , threshold value is a distance between hyperplane and the boundary line. slack variables are the support vectors to determine the structure position of the tube.

C is the cost function, will tune it to avoid over or underfitting (since in linear kernel function, if cost function is high, leads to get low bias and high variance so and will overfit and vice versa)

NON-Linear kernel SVR (support vector regression):

The RBF kernel is commonly used in SVR for its flexibility in modeling non-linear patterns.

f(x) = Σᵢ αᵢ \* K(xᵢ, x) + b

αᵢ: Lagrange multipliers for optimization process.

K: K(xᵢ, x) = exp(-γ \* ||xᵢ - x||²)

γ: controls the width of the kernel and flexibility of the model (higher value of gamma indicate narrow kernel and sensitive to scale variation)

C: Controls the trade-off between maximizing the margin and minimizing the error

**XGBoost:**

Inspired by the study of Mahdi S. Alajmi and Abdullah M. Almeshal (Alajmi, M. S., & Almeshal, A. M. (2020). Predicting the tool wear of a drilling process using novel machine learning XGBoost-SDA), where XGBoost demonstrated high predictive accuracy in predicting tool wear, I use XGBoost for building one of the models. XGBoost is a robust machine learning algorithm known for its effectiveness in complex modeling tasks. In my research, I apply it to forecast GDP values, considering variable importance and regularization. This has ability to capture the relationships among multiple variables and GDP. XGBoost builds an ensemble of decision trees to make predictions. It assumes that the target variable can be explained as the sum of the contributions from individual features and interactions between features. It builds decision trees to capture these additive relationships and interactions. While it can handle correlated features, it's generally better to preprocess the data to remove or reduce multicollinearity. Highly correlated features may lead to unstable feature importance rankings. XGBoost provides feature importance scores, which can help you identify which variables are the most relevant for making GDP predictions. Common hyperparameters to tune include the learning rate, tree depth, and the number of trees in the ensemble is the only challenge which can solved it by cross validation to have better prediction. objective function is minimized MSE during training.

to updates for each tree in the ensemble, I should calculate gradient and Hessian of the loss function ∂L/∂ŷᵢ = -2 \* (yᵢ - ŷᵢ)

prediction for a single tree is a sum of leaf values which added to the previous ensemble's prediction F\_t(x) = Σ w\_i \* I(x ∈ R\_i)

Ω(f) is the regularization term which control overfitting

Ω(f) = γ \* T + ½ \* λ \* Σ w²

γ : the regularization parameter for the number of leaves.

λ : the regularization parameter for the weights.

T : the number of leaves in the tree.

Objective Function:

Obj = Σ L(y, ŷ) + Ω(f)

To minimize the objective function, XGBoost computes updates for the leaf values

w\_i = -G\_i / (H\_i + λ)

w : the updated value.

G : the sum of the gradients for the data points.

H : the sum of the Hessians for the data points.

λ : the regularization parameter for the weights.

XGBoost uses Second-Order Taylor Approximation:

A math equation with black text

Description automatically generated

**Model Evaluation:**

My empirical method concludes with model evaluation. I assess the predictive performance of each modeling technique using standard evaluation metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared values. To determine the model that best captures the relationships between labor market policies, taxation policies, social welfare programs, and GDP, I conducted a comparative analysis of the evaluation metrics for the three techniques.

The technique with the lowest MAE demonstrates the smallest average absolute prediction error and offers the best point prediction performance.

The technique with the lowest RMSE exhibits superior performance in terms of accurately capturing the variability in the target variable.

The technique with the highest R-squared indicates the best overall fit to the data, explaining a greater proportion of variance in the target variable. By comparing these metrics, I select the model that best captures the intricate relationships between labor market policies, taxation policies, social welfare programs, and GDP. After that, I apply the best technique from the train set to the test set data. Once I have predictions for the test data, I evaluate the model's performance using the same evaluation metrics (e.g., Mean Absolute Error, Root Mean Squared Error, R-squared) that I used during model evaluation on your training data. This will show of how well the best model generalizes to unseen data.

Afterwards, I plan to use the model to predict the GDP of different countries. By figuring out which factors are most important, I aim to create a model that can offer useful advice for different nations. Using the model insights, I'll conduct tests to see what might happen to a country's GDP if it decides to change certain policies. This will help in suggesting practical policy changes, especially for developing countries looking for guidance.

A graph of a graph of the growth of the company

Description automatically generated with medium confidenceA graph showing the growth of the gdp per million

Description automatically generated

Two above graphs show GDP and GDP per capita per countries and show importance of population. the above graphs also show OECD total GDP and GDP per capita for total OECD’s countries. As a result, by considering population, if imagine that total OECD’s works as a country, the slope is increasing over time, that’s why this data is suitable to extend the results and policy implication to undeveloped countries.

A graph of a graph showing the amount of tax revenue

Description automatically generated with medium confidenceA graph of a tax revenue

Description automatically generated with medium confidenceUS has the most tax revenue among all other countries which is make sense because it’s a biggest economy, but the revenue as a percentage of GDP, US is bellow most of these countries. Most of the countries tax revenue is smooth over time (except few countries like Iceland that jumps in 2016)

A graph of the us government

Description automatically generated with medium confidenceA graph of a graph showing the average wage

Description automatically generatedsocial spending by governments per capita is not fluctuating over the time for these countries and for total OECD as well. Here, Iceland has the most average wages fluctuation among other countries on average.

A graph of different colored lines

Description automatically generatedA graph showing the growth of the european union

Description automatically generated

Labor force participation rate for 25-64 years old over time for US and total OECD is similar specially after 2014. Inflation rate for all OECDS, US and EU behave similar like each other. This shows how countries economics are dependent on each other. So by detecting what influence countries economics and help other countries to developed, it is helpful for both kinds of countries.

Works cited:

Bird, G. (2004). Growth, Poverty and the IMF. *Journal of International Development*, *16*(4), 621-636.

Kitov, I. O. (2008). GDP growth rate and population. *arXiv preprint arXiv:0811.2125*.

Sunal, O., & Alp, O. S. (2016). Effect of different price indices on linkage between real GDP growth and real minimum wage growth in Turkey. *Journal of Economic & Financial Studies*, *4*(01), 01-10.

BAHLOULA, R. (2022). Estimating Optimal Level of Taxation for Growth Maximization in MOROCCO. *International Journal of Multidisciplinary Studies on Management, Business, and Economy*, *5*(1), 14-31.

Ahmed, M. R., & Shafin, A. A. (2020). Statistical and machine learning analysis of impact of population and gender effect in GDP of Bangladesh: a case study.

MORALES, N. Immigration Enforcement and Labor Supply: Hispanic Youth in Mixed-Status Families.

Paruchuri, H. (2021). Conceptualization of machine learning in economic forecasting. *Asian Business Review*, *11*(1), 51-58.

Ye, S., Zhu, G., & Xiao, Z. (2012). Long term load forecasting and recommendations for China based on support vector regression. *Energy and Power Engineering*, *4*(5), 380-385.

Alajmi, M. S., & Almeshal, A. M. (2020). Predicting the tool wear of a drilling process using novel machine learning XGBoost-SDA. *Materials*, *13*(21), 4952.