Volume 30 (2) October 2023 ISSN 0216-423X (Print)
ISSN 2622-2667 (Online)

JABM
JOURNAL of ACCOUNTING, BUSINESS and MANAGEMENT

Journal of Accounting, Business and Management (JABM) vol. 30 no. 2 (2023) 27-36

Can the WEKA Data Mining Tool be Used in Developing an Economic Growth Model?

Zahid B. Zamir*

Abstract

WEKA is a free and open-source software offering visual outputs and a diverse range of machine-learning algorithms. Its rich collection of algorithms sets it apart from commercial data mining systems. Inflation and economic growth, encompassing entire economies, have far-reaching impacts on individuals, either directly or indirectly. While conventional economic theories often assert a negative relationship between inflation and economic growth, researchers over the last seven decades have encountered various models and datasets both supporting and contradicting these traditional notions. In order to forecast economic growth or inflation, seven key attributes from the World Bank dataset were utilized. These attributes included variables such as war, terms of trade log change, consolidated public sector surplus, and more. The objective was to develop a decision support model using the WEKA data mining tool. Three algorithms within WEKA were employed: linear regression, ZeroR, and RepTree. Results indicate that the linear regression algorithm consistently outperformed the others. It exhibited superior predictive abilities for expected growth, confirmed through both 10-fold cross-validation and a 75% split test. WEKA's versatility, coupled with its machine learning algorithms, especially the linear regression model, provides a potent resource for exploring and predicting the intricate relationship between inflation and economic growth.

Keywords: data modeling, WEKA, growth, inflation, machine learning.

I. INTRODUCTION

Inflation and economic growth are economy-wide phenomena that affect positively or negatively everyone in an economy, either directly or indirectly. Inflation means a general increase in the price level. As inflation rises, every dollar one owns buys a smaller percentage of a good or service. For example, during world war II, one could buy a loaf of bread for $0.15, a new car for less than $1,000, and an average house for around $5,000 in the United States. In the twenty-first century, bread, autos, homes, and everything else cost more. As a result, we have experienced a significant amount of Inflation over the last 60 years. For example, during the time of president Gerald Ford in the 1970s, the united states had a very high inflation rate. As a result, it was dubbed public enemy number 1 by the president of the united states (Mankiw, 2018).

On the other hand, economic growth is generally measured by the amount of production in a country or region over a certain period. Some economists like to put it as the growing ability of the economy to produce goods and services (Krugman & Wells, 2014). In addition to production, measured through the gross domestic product, GDP, local governments and individuals may use different standards to measure economic growth. Most consider economic growth one of the surest signs of a country's overall health. More commerce means more jobs, and more jobs mean more consumption,

* Associate professor (MIS/ERP). Department of Business Administration, College of Business, Delaware State University. 1200 N. DuPont Highway, Dover, DE 19901. 302.857.6937. E-mail: zzamir@desu.edu.

leading to more production. This can be a perfect circle to get into. However, like most things, economic growth tends to occur in cycles.

On the other hand, WEKA is an open-source data mining software package. WEKA, formally called Waikato Environment for Knowledge Analysis, is a computer program developed at the University of Waikato, New Zealand, to identify information from raw data gathered from the agricultural domains. WEKA is a state-of-the-art facility for developing machine learning techniques and their application to real-work data mining problems. WEKA supports standard data mining tasks such as pre-processing, classification, clustering, regression, visualization, and feature selection (Kiranmai & Laxmi, 2018). The basic premise of the application is to utilize a computer application that can be trained to perform machine learning capabilities and derive useful information in the form of trends and patterns (Kiranmai & Laxmi, 2018).

II. LITERATURE REVIEW

In the open source scenario, some recently appeared tools like Mahout and Vowpal Wabbit can perform big data analytics in large computer clusters. On the other hand, desktop-based tools like R, WEKA, and RapidMiner are usually employed in solving smaller yet important problems (Engel et al., 2014). WEKA's rich functionality also allows its use for text and web document pre-processing and mining. In addition, the WEKA system provides a rich set of robust machine learning algorithms for data mining tasks not found in commercial data mining systems. These include basic statistics and visualization tools and tools for pre-processing, classification, and clustering, all available through an easy-to-use graphical user interface (Russell & Markov, 2017).

There have been many studies on the relationship between inflation and economic growth by considering datasets from one country or a group of countries in the past. So the existence and nature of the relationship between inflation and development have had a long history. Although conventional economic theories assert that inflation and economic growth are negatively related, researchers have found different models and datasets in favor and against conventional beliefs throughout the last five decades. Mallick study on "Inflation and growth dynamics: The Indian experience," published in the Journal of Economic Policy Reform (2003 September, vol. 11, no. 3, p.163-172), investigated the impact of inflation on economic growth in India from 1960 to 2005. Hrushkesh's study found significant adverse effects of inflation and the favorable impact of investment on economic growth in India's context. Therefore, according to Hrushkesh, price stability is necessary for India to achieve higher economic growth. Hrushkesh's study refutes the earlier study by Mallik and Chowdhury (2001) on the inflation and economic growth based on four south Asian Countries (India, Pakistan, Bangladesh, and Sri Lanka) that found (based on the empirical evidence obtained from the cointegration and correction models using annual data collected from the IMF) evidence of a long-run positive relationship between GDP growth rate and inflation for all four countries. Mallik and Chowdhury's (2001) study further found that moderate inflation is helpful to growth, but faster economic growth feeds back into inflation. Jin (2009) study "A note on inflation, economic growth, and income inequality" found a similar result that shows an increase in the long-run money growth rate raises inflation and reduces growth. While analyzing the relationship between inflation, economic development, and income inequality, Jin Yi's model also found that income inequality shrinks with inflation.

A large body of economic theory asserts that moderate and stable inflation rates promote development and economic growth. Saaed (2007) conducted research using an annual dataset on real GDP and CPI (CPI or Consumer Price Index used to measure the Inflation rate) from 1985 to 2005 in Kuwait. Saaed's (2007) empirical research also demonstrated a statistically significant long-run negative relationship between inflation and economic growth. Examining the threshold effects in the inflation-growth relationship using the datasets from 1970 to 2003 in Nigeria, Fabayo and Ajilore's (2006) research found an inflation threshold level of 6 percent. Below this level, there exists a significant positive relationship between inflation and economic growth, while above this threshold level, inflation impedes growth performance. Therefore, Fabayo and Ajilore (2006) suggest bringing inflation down to single digits should be Nigeria's macroeconomic management goal.

Pollin and Zhu (2006) conducted a nonlinear cross-country analysis of inflation and economic growth for 80 countries over the 1961 to 2000 period using middle-income and low–income countries. Pollin and Zhu's paper presents new nonlinear regression estimates of the relationship between inflation and economic growth for 80 middle and low-income countries. Their paper finds that higher inflation is associated with moderate gains in gross domestic product (a measure of economic growth) growth up to a roughly 15-18 percent inflation threshold, much higher than

Fabayo and Ajilore's (2006) six percent of inflation threshold. However, Pollin and Zhu's (2006) findings diverge when complete datasets are divided according to income levels. An interesting study by Burdekin et al. (2004) on inflation and growth found that nonlinearities are pretty different for industrial economies than for developing countries. Their study shows that failure to account for nonlinearity biases downward the estimated effects of inflation on growth, mixing industrial and developing economies produces unreliable results. The nonlinear view concerning inflation-growth relationships explains the empirical findings. Fernandez and Carlos's (2003) time series analysis on inflation and economic growth tests the proposition that the growth rate of the economy and the level of inflation are negatively correlated in the long run.

III. RESEARCH METHODOLOGY

3.1. Problem Description and Approach

The primary purpose of this research is to develop a decision support system using real-world datasets. By considering the dataset available in the world bank database concerning inflation and economic growth, this research uses the explorer interface of the WEKA data mining tool and data mining algorithms such as linear regression REPTree and ZeroR rules to help predict the expected growth and inflation. Finally, based on different modeling methods and 10-fold cross-validation, and 75 percentage split test options, a recommendation will be made on the best model for decision support.

3.1.1. 10-Fold cross-validation test option

When 10-fold cross-validation is used, it will split the database into ten different subsets, and then it will use nine subsets to train the data mining algorithm and one subset (the 10th subset) to test it. This process will be repeated by using different subsets for training and testing ten times. Thus, 10- fold cross-validation is the most commonly used process for developing and testing data mining models.

3.1.2. Percentage Split Test Option

The percentage Split test will split the datasets into training and test sets. So if I use a 75 percentage split test, then 75 percentage of the dataset will be used as a training test, and the rest will be used as a test set.

IV. RESULTS AND DISCUSSIONS

4.1. Evaluation

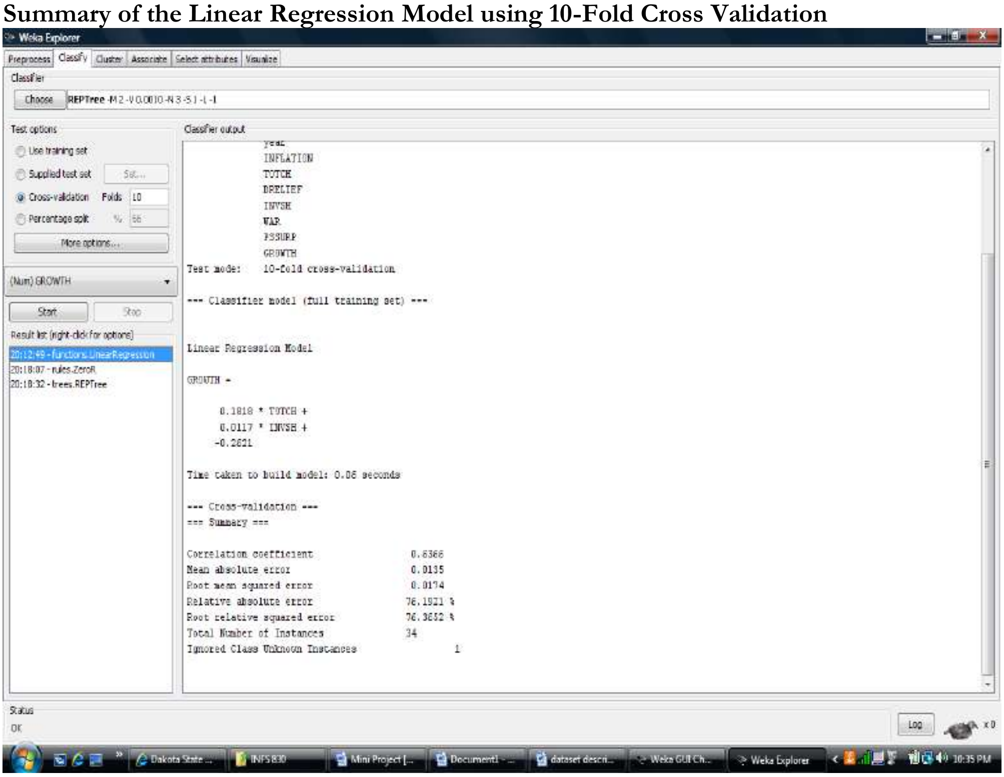
The dataset was taken from Bruno and Easterly (1996) collected data from 136 countries over the 1960-1995 period. The variables that are included in that dataset are as follows:

1) WAR: this variable is used as dummy for the war taking place on national territory.
2) TOTCH: terms of trade log change.
3) PSSURP: consolidated Public sector surplus.
4) DRELIEF: dummy for debt rescheduling (it takes the value of 1 the first year of debt rescheduled and 0 otherwise; only available from 1980 to 2016).
5) INFLATION: december over december; consumer price index-% change.
6) GROWTH: log per capita growth rate.
7) INVSH: investment share in PPP constant prices.

In order to test and develop a decision support model using the WEKA data mining tool to help predict the economic growth with the help of the above seven variables, I have used the linear regression algorithm from the functions classifier, ZeroR algorithm from the rules classifier, RepTree algorithm from the trees classifier. All these different algorithms and models were evaluated using 75% train-test split and 10-fold cross-validation test options.

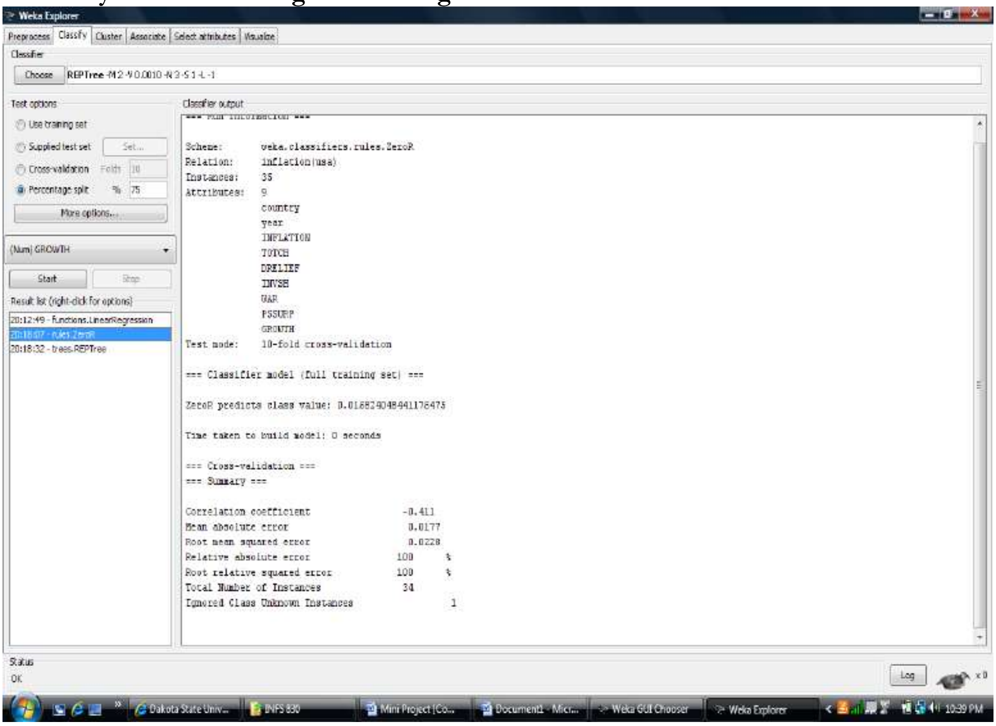
Figure 1

Summary of the Linear Regression Model using 10-Fold Cross Validation



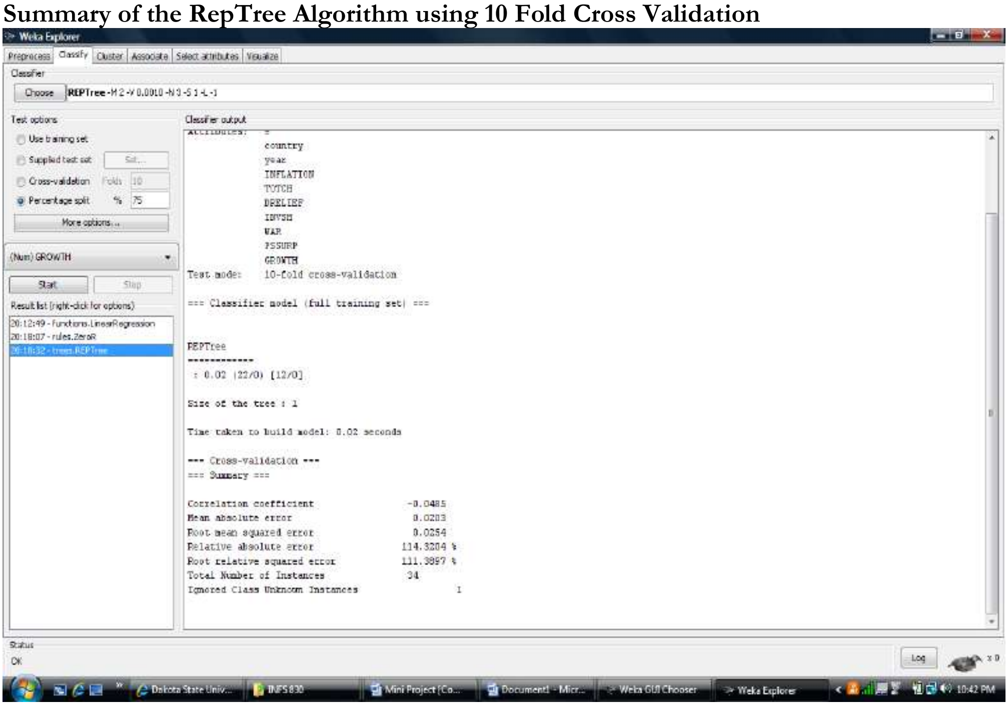
Notes: linear regression= correlation coefficient: 0.6366, mean absolute error (MAE): 0.0135, root MAE: 0.0174, relative absolute error (RAE): 76.1921%.

Figure 2
Summary of the ZeroR Algorithm using 10-Fold Cross Validation



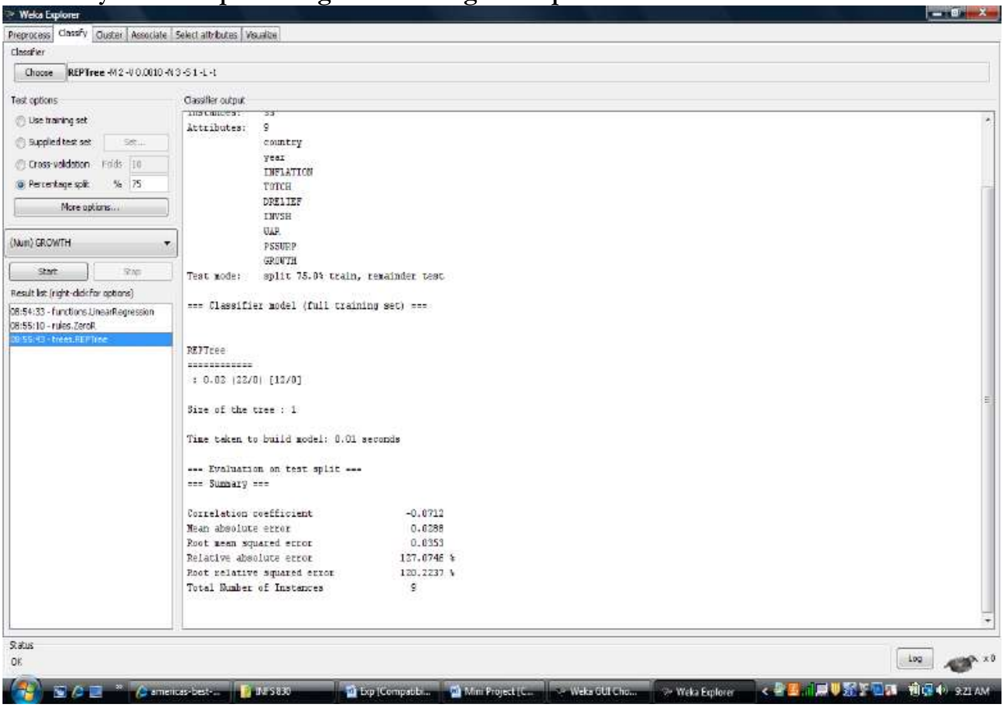
Notes: ZeroR= correlation coefficient: -0.411, mean absolute error (MAE): 0.0177, root MAE: 0.0228, relative absolute error (RAE): 100%.

Figure 3
Summary of the RepTree Algorithm using 10 Fold Cross Validation



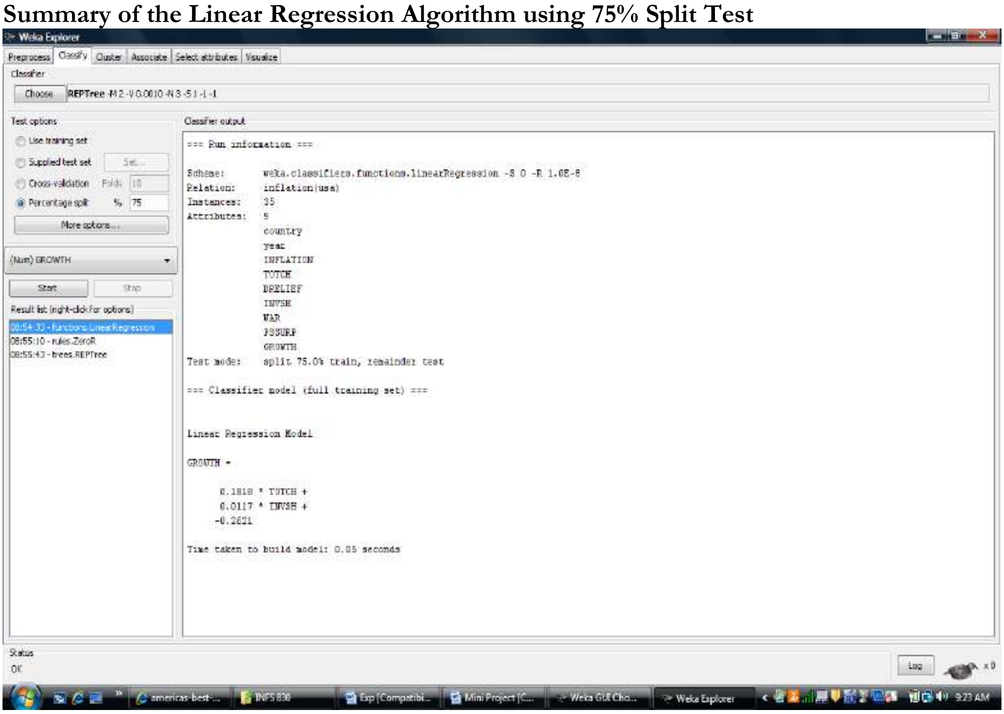
Notes: RepTree= correlation coefficient: -0.0485, mean absolute error (MAE): 0.0203, root MAE: 0.0254, relative absolute error (RAE): 114.3204%.

Figure 4
Summary of the RepTree Algorithm using 75% Split Test



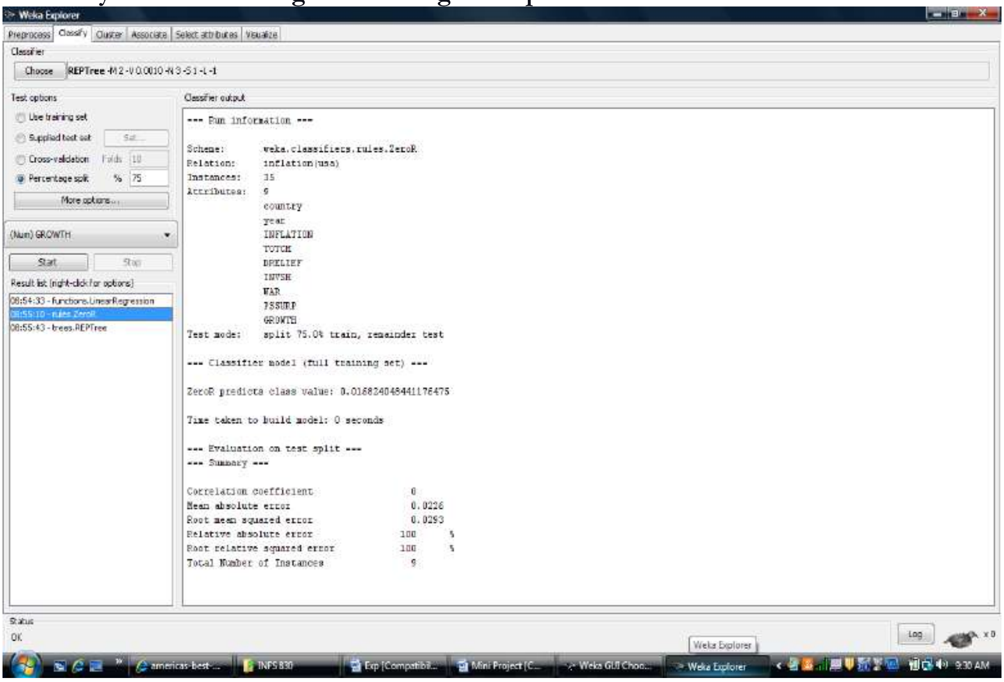
Notes: RepTree= correlation coefficient: -0.0712, mean absolute error (MAE): 0.0288, root MAE: 0.0353, relative absolute error (RAE): 127.2237%.

Figure 5
Summary of the Linear Regression Algorithm using 75% Split Test



Notes: linear regression= correlation coefficient: 0.8523, mean absolute error (MAE): 0.0166, root MAE: 0.0194, relative absolute error (RAE): 73.2584%.

Figure 6
Summary of the ZeroR Algorithm using 75% Split Test



Notes: ZeroR= correlation coefficient: -0, mean absolute error (MAE): 0.0226, root MAE: 0.0293, relative absolute error (RAE): 100%.

V. CONCLUSION

Evaluation criteria are explained below before those results are compared. In addition, the reasons why some essential evaluation criteria are missing have also been explained here.

1) Correlation coefficient: The correlation coefficient, a concept from statistics, measures how well trends in the predicted values follow trends in actual past values. It measures how well the predicted values from a forecast model "fit" with the real-life data. A correlation coefficient is a number between 0 and 1. If there is no relationship between the predicted and actual values, the correlation coefficient is 0 or very low (the predicted values are no better than random numbers). As the strength of the relationship between predicted and actual values increases, so does the correlation coefficient. A perfect fit gives a coefficient of 1.0. Thus the higher the correlation coefficient, the better.
2) Mean absolute error: The MAE measures the average magnitude of the errors in a forecast set without considering their direction. It measures accuracy for continuous variables. Expressed in words, the MAE is the average over the verification sample of the absolute values of the differences between the forecast and the corresponding observation.
3) Root mean squared error (RMSE): The RMSE is a quadratic scoring rule that measures the error's average magnitude. The difference between forecast and corresponding observed values are each squared and then averaged over the sample. Finally, the square root of the average is taken. Since the errors are squared before they are averaged, the RMSE gives a relatively high weight to large errors. This means the RMSE is most useful when large errors are particularly undesirable. The MAE and the

RMSE can be used together to diagnose the variation in the errors in a set of forecasts. The RMSE will always be larger or equal to the MAE; the greater difference between them, the greater the variance in the individual errors in the sample. If the RMSE= MAE, then all the errors are of the same magnitude. Both the MAE and RMSE can range from 0 to ∞. They are negatively-oriented scores: Lower values are better.

4) Mean absolute error: The Mean absolute error is a quantity used to measure how close forecasts or predictions are to the eventual outcomes. The mean absolute error is a standard measure of forecast error in time series analysis.
5) Relative absolute error: The Relative absolute error is very similar to the relative squared error because it is also relative to a simple predictor, which is just the average of the actual values. In this case, though, the error is just the total absolute error instead of the total squared error. Thus, the relative absolute error takes the total absolute error and normalizes it by dividing it by the total absolute error of the simple predictor.

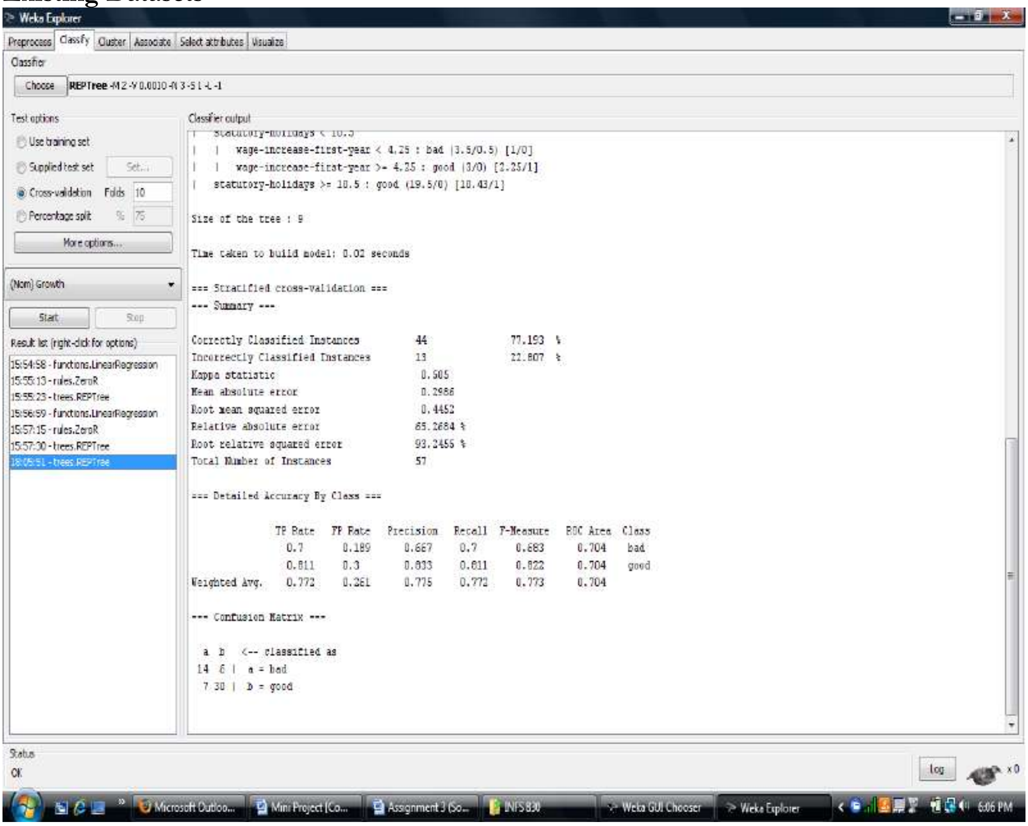
To help predict the expected economic growth or inflation using seven attributes such as war, totch, pssurp, drelief, inflation, invsh, and growth found in the world bank dataset, three models/algorithms and two testing options were used to find out the best model for decision support. As far as the correlation coefficient or measurement of how well the predicted values from a forecast model "fit" with the real-life data is concerned, the linear regression algorithm seems to predict the expected growth much better than the other two algorithms in this study, namely ZeroR and REPTree using both 10 fold cross-validation as well as 75 percentage split test options since coefficient correlation is found to be higher in the linear regression model (0.6366 using 10 fold cross-validation test option and 0.8523 using 75 percentage split test option). One of the most important evaluation criteria to determine the best model is measuring the Mean absolute error. Mean absolute error measures the average magnitude of the errors in a forecast set without considering their direction. The lower the mean absolute error, the better the model. The linear regression algorithm using 10-fold cross-validation gives the lowest mean absolute error, which is only 0.0135, compared to other models and test options used in this study. As far as the other error measurements are concerned, the linear regression model using 10 fold cross-validation test option gives a lower error rate (root mean squared error: 0.0174, relative absolute error: 76.1921%, root relative squared error: 76.3652%) compared to ZeroR and REPTree algorithms.

Therefore, the best data mining model out of the three models used above that best predicts the expected growth is found to be a linear regression algorithm using 10 fold cross-validation test option. The linear regression model works better than other models in this study because all the variables (input and output) in the case of inflation and growth analysis contain continuous data, and the output is a linear combination of input variables. Therefore, a usual linear regression model for this type of data better predicts the outcome.

5.1. Future Research

As mentioned above, all models could not provide some important evaluation criteria using this dataset. However, models predicted the outcome far better when I added other variables like wage increase, cost of living adjustments, average annual working hours, and unemployment benefits with the existing datasets; here is an example of such manipulation.

Figure 7
Manipulation results with Added variables like Wage Increase, Cost of Living Adjustments, Average Annual Working Hours, and Unemployment Benefits with the Existing Datasets



Notes: correctly classified instances: 77.193%, incorrectly classified instances: 22.807%, mean absolute error: 0.2984, relative absolute error: 65.2684%, TP rate: 0.7, FP rate: 0.189, precision: 0.667, recall: 0.7, F-measure: 0.683, and ROC area: 0.704.

Along with all different types of errors, models also give several correctly/incorrectly classified instances along with true positive, FP, precision, F-statistical measure, and ROC. I hope that for future research, one can add those variables to further develop a data mining model.

REFERENCES

Bruno, M., & Easterly, W. R. (1996, May). *Inflation and growth: In search of a stable relationship* (pp. 139-146). Proceedings of the Federal Reserve Bank of St. Louis, vol. 78.

Burdekin, R. C. K., Denzau A. T., Keil, M. W, Sitthiyot T., & Wilett, T. D. (2004, September). When does inflation hurt economic growth? Different nonlinearities for different economies. *Journal of Macroeconomics*, *26*(3), 519-532.

Engel, T. A., Charão, A. S., Kirsch-Pinheiro, M., & Steffenel, L. –A. (2014). Performance improvement of data mining in WEKA through GPU acceletation. *Procedia Computer Science*, *32*, 93-100. Https://doi.org/10.1016/j.procs.2014.05.402.

Fabayo, J. A., & Ajilore, O. T. (2006, July-December). Inflation: How much is too much for economic growth in Nigeria. *Indian Economic Review*, *41*(2), 129-147.

Fernandez, V., & Carlos, G. (2003, August). Inflation and economic growth in the long run. *Economics Letters*, *80*(2), 167-173.

Jin, Y. (2009, February). A note on inflation, economic growth and income inequality. *Macroeconomic Dynamics*, *13*(1), 138-147.

Kiranmai, S. A., & Laxmi, A. J. (2018). Data mining for classification of power quality problems using WEKA and the effect attributes on classification accuracy. *Protection & Control of Modern Power Systems*, *3*(3), 303-314.

Krugman, P., & Wells, R. (2014). *Economics*. New York, NY, USA: Worth Publishers.

Mallick, H. (2008, September). Inflation and growth dynamics: The Indian experience. *Journal of Economic Policy Reform,* *11*(3), 163-172. Https://doi.org/10.1080/17487870802327249.

Mallik, G., & Chowdhury, A. (2001, June). Inflation and economic growth: Evidence from four south Asian countries. *Asia-Pacific Development Journal*, *8*(1), 123-135.

Mankiw, N. G. (2018). *Principles of economics* (4th ed.). Mason, Ohio, USA: Thomson South-Western.

Pollin, R., & Zhu, A. (2006). Inflation and economic growth: A cross-country nonlinear analysis. *Journal of Post Keynesian Economics*, *28*(4), 593-614.

Russell, I., & Markov, Z. (2017, March 8-11). *An introduction to the WEKA data mining system* (pp. 742). Proceedings of the 2017 ACM SIGCSE Technical Symposium on Computer Science Education, Seattle Washington, USA.

Saaed, A. A. J. (2007, January-June). Inflation and economic growth in Kuwait: 1985-2005 evidence from co-integration and error correction model. *Applied Econometrics & International Development*, *7*(1), 143-150.