**Logotipo, nombre de la empresa

Descripción generada automáticamente**

**Predicting job creation in Colombian cities with key economic, social, and demographic information**

**Multivariate and univariate regression approach to job growth prediction**

Supervised by

**Prof. Dr. Simon Munzert**

Professor of Data Science and Public Policy | Director, Data Science Lab

**Alvaro Jose Guijarro May**

Master of Data Science for Public Policy

2024

**Word count:**

**HERTIE SCHOOL**

Berlin, Germany

*If man can predict, almost with certainty, those appearances of which he understands the laws; if, even when the laws are unknown to him, experience of the past enables him to foresee, with considerable probability, future appearances; why should we suppose it a chimerical undertaking to delineate, with some degree of truth, the picture of the future destiny of mankind from the results of its history?*

(De Condorcet, 1795)

*Para Rita, Alvaro, Camilo y Coco. Valió la pena.*

Table of Contents

[1. Introduction 5](#_Toc165200795)

[2. Literature review 6](#_Toc165200796)

[Do local factors have an effect on employment growth? 6](#_Toc165200797)

[Cities development metrics 6](#_Toc165200798)

[Economic growth forecast 6](#_Toc165200799)

[Modelos predictivos de empleo en Colombia 6](#_Toc165200800)

[3. Research Question 6](#_Toc165200801)

[4. Methodology 7](#_Toc165200802)

[Data 7](#_Toc165200803)

[Preprocessing 11](#_Toc165200804)

[Models 12](#_Toc165200805)

[5. Results 13](#_Toc165200806)

[Overall Results 13](#_Toc165200807)

[City Specific Results 14](#_Toc165200808)

[6. Discussion 14](#_Toc165200809)

[Limitations 14](#_Toc165200810)

[Future Work 14](#_Toc165200811)

[7. Conclusion 14](#_Toc165200812)

[8. Bibliography 14](#_Toc165200813)

[9. Statement of Authorship 17](#_Toc165200814)

[10. Annex 17](#_Toc165200815)

List of tables

# Introduction

Public administration and its effects have been evolving in parallel with the societies it has been affecting. At the beginning of the XX century, Max Weber highlighted the necessity of a stable distribution of official labors with the end goal of achieving public objectives in his studies about the rationalization of bureaucracy. He also insisted on the importance of stablishing a system of rules that clearly specifies the authorities in charge of carrying out said duties, its corresponding functions, and the coercive methods available to guarantee effective policy management (Weber, 2024). He laid out the foundations on which the measurement of public policies was built upon.

To connect these historical insights with contemporary governance, it's important to define that Colombia is a social State under the rule of law, structured under the category of a Unitary Republic, whose central authority is the President of the Republic. Its territorial organization has four autonomous territorial divisions: departments, districts, municipalities, and indigenous territories. The municipalities are the primary and essential units of the territorial organization and, together with the departments, enjoy autonomy, based on the decisions that the decentralized regime allows their local authorities to make (Asamblea Nacional Constituyente, 1991). Each municipality must offer the public services established by law, develop the necessary infrastructure for local progress, plan the growth of its territory, foster community participation, and improve the social and cultural welfare of its residents.

Mayors are conceived in Colombia as heads of local administration, legal representatives, and first political authority (Congreso de la República de Colombia, 1994) as well as the police authority of the municipality. They are democratically elected for four-year terms, without the possibility of reelection (Asamblea Nacional Constituyente, 1991). By virtue of their role as authorities with constitutional functions, mayors act as economic agents, making decisions on the demand and supply of goods and services within their territories and as joint representatives of the authorities of the governors and presidents, who in turn define the policies for the regulation and promotion of labor in Colombia (Dorado, 2021). They must direct the administrative action of their municipalities, ensuring the provision of public services and the proper functioning of local industrial or commercial enterprises. In order to determine their actions and decisions, they must present plans and programs for economic and social development, consistent with municipal expenditure and investment, collection, and budget plans (Asamblea Nacional Constituyente, 1991).

Government officials have the opportunity to guide their communities positively, or negatively, and affect the lives of their citizens with their public policy priorities, implementations, and executions. Leaders specially have the capacity of affecting economic growth in their countries (Jones & Olken, 2005). The actions taken during their administrative periods have effects on current and future economic, social, and demographic metrics, and this relationship is not exclusive for one community; any society that has this type of government structure has and will experience it (Jones & Olken, 2005). A way of determining how positive or negative the performance of public officials has been on their societies is to evaluate key metrics that track important development factors through time. Being able to measure their overall performance can help public official to “evaluate, control, budget, motivate, promote, celebrate, learn, and improve”(Behn, 2003).

# Literature review

Do local factors have an effect on employment growth?

There have been previous attempts to try to quantify the effects social, demographic, economic, or geospatial factors have in local employment growth, like the research conducted by Richard Shearmur and Mario Polèse, in which they analyzed the impact of local and structural factor on employment growth in Canada (Shearmur & Polèse, 2007). By analyzing why employment growth occurs on some regions of the country and not in others, they were able to determine that “local (endogenous) and structural (exogenous) factors retain significant explanatory power”(Shearmur & Polèse, 2007) regarding employment growth. Some of the factors used in this study were education levels, population growth, workers’ wages, and geographic locations of Canadian regions.

The International Monetary Fund has pointed out that job growth and creation are within countries and cities top priorities, but the outlook for growth and creation remain as an important concern (International Monetary Fund, 2013). The set of forces that influence growth and job creation in developing countries in recent decades are technological change, demographic changes, poverty rates, GDP, income inequality, and fiscal maneuverability (International Monetary Fund, 2013).

In the Colombian context, researchers have looked into the effect in employment rates and the economy from fiscal policies (Gerardo et al., 2014), educational and health reforms (Martínez-Álvarez, 2015),

Falta lo que dijo Julian

Cities development metrics

Economic growth forecast

Modelos predictivos de empleo en Colombia

# Research Question

This paper will look to evaluate if the evolution of key economic, social, or demographic metrics in Colombian cities can be used to predict economic wellbeing in said locations, reflected in citizen’s employment levels in the formal economy.

# Methodology

Data

As of 2019, Colombia had a population of around 49 million people, distributed in 32 states and 1103 municipalities (DANE, 2019). 21 million people, or around 45% of the total population of the country is distributed along its 13 biggest cities and their metropolitan areas (referred to as “A.M” for their Spanish definition of “Area Metropolitana”). These are: Bogota A.M., Medellín A.M., Cali A.M., Barranquilla A.M., Cartagena, Cúcuta A.M., Bucaramanga A.M. Villavicencio, Ibagué, Monteria, Pereira A.M., Manizales A.M, and Pasto. The following map shows the distribution of these cities on the Colombian territory with their official names:

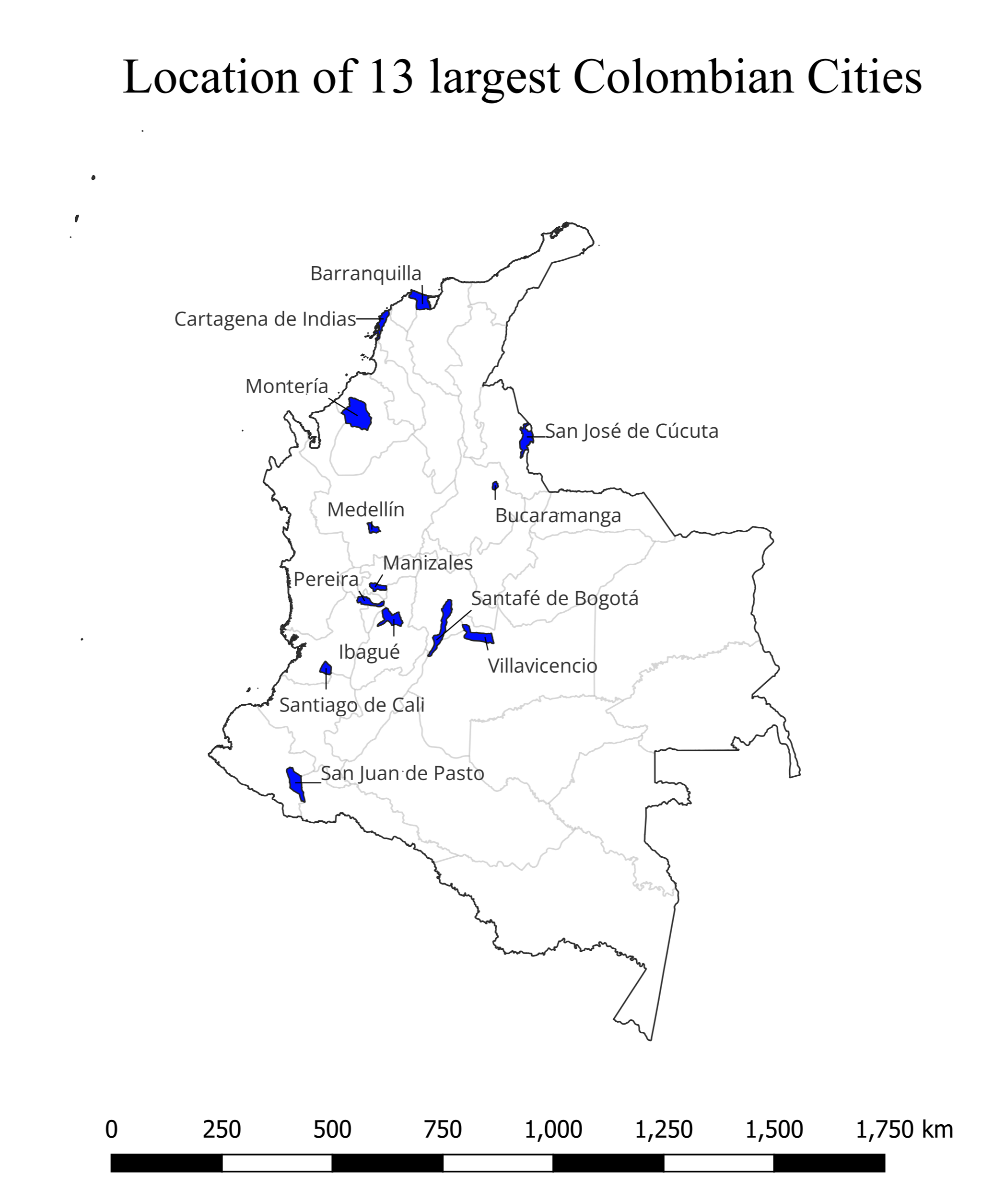


Illustration 1Map of selected Colombian cities

The country’s unique socio-economic landscape has changed through the years by various demographic, political, economic, geospatial, and social factors. Most of the population has historically been based in the center of the country, some 2,625m above sea level in the capital city of Bogotá A.M. in the Andes mountains, with other population centers scattered throughout the country’s diverse landscape. Here, each city has found their own cultural, economic, and industrial identity, and has managed to provide stability and community for their local populations. Bogotá A.M. and Medellín A.M. are considered the commercial and administrative centers of the country, with Barranquilla A.M. and Cartagena being major port cities filled with trade, industry and tourism, Pereira A.M., Manizales A.M., and Ibagué focusing on coffee and rich agricultural activities, in addition to Cali A.M. and Bucaramanga concentrating on manufacturing and industrial activities. Each city has had its own population evolution during the last decades, and this has also affected their job markets. In the following plot we can observe the evolution of employed citizens vs population growth of each city.

A graph of a person with red and blue lines

Description automatically generated with medium confidence

Illustration 2. Distribution of population vs employed citizens by city, 2015-2024

All of the cities analyzed in this study suffered from a drop in employed workers after March 2020, which coincides with the beginning of the COVID-19 pandemic in Colombia (Ministerio de Salud y Protección Social, 2020). Most of the cities suffered from a decrease in their population growth, and since we will be looking into fitting regression models for prediction, the time limit chosen for this study is the end of 2019.

Understanding the diverse and unique context of each city is fundamental for attempting to predict their job growth in previous years. Taking inspiration from the literature review, a search for possible datasets was conducted in order to identify the different economic, social, and demographic variables that could be used for predictions. The details of these can be found in Annex 1. In summary, the different dataset gathered for this analysis were:

|  |  |  |  |
| --- | --- | --- | --- |
| **Name of Dataset** | **# Of Variables** | **Source** | **Description** |
| “*Gran Encuesta Integrada de Hogares*” | 4 | (DANE, 2023a) | Survey that contains the information of Colombian’s employment conditions, in addition to general characteristics of the population such as sex, age, marital status and educational level, and asks about their sources of income. The GEIH provides the country with information at the national level, head, regional, departmental, and for each of the departmental capitals. |
| Population | 4 | (DANE, 2019) | Population projections taking as base the 2018 Census methodology. |
| Consumer Price Index | 6 | (Banco de la República, 2024) | The consumer price index (CPI) measures the evolution of the average cost of a basket of goods and services representative of households’ final consumption, expressed in relation to a base period. Calculated with data from DANE. |
| Education | 24 | (Ministerio de Educación Nacional, 2024) | Contains statistical information on preschool, primary, secondary, and high school levels related to sector indicators by municipality without outliers, from 2011 to 2022. |
| Monetary Poverty | 8 | (DANE, 2023b) | Contains statistical information on preschool, primary, secondary, and high school levels related to sector indicators by municipality without outliers, from 2011 to 2022. |
| MDM Cities Indicators | 24 | (Departamento Nacional de Planeación, 2021) | Municipal Performance Measurement (MDM) aims to measure, compare, and rank municipalities according to their municipal performance, understood as management capacity and development results, taking into account their initial states. |
| Fiscal Performance Amounts | 35 | (Departamento Nacional de Planeación, 2022a) | Municipal and departmental budget execution information aggregated in the Cash Flow Statement. |
| Fiscal Performance Scores | 10 | (Departamento Nacional de Planeación, 2022b) | Fiscal performance Score of the territorial entities for different fiscal years |

Table 1 Description of datasets gathered

The information contained in each of these datasets could be affected by the public policy decisions of government officials, and following the literature review, have highlighted some influence in job growth and generation indicators (agregar fuentes). From this initial assessment, a group of 34 variables was selected to create a more manageable dataset for each city. These were:

|  |  |  |
| --- | --- | --- |
| **#** | **Variables** | **Description** |
| 1 | workers.geih | Employed population |
| 2 | date | Year and Month |
| 3 | year | Year |
| 4 | month | Month |
| 5 | population\_month.pop | Total Population in monthly frecuency (interpolated) |
| 6 | population\_year.pop | Total Population in yearly frecuency |
| 7 | CPI.cpi | Consumer Price Index, The Consumer Price Index (CPI) is a measure that examines the weighted average of prices of a basket of consumer goods and services, such as transportation, food, and medical care. The CPI is calculated by taking price changes for each item in the predetermined basket of goods and averaging them. |
| 8 | CPI\_month\_var.cpi | The CPI (Consumer Price Index) monthly variance refers to the change in the CPI from one month to the next, expressed as a percentage. This measure provides an indication of how consumer prices have moved within a month, reflecting short-term inflation or deflation trends. |
| 9 | Enrollment\_Rate\_5\_16.edu\* | A proportion of the population between 5 and 16 years old are attending the educational system. When DANE's population projections do not adequately capture internal migratory flows, it can reach values greater than 100%." |
| 10 | Net\_Coverage.edu\* | The ratio between the number of students enrolled in transition, primary, secondary, and high school who have the theoretical age (5 to 16 years) and the total population of that same age. |
| 11 | Pass\_Rate.edu\* | Pass rate of students in the official sector. Identifies the percentage of students in preschool, basic, and high school education who pass according to current educational plans and programs. |
| 12 | I\_PM.mp\* | % of population - Monetary Poverty Rate |
| 13 | I\_PME.mp\* | % of population - Extreme Monetary Poverty Rate |
| 14 | Gini.mp\* | Gini Coefficient (values between 0-1) |
| 15 | IPUG.mp\* | Average Per Capita Income of the Household Spending Unit in Colombian Pesos |
| 16 | LP.mp\* | Monetary Poverty Lines (monthly values per person) |
| 17 | LPE.mp\* | Extreme Monetary Poverty Lines (monthly values per person) |
| 18 | MDM\_Resource\_Mobilization.ci\* | Score between 1-100 - Measures mobilization of financial resources |
| 19 | MDM\_Execution\_Of\_Resources.ci\* | Score between 1-100 - Execution of financial resources |
| 20 | MDM\_Open\_Government\_And\_Transparency.ci\* | Score between 1-100 - Measures of open government and transparency practices |
| 21 | MDM\_Territorial\_Ordering.ci\* | Score between 1-100 - Territorial ordering and planning measures |
| 22 | MDM\_Education.ci\* | Score between 1-100 - Educational coverage and quality in middle education |
| 23 | MDM\_Health\_Coverage.ci\* | Score between 1-100 - Health coverage and services |
| 24 | MDM\_Services.ci\* | Score 1-100 - Coverage and quality of public services |
| 25 | MDM\_Security\_And\_Coexistence.ci\* | Score 1-100 - Security and social coexistence indicators |
| 26 | TotalIncome.fp\* | $ Millions of Pesos - Total income received in the municipality |
| 27 | TotalExpenses.fp\* | $ Millions of Pesos - Total expenses of the municipality\* |
| 28 | Self\_financing\_of\_operating\_expenses.sfp\* | Score 1-100 - Self-financing of operating expenses: the ability to cover the operating expenses of the central administration with unrestricted income (Law 617 of 2000) |
| 29 | Debt\_service\_support.sfp\* | Score 1-100 - Debt service support: the ability to support debt service with perceived revenues. |
| 30 | Dependence\_on\_transfers\_from\_the\_Nation\_and\_Royalties.sfp\* | Score 1-100 - Dependence on transfers from the Nation and Royalties: measures the importance of national transfers and royalties (SGR) in total revenues. |
| 31 | Generation\_of\_Own\_Resources.sfp\* | Score 1-100 - Generation of Own Resources: the ability to generate resources complementary to the transfers. |
| 32 | Magnitude\_of\_Investment.sfp\* | Score 1-100 - Magnitude of Investment: quantifies the magnitude of the investment executed by the territorial entity. |
| 33 | Saving\_Capacity.sfp\* | Score 1-100 - Saving Capacity: determines the degree to which surpluses are freed up to finance investment. |
| 34 | Fiscal\_Performance\_Indicator.sfp\* | Score 1-100 - Fiscal Performance Indicator |

**\*(interpolated to match GEIH employed workers monthly frequency)**

Table 2, Selected variables for model fitting and analysis

Preprocessing

The first decision was to define the scope of the research and the timeframe of interest for the evaluation. In Colombia, administrative periods for Mayors and Governors are for 4 years since 2008 (Asamblea Nacional Constituyente, 1991, art 315). Our datasets span between several administrative time periods (2008-2011 / 2012-2015 / 2016-2019 / 2020-2023 / 2024-2027). Taking into consideration the start of the 2020 Covid-19 pandemic and the disruption it brought to global health, economic, logistical, and productive systems, as well as the introduction of the MDM statistic in 2016 ("*Medición de Desempeño Municipal*" = Municipal Performance Measurement, which ranks Colombian cities by their performance on various key economic, health, safety, and demographic indicators) (Departamento Nacional de Planeación, 2021), the timeframe selected for our predictions analysis was the 2016-2019 administrative time period.

Our variable of interest is the total number of employed civilians for each city, stored in the *Gran Encuesta Integrada de Hogares* (GEIH) dataset. This information is stored in a monthly basis, so in order to be able to fit and define our prediction models, our dependent variables have to match that timescale (Brockwell & Davis, 2010). In order to do so, for all of the variables that were originally stored in a yearly frequency, interpolation was applied. Here, the unknown monthly values were estimated using spline interpolation. In this method of interpolation the interpolant is a piecewise polynomial known as a spline, that fits low-degree polynomials to segments of data points to enhance accuracy and prevent oscillatory errors typical of high-degree polynomial interpolation (Wolber & Alfy, 1999). In order to avoid overestimations in the dependent variables, especially those related to yearly scores, the interpolation was set within unique minimums and maximums for each, so it followed mostly the trends of the historical information.

Models

In the realm of statistical models, diverse methodologies cater to specific data characteristics and analytical needs. Linear regression models are foundational in statistical analysis and are used primarily to predict a dependent variable based on linear relationships with one or more independent variables. For this type of model, it is assumed that the relationship between the dependent and independent variables is linear and that the residuals are normally distributed and homoscedastic (Yan & Su, 2009). Time series models are specialized for analyzing data structured in time order, they are crucial for forecasting where data show seasonality, trends, and autocorrelation, and they focus on dependencies within the time series itself rather than external variables (Brockwell & Davis, 2010). Machine learning models provide more robust predictions by learning complex patterns from large datasets without explicitly predefined equations, they can model nonlinear interactions and are particularly useful in scenarios where relationships between variables are highly complex or are not well understood (Hurwitz, 2018). Evaluating these different types of models will provide insights into the feasibility of accurately predicting our variable of interest.

**OLS Model**

The Ordinary Least Squared (OLS) model is used to estimate the relationship between a dependent variable and one or more independent variables, looking to minimize the sum of the squares of the differences between observed and predicted values. It ensures the best fit line through data points, aiming to yield unbiased and efficient estimates as well as to reduce prediction error (Yan & Su, 2009).

= Dependent Variable  
 = Intercept of model  
= Coefficients of independent variables  
 = Independent variables  
 = Residuals.

Linear regression makes several key assumptions:

1. There must be a linear relationship between the independent and dependent variables
2. All variables need to be multivariate normal
3. There must be little or no multicollinearity in the data
4. No autocorrelation
5. Homoscedasticity

**ARIMA & SARIMA Model**

The Autoregressive Integrated Moving Average (ARIMA(*p,d,q*)), model is designed to forecast data based on its own past values and it is a cornerstone of univariate time series analysis (Box & Jenkins, 1976). It encapsulates three key components:

1. Autoregression (AR): Relationship between an observation and a number of lagged observations (*p*)
2. Integration (I): representing the differencing steps to make the series stationary (*d)*
3. Moving Average (MA): models the error term as a combination of previous errors (*q*)

.

asda

asda

asda

AR

MA

I

= Time series data at time *t*, dependent variable we are trying to forecast  
*p* = # of lag observations included in the model  
 = Coefficients of the autoregressive terms  
*L =* Lag operator  
*d =* Degree of differencing  
*q =* Order of the moving average  
= Coefficients of the moving average terms  
= Error terms

Seasonal Autoregressive Integrated Moving Average (SARIMA) extends the ARIMA model by specifically addressing and modeling seasonal variations in data. It is particularly useful for modeling data with strong seasonal effects. The model is expressed as SARIMA(*p,d,q*)(*P,D,Q*)[*S*], where (*p,d,q*) have the same definitions as the ARIMA model, *P* represent the seasonal autoregressive, *D* the seasonal differencing, *Q* the moving average terms, and finally *S* indicates the length of the seasonal cycle (Brockwell & Davis, 2010).

asda

asda

asda

asda

asda

asda

Seasonal MA

Non-seasonal MA

Non-seasonal AR

Seasonal AR

Non-seasonal  
diferencing

Non-seasonal  
diferencing

= Time series data at time *t*, dependent variable we are trying to forecast  
*p* = # of lag observations included in the model for non-seasonal autoregressive part  
*P* = # of seasonal autoregressive terms  
 = Coefficients of the non-seasonal autoregressive terms  
 = Coefficients of the seasonal autoregressive terms  
*L =* Lag operator  
*d =* Degree of differencing on a non-seasonal level  
*D =* Number of seasonal differencing  
*q =* Order of the non-seasonal moving average  
*Q =* Order of seasonal moving average  
= Coefficients of the non-seasonal moving average terms  
 = Coefficients of seasonal moving average terms  
*s =* length of seasonal cycle in the data  
= Error terms

**Random Forest**

Random forest is a machine learning algorithm utilized for predictive modeling, which enhances decision tree methods by generating a “forest” of trees and aggregating their predictions. This technique constructs numerous decision trees during training and determines the outcome by outputting the mean prediction of these trees. Randomness is introduced by using different subsets of the data to build each tree (bootstrap aggregating) and by selecting a random subset of features for each split. It effectively handles high dimensionality and maintains accuracy even with missing values, also known for its robustness against overfitting (Breiman, 2001). It was also selected in this study for its superior capability to model complex interactions and process large datasets efficiently.

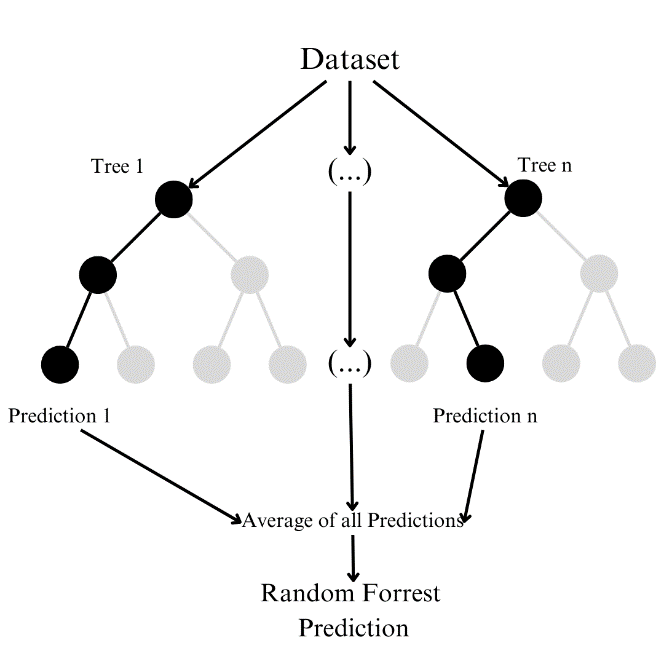


Figure 2, Random Forrest Prediction Algorithm

In order to be able to assess the capability of our selected variables in effectively predicting our employed workers in the different Colombian cities, 8 different models were fitted for each city. As a base, an OLS model with all of our selected variables was fitted first, followed by a more concise and robust OLS model only with the most significant variables selected. From here, two ARIMA and one SARIMA model were fitted manually and also by using the auto.arima function in R, which determines the most optimal (p,d,q) elements (Hyndman & Khandakar, 2008). Finally, 3 random forest models were also fitted for our benchmark, one using all of our selected variables, one focusing on obtaining the best RMSE, and one focusing on obtaining the best R2.

For evaluating our models, the following performance metrics were selected for our benchmark:

* MAE: Mean Absolute Error,
* RMSE:
* R Squared:
* Residual SE:

Fitting the models and importance of comparing

# Results

* Explanation of Result Metrics

Overall Results

* Table with results for 13 cities, OLS, ARIMA, SARIMA Results, Random Forrest results

City Specific Results

* Examples of prediction for Bogota, Barranquilla, and Medellín.

# Discussion

Limitations

* Complexity of variable y prediction
* Limitations in current available data, “Higher education data, Monthly frequency data, etc etc)
* Interpolation and transformation
* Descriptive analysis only, causation component needs to be explored

Future Work

* Identifying more suitable models
* Scaling and transforming variables.
* More precise and trustworthy Data gathering
* City specific, Economic Sector Specific

# Conclusion

* Prediction is possible but complex.
* Correlation does not mean causality
* Public policy must be based on empirical data, and decisions have to be made according to facts but guided by continuous improvement.
* Public officials have the responsibility and duty of fostering an air of improvement, possibilities, and wellbeing in their jurisdiction, and must have plans in place that will lead to this.
* Fortalecer el Sistema de accesos de datos para poder utilizar las ventajas de las nuevas tecnologías.

# Bibliography

Asamblea Nacional Constituyente. (1991). *Constitución Política de Colombia de 1991.* Imprenta Nacional.

Banco de la República. (2024). *Indice de Precio al Consumidor IPC Por Ciudades* [dataset]. https://totoro.banrep.gov.co/analytics/saw.dll?Portal

Behn, R. D. (2003). Why Measure Performance? Different Purposes Require Different Measures. *Public Administration Review*, *63*(5), 586–606. https://doi.org/10.1111/1540-6210.00322

Box, G. E. P., & Jenkins, G. M. (1976). *Time Series Analysis: Forecasting and Control*. Holden-Day. https://books.google.de/books?id=1WVHAAAAMAAJ

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Brockwell, P. J., & Davis, R. A. (2010). *Introduction to time series and forecasting* (Second edition, (corrected at 8th. printing 2010)). Springer.

Congreso de la República de Colombia. (1994). *Ley 136 de 1994*.

DANE. (2019). *PROYECCIONES DE POBLACIÓN A NIVEL MUNICIPAL. PERIODO 2005—2019.* [dataset]. https://www.dane.gov.co/index.php/estadisticas-por-tema/demografia-y-poblacion/proyecciones-de-poblacion

DANE. (2023a). *Gran Encuesta Integrada de Hogares* [dataset]. https://www.dane.gov.co/index.php/estadisticas-por-tema/mercado-laboral/empleo-y-desempleo/mercado-laboral-historicos

DANE. (2023b). *Pobreza Monetaria* [dataset]. https://www.dane.gov.co/index.php/estadisticas-por-tema/pobreza-y-condiciones-de-vida/pobreza-monetaria

De Condorcet, M. (1795). Future progress of mankind. In *Outlines of an historical view of the progress of the human mind: Being a posthumous work of the late M. De Condorcet.* (pp. 316–372). J Johnson. https://doi.org/10.1037/11670-010

Departamento Nacional de Planeación. (2021). *Medición de Desempeño Municipal* [dataset]. https://portalterritorial.dnp.gov.co/AdmInfoTerritorial/MenuInfoTerrEstMDM

Departamento Nacional de Planeación. (2022a). *Historico Operaciones Efectivas de Caja* [dataset]. https://2022.dnp.gov.co/programas/desarrollo-territorial/Estudios-Territoriales/Informacion-Presupuestal/Ejecuciones-Presupuestales/Paginas/Operaciones%20Efectivas%20de%20Caja.aspx

Departamento Nacional de Planeación. (2022b). *Información fiscal y financiera* [dataset]. https://www.dnp.gov.co/LaEntidad\_/subdireccion-general-descentralizacion-desarrollo-territorial/direccion-descentralizacion-fortalecimiento-fiscal/Paginas/informacion-fiscal-y-financiera.aspx

Dorado, D. (2021). *El Rol del Alcalde en el Desarrollo de los Mercados*.

Gerardo, C.-T., Robinson, C.-Á., & Asdrúbal, R.-M. (2014). *Impact of fiscal policy on savings and employment in colombia (1970-2011)*. *21*.

Hurwitz, J. (2018). *Machine Learning For Dummies®, IBM Limited Edition*.

Hyndman, R. J., & Khandakar, Y. (2008). Automatic Time Series Forecasting: The forecast Package for R. *Journal of Statistical Software*, *27*(3), 1–22. https://doi.org/10.18637/jss.v027.i03

International Monetary Fund. (2013). Jobs and Growth—Analytical and Operational Considerations for the Fund. *Policy Papers*, *2013*(18). https://doi.org/10.5089/9781498342148.007

Jones, B. F., & Olken, B. A. (2005). *DO LEADERS MATTER? NATIONAL LEADERSHIP AND GROWTH SINCE WORLD WAR II*.

Martínez-Álvarez, J. J. (2015). *IMPACTO DE LAS REFORMAS ECONÓMICAS NEOLIBERALES EN COLOMBIA DESDE 1990.* *8*(1).

Ministerio de Educación Nacional. (2024). *MEN ESTADISTICAS EN EDUCACION EN PREESCOLAR, BÁSICA Y MEDIA POR MUNICIPIO* [dataset]. https://www.datos.gov.co/Educaci-n/MEN\_ESTADISTICAS\_EN\_EDUCACION\_EN\_PREESCOLAR-B-SICA/nudc-7mev/about\_data

Ministerio de Salud y Protección Social. (2020, June 3). Colombia confirma su primer caso de COVID-19. *Colombia confirma su primer caso de COVID-19*. https://www.minsalud.gov.co/Paginas/Colombia-confirma-su-primer-caso-de-COVID-19.aspx#:~:text=Bogot%C3%A1%2C%206%20de%20marzo%20de,una%20paciente%20de%2019%20a%C3%B1os.

Shearmur, R., & Polèse, M. (2007). Do Local Factors Explain Local Employment Growth? Evidence from Canada, 1971–2001. *Regional Studies*, *41*(4), 453–471. https://doi.org/10.1080/00343400600928269

Wolber & Alfy. (1999). Monotonic cubic spline interpolation. *Proceedings Computer Graphics International CGI-99*, 188–195. https://doi.org/10.1109/CGI.1999.777953

Yan, X., & Su, X. (2009). *Linear regression analysis: Theory and computing*. World Scientific.

# Statement of Authorship

# Annex

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Name of Variable | Source | Variables of Interest | Frequency & Time Frame | Description |
| Gran Encuesta Integrada de Hogares | (DANE, 2023a) | * **Workers** * Economic sector * City * Date | Monthly 2015-03/2023-12 | Survey that contains the information of Colombian’s employment conditions, in addition to general characteristics of the population such as sex, age, marital status and educational level, and asks about their sources of income. The GEIH provides the country with information at the national level, head, regional, departmental, and for each of the departmental capitals. |
| Population | (DANE, 2019) | * Population monthly * Population Yearly * City * Month | Yearly 2010 - 2023 | Population projections taking as base the 2018 Census methodology. |
| Consumer Price Index | (Banco de la República, 2024) | * Consumer Price Index (CPI) * CPI year to date variation * CPI yearly variation * CPI monthly variation * City * Date | Monthly 1973-01/2024-03 | The consumer price index (CPI) measures the evolution of the average cost of a basket of goods and services representative of households’ final consumption, expressed in relation to a base period. Calculated with data from DANE. |
| Education | (Ministerio de Educación Nacional, 2024) | * City * Date * Year * Enrollment Rate 5-16 y.o * Net Coverage * Net Coverage Transition * Net Coverage Primary * Net Coverage Secondary * Net Coverage High School * Dropout Rate * Dropout Rate Transition * Dropout Rate Primary * Dropout Rate Secondary * Dropout Rate High School * Pass Rate * Pass Rate Transition * Pass Rate Primary * Pass Rate Secondary * Pass Rate Highschool * Fail Rate * Fail Rate Transition * Fail Rate Primary * Fail Rate Secondary * Fail Rate High School | Yearly  2011 - 2022 | Contains statistical information on preschool, primary, secondary, and high school levels related to sector indicators by municipality without outliers, from 2011 to 2022. |
| Monetary Poverty | (DANE, 2023b) | * I\_PM | Monetary Poverty Rate * I\_PME | Extreme Monetary Poverty Rate * Gini | Gini Coefficient * IPUG | Average Per Capita Income of the Household Spending Unit * LP | Monetary Poverty Lines (monthly values per person) * LPE | Extreme Monetary Poverty Lines (monthly values per person) * City * Date | Yearly  2012-2022 | Contains official monetary poverty figures of the Colombian population, corresponding to the methodological update based on information from the GEIH. |
| MDM Cities Indicators | (Departamento Nacional de Planeación, 2021) |  | Yearly 2015-2022 | Municipal Performance Measurement (“Medición de Desempeño Municipal” MDM) aims to measure, compare, and rank municipalities according to their municipal performance, understood as management capacity and development results, taking into account their initial states. |
| Fiscal Performance Amounts |  |  |  |  |
| Fiscal Performance Scores |  |  |  |  |

Annex 2 Description of variables gathered