correlation one

Behavior of the operating expenses of the territorial entities and capital cities of Colombia

David Pardo, Edward Sandoval, Isabela Mercado, Jairo Urrego,
Julián Orrego, Sebastian Gallon, Yosef Guevara

Team - 194

Asocapitales - DS4A by Correlation One
July 7, 2022

Contents

1 Abstract

Entities are required to submit budget information in accordance with the provisions of Title II and VI of Resolution No. 0007 of 2016 issued by the Comptroller General of the Republic. With this information, entities like the National Planning Department make the fiscal performance index, a measurement of the performance of the financial management of the territorial entities that account for their financial sustainability. For the calculation of the said indicator, information such as the level of income, operating expenses, and the category to which each municipality belongs, according to its annual income and population, are considered.

Keywords - Operational Expenses, incomnes, budget, Fiscal performance index, municipalities

2 Introduction

The Colombian Association of Capital Cities - Asocapitales is a non-profit organization that aims to work in the preparation, consolidation and management of a common agenda built from the territories and consisting of issues of national, regional and local scope and interest. To achieve this purpose, the Association works mainly in the generation of spaces for dialogue, integration, coordination and collaboration between municipal and national authorities.

Law 617 of 2000 defined a limit to the operating expenses of territorial entities in order to generate a framework of fiscal responsibility. On the other hand, Law 2082 of 2021 defined a framework for the allocation of competencies, where the Fiscal Performance Index, which incorporates the results of the operating expenditure indicator defined by Law 617 of 2000, is taken as a starting point. In this sense, as of 2019 there were some territorial entities that did not comply with the operating expenditure limit. Additionally, the Fiscal Performance Index is published 5 months after the end of the year, so the information is not available quickly for decision-making in terms of expenditure execution of territorial entities. In this sense, it is necessary to create a set of metrics to determine whether territorial entities are efficient in the execution of operating expenses.

Achieving the construction of these metrics is important because it allows to determine the efficiency in the execution of operating expenses for territorial entities and to have timely information for decision-making by the authorities. Greater efficiency allows more resources to be available for investment in the cities and to comply with the indicators established by law.

3 Exploratory analysis

3.1 Historical CGR Category

From the first 5 row from the Historical CRG category, it is possible to confirm that deals with structured data composed have only by categorical ordinal data, from the year 2012 to 2022, where the cat columns represent the category of the municipality been 1 the highest category and 6 the lowest, the one with the largest and the one with the smallest budget, respectively.

	codigo_cgr	divipola	departamento	MUNICIPIO	cat_2012	cat_2013	cat_2014	cat_2015	cat_2016	cat_2017	cat_2018	cat_2019	cat_2020	cat_2021	cat_2022
0	210105001.0	5001.0	ANTIOQUIA	MEDELLIN	ESP	ESP	ESP	ESP	ESP	ESP	ESP	ESP	ESP	ESP	ESP
1	210205002.0	5002.0	ANTIOQUIA	ABEJORRAL	6	6	6	6	6	6	6	6	6	6	6
2	210405004.0	5004.0	ANTIOQUIA	ABRIAQUI	6	6	6	6	6	6	6	6	6	6	6
3	212105021.0	5021.0	ANTIOQUIA	ALEJANDRIA	6	6	6	6	6	6	6	6	6	6	6
4	213005030.0	5030.0	ANTIOQUIA	AMAGA	6	6	6	6	6	6	6	6	6	6	6

Figure 1: Historical CGR Category

There are 1014 individuals, one by each Colombian municipality and 15 variables. By checking the dataset, only 0.0005 of the data are nan values, this is a great advantage because the imputation methods will not significantly affect the distribution of the data. For apply the imputation methods first it is necessary to transform the categorical ordinal variables to continuous variables.
Since the dataset is only composed of categorical data, where only some of the municipalities have the special category, we considered analysing the behaviour of these municipalities independently. Therefore, the data set will be divided in 2, one data set for the municipalities between categories 1 to 6 and another for the municipalities that were once classified as "ESP" special.

	codigo_cgr	divipola	departamento	MUNICIPIO	cat_2012	cat_2013	cat_2014	cat_2015	cat_2016	cat_2017	cat_2018	cat_2019	cat_2020	cat_2021	cat_2022
0	210105001.0	5001.0	ANTIOQUIA	MEDELLIN	ESP	ESP	ESP	ESP	ESP	ESP	ESP	ESP	ESP	ESP	ESP
125	210108001.0	8001.0	ATLANTICO	BARRANQUILLA, DISTRITO ESP, INDUSTRIAL Y PORTU...	ESP	ESP	ESP	ESP	ESP	ESP	ESP	ESP	ESP	ESP	ESP
148	210111001.0	11001.0	CUNDINAMARCA	BOGOTA D.C.	ESP	ESP	ESP	ESP	ESP	ESP	ESP	ESP	ESP	ESP	ESP
149	210113001.0	13001.0	BOLIVAR	CARTAGENA DE INDIAS, DISTRITO TURISTICO Y CULT...	ESP	ESP	ESP	ESP	ESP	ESP	ESP	ESP	ESP	ESP	ESP
845	210168001.0	68001.0	SANTANDER	BUCARAMANGA	ESP	ESP	ESP	ESP	ESP	ESP	1	ESP	1	ESP	1
1005	210176001.0	76001.0	VALLE DEL CAUCA	CALI	ESP	ESP	ESP	ESP	ESP	ESP	ESP	ESP	ESP	ESP	ESP
779	210154001.0	54001.0	NORTE DE SANTANDER	CUCUTA	1	1	ESP	ESP	1	1	1	1	1	1	1

Figure 2: Municipalities of special category

As can be seen, all the municipalities, except for Cúcuta and Bucaramanga, have maintained their classification as special districts.

The conversion of these numerical data shows that the data are skewed to the left, i.e., there are much more municipalities in category 6 than in the other categories. This is easier to appreciate by using a bar chart with the frequency of each category for each year.

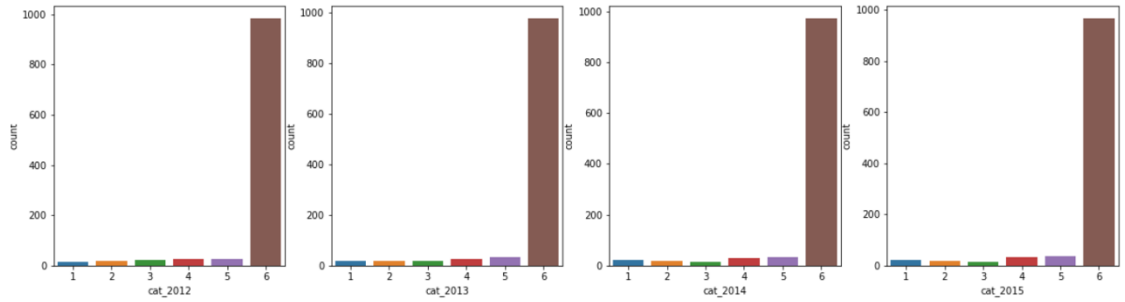


Figure 3: Distribution of municipalities 2012 - 2015

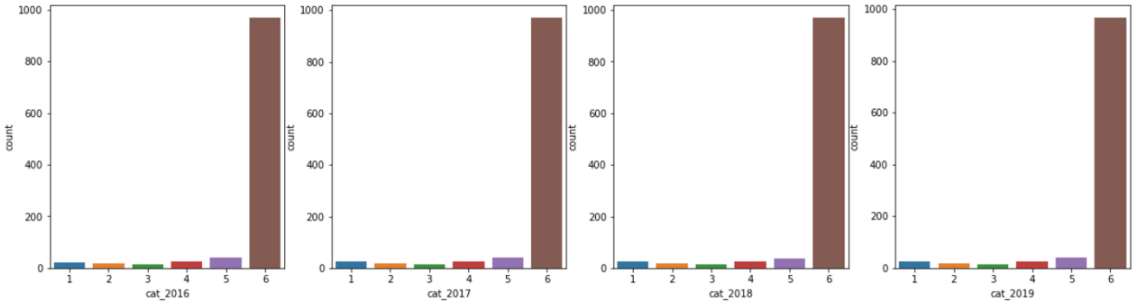


Figure 4: Distribution of municipalities 2016 – 2019

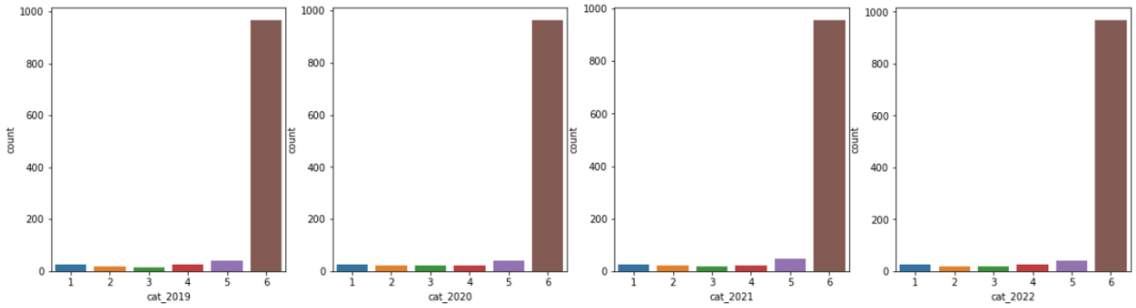


Figure 5: Distribution of municipalities 2019 – 2022

In general, we see a growing trend in category 1, that these municipalities are increasing their population and their total income, as well as the number of municipalities in category 2, which suffered a small drop in 2012, so it is expected that some of these municipalities moved up their classification to category 1. On the other hand, there is a primarily increasing trend in the municipalities classified in category 5, i.e., apart from the last year, many municipalities fell from category 4 to 5.

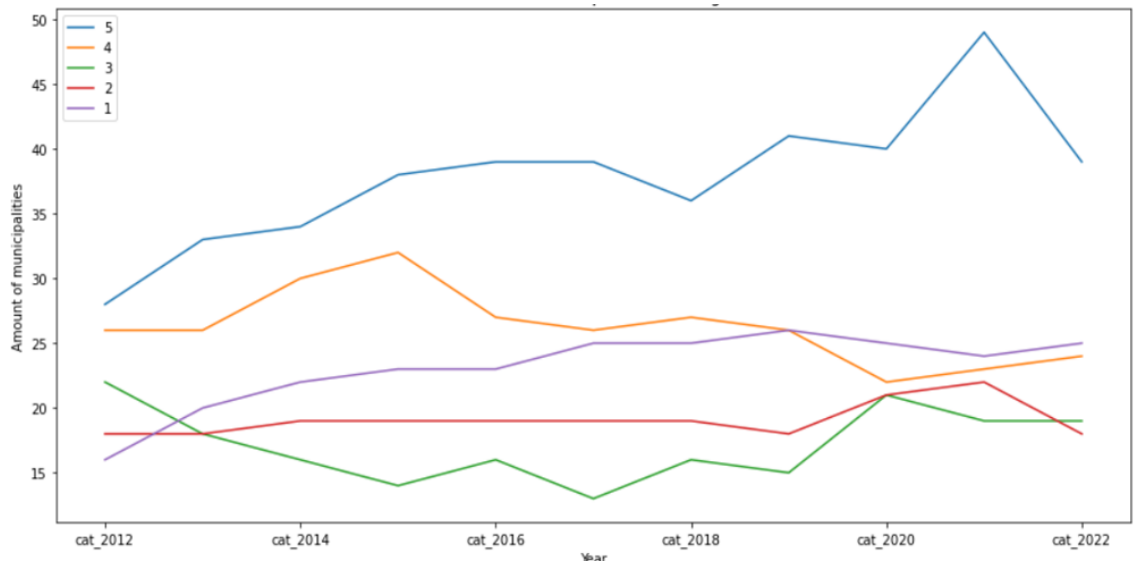


Figure 6: Historical number of municipalities on categories 5 to 1.

In other hand, the number of municipalities in category 6 has shown a decreasing trend, however, the drop in the number of municipalities in category 5 by 2022 is reflected in the increase of municipalities in category 6, which indicates that these municipalities now have a smaller population or lower income, or both by 2022, indicating a return in the growth of these municipalities to values close to those of 2016.

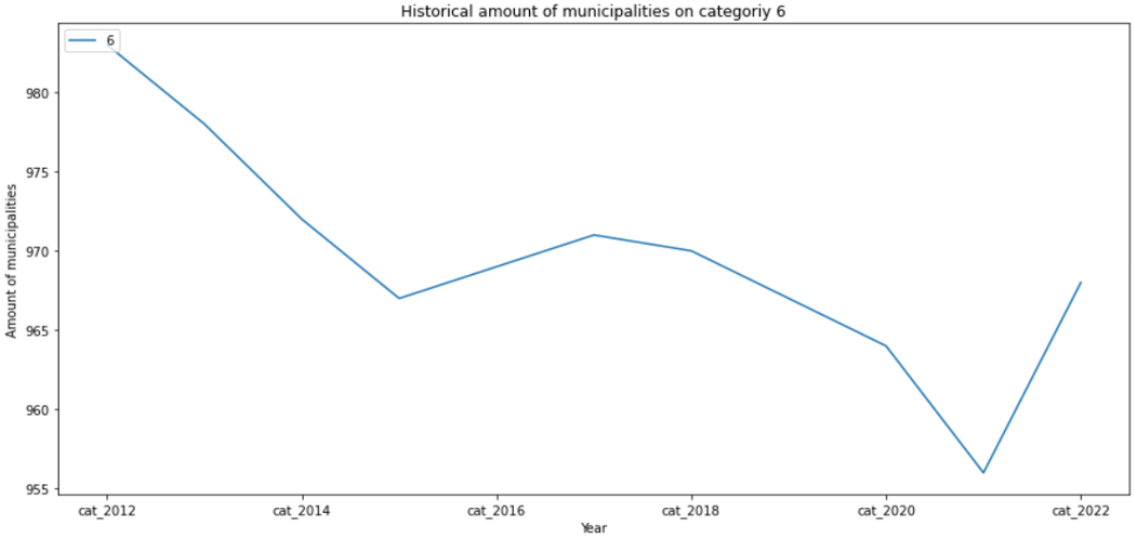


Figure 7: Historical number of municipalities on category 6.

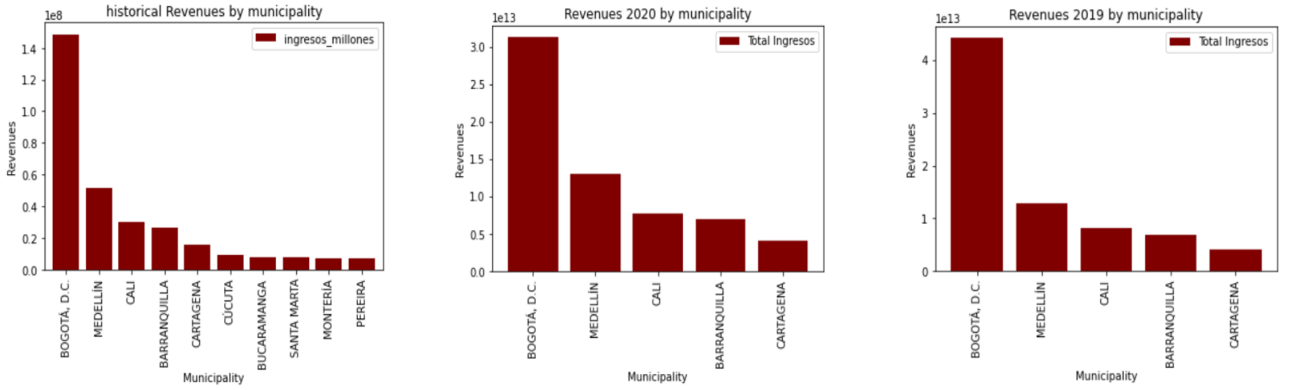
3.2 Revenues Analysis

This dataset contains 17 fields, that include historical information (2012-2021) of the income of the country's municipalities. Each row contains the budget and the incomes of the municipality, associated with its source (taxes, contributions, and fees).

	Código FUT	Nombre Entidad	Cód. DANE Departamento	Nombre DANE Departamento	Cód. DANE Municipio	Nombre DANE Municipio	Código Concepto	Concepto	Presupuesto Inicial	Presupuesto Definitivo	Recaudo	Sin Situación Fondos	Total Ingresos	Tiene Documento
326	210641206	COLOMBIA	41	HUILA	41206.0	COLOMBIA	TI.A.1.10	IMPUESTO DE ESPECTÁCULOS PÚBLICOS MUNICIPAL	200	0	0	0	0	No
558	210641206	COLOMBIA	41	HUILA	41206.0	COLOMBIA	TI.A.1.2	VEHÍCULOS AUTOMOTORES	3.500	750	724	0	724	No
647	210641206	COLOMBIA	41	HUILA	41206.0	COLOMBIA	TI.A.1.2.1	VEHÍCULOS AUTOMOTORES VIGENCIA ACTUAL	3.000	750	724	0	724	No

Figure 8: Municipalities overview.

We can observe that for the 'Total Ingresos' field, some rows contain null values and other characters like ' ', and some rows contain the sum of the total income for each municipality. Additionally, if we analyze the historical behavior of income by municipality, it can observe that the first municipalities with the highest historical revenues are Bogotá, Medellín, Cali, Barranquilla, and Cartagena. Moreover, these cities have maintained the same position according to the level of annual income, how is showed in the figure below.



(a) Municipalities overview (b) Municipality incomes 2020 (c) Municipality incomes 2019

Figure 9: Historical annual revenues by municipality

3.3 Operating expenses analysis

This dataset presents the operating expenses of the different departments and municipalities in billions of pesos. The key columns to determine the proper execution of each department's budget are:

- **Initial budget:** Estimated budget at the beginning of each year per municipality.
- **Final Budget:** Budget collected at the end of each year by municipality.
- **Commitments and obligations:** Represent debts payable and other financial commitments.
- **Payments:** The total amount of payments made on financial commitments and obligations.

Therefore, data on those columns are highly correlated.

Cód. DANE Departamento	Nombre DANE Departamento	Código Concepto	Concepto	Unidad Ejecutora	Presupuesto Inicial	Presupuesto Definitivo	Compromisos	Obligaciones	Pagos	Año
41	HUILA	1.1.1	SERVICIOS PERSONALES ASOCIADOS A LA NOMINA	ADMINISTRACIÓN CENTRAL	13488188000	11643058542	11643058542	11643058542	11643058542	2017
20	CESAR	1.1.1	SERVICIOS PERSONALES ASOCIADOS A LA NOMINA	ADMINISTRACIÓN CENTRAL	16706372655	15806177426	15062956022	15062956022	15062956022	2017
50	META	1.1.1	SERVICIOS PERSONALES ASOCIADOS A LA NOMINA	ADMINISTRACIÓN CENTRAL	20525824434	17259597520	17030332360	17030332360	17030332360	2017
47	MAGDALENA	1.1.1	SERVICIOS PERSONALES ASOCIADOS A LA NOMINA	ADMINISTRACIÓN CENTRAL	12753724635	15693247975	15693247975	15693247975	15002638259	2017
95	GUAVIARE	1.1.1	SERVICIOS PERSONALES ASOCIADOS A LA NOMINA	ADMINISTRACIÓN CENTRAL	6149719228	6520400692	6486522778	6465229758	6465229758	2017

Figure 10: Operating expenses overview.

When grouping the data by concept and year during the last 5 years the concept that generates the most budget is personnel expenses, being 2018 the year in which the budget reached its highest level. Similarly, we see that if we group this data by concept and total it, the largest amounts associated with payroll payments.

Año	Concepto	Presupuesto Inicial
2018	GASTOS DE PERSONAL	13607982976659
2020	GASTOS DE PERSONAL	9148477591924
2019	GASTOS DE PERSONAL	8623217205437
2018	SERVICIOS PERSONALES ASOCIADOS A LA NOMINA	8217466450269
2017	GASTOS DE PERSONAL	7076147664303
2018	TRANSFERENCIAS CORRIENTES	6589201132879
2020	TRANSFERENCIAS CORRIENTES	6168721272682
2019	TRANSFERENCIAS CORRIENTES	6012037144301
2020	SERVICIOS PERSONALES ASOCIADOS A LA NOMINA	5913742152564
2017	TRANSFERENCIAS CORRIENTES	5817256916957

Concepto	Presupuesto Inicial
TRANSFERENCIAS CORRIENTES	24587216466819
SERVICIOS PERSONALES ASOCIADOS A LA NOMINA	24115858241825
SUELDOS DE PERSONAL DE NOMINA	16336447687623
GASTOS GENERALES	12415157007478
MESADAS PENSIONALES	8905543285092
ADQUISICIÓN DE SERVICIOS	8438121159435
DE FUNCIONARIOS	8398006914521
CONTRIBUCIONES INHERENTES A LA NOMINA	7928986376243
SERVICIOS PERSONALES INDIRECTOS	6309848441952
APORTES DE PREVISIÓN SOCIAL	6114531922853

Figure 11: Initial Budget by year and concept.

Similarly, the financial obligations follow the same pattern as the initial budget; however, not all the budget executed by the end of each year. This is an indicator that in general, there are no cost overruns related to salary payments.

Año	Concepto	Obligaciones
2018	GASTOS DE PERSONAL	13108827427459
2020	GASTOS DE PERSONAL	8558805909974
2019	GASTOS DE PERSONAL	8171574663138
2018	SERVICIOS PERSONALES ASOCIADOS A LA NOMINA	7802640992710
2017	GASTOS DE PERSONAL	6743566615399
2018	TRANSFERENCIAS CORRIENTES	5714172443239
2019	TRANSFERENCIAS CORRIENTES	5691659229991
2018	SUELDOS DE PERSONAL DE NOMINA	5516455335871
2017	TRANSFERENCIAS CORRIENTES	5398253618543
2020	SERVICIOS PERSONALES ASOCIADOS A LA NOMINA	5380229599291

Concepto	Obligaciones
SERVICIOS PERSONALES ASOCIADOS A LA NOMINA	22369070014049
TRANSFERENCIAS CORRIENTES	21734772524656
SUELDOS DE PERSONAL DE NOMINA	15411610911165
GASTOS GENERALES	10472824705623
MESADAS PENSIONALES	7719709612882
DE FUNCIONARIOS	7677158584369
ADQUISICIÓN DE SERVICIOS	7577407385644
CONTRIBUCIONES INHERENTES A LA NOMINA	7259957719548
SERVICIOS PERSONALES INDIRECTOS	6863634782426
APORTES DE PREVISIÓN SOCIAL	5608471750998

Figure 12: Obligations by year and concept.

Even though a large part of the budget is allocated to payroll, in all years the amount of money paid for this item is much lower than the financial obligations. This is indicative of poor budget execution, as this value is always much higher than obligations and payments.

Año	Concepto	Saldos
2018	GASTOS DE PERSONAL	384846311780
	GASTOS GENERALES	324905180474
	TRANSFERENCIAS CORRIENTES	320504339153
2019	TRANSFERENCIAS CORRIENTES	304161852116
2017	TRANSFERENCIAS CORRIENTES	284082884337
2020	TRANSFERENCIAS CORRIENTES	269702186616
2017	GASTOS DE PERSONAL	219046047642
2020	GASTOS DE PERSONAL	204895942900
2019	GASTOS DE PERSONAL	200580735238
2018	ADQUISICIÓN DE SERVICIOS	183084183586

Figure 13: Balance by year and concept.

On the other hand, when analyzing the executing entities that have a greater financial obligation, we see that the central administration of the municipalities occupies a large part of these obligations, followed by the council and the "personería", the latter two being the planning and control agencies, respectively. To study the behavior of expenses, we proceed to look at the 3 concepts that make up 90% of their composition. Starting with staff costs, where the graph does not show a clear trend over time, however, there is an oscillation over time between 1.3 and 0.4 approximately.

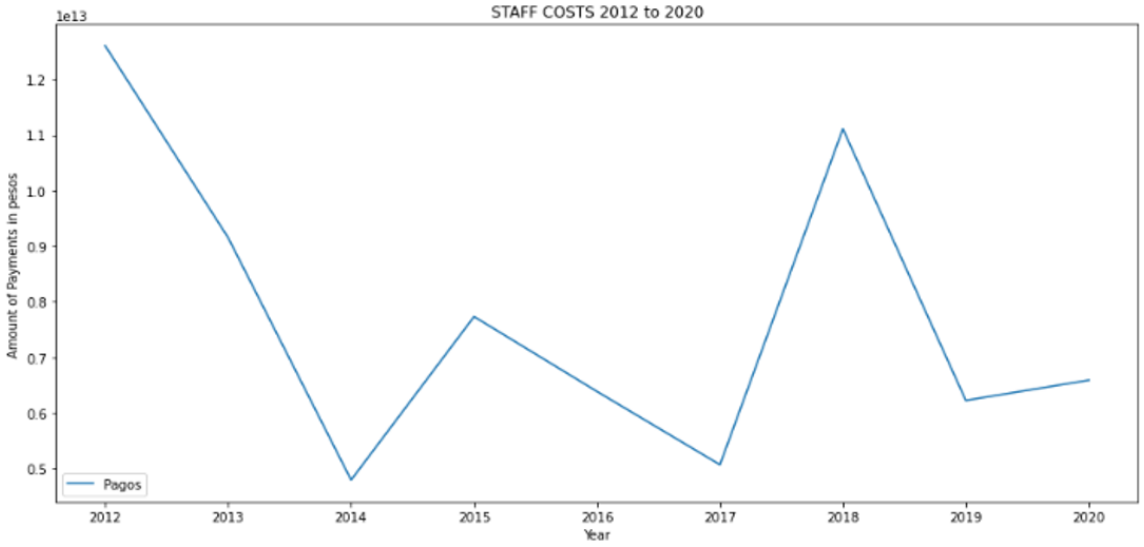


Figure 14: Staff costs 2012 to 2020.

The previously mentioned oscillation with lack of trend is also present in general expenses. In addition, it is observed that there is similarity in the oscillation, since both time series maintain peaks in years either close or the same, which would allow us to affirm that approximately 80% of the costs behave similarly in these two series.

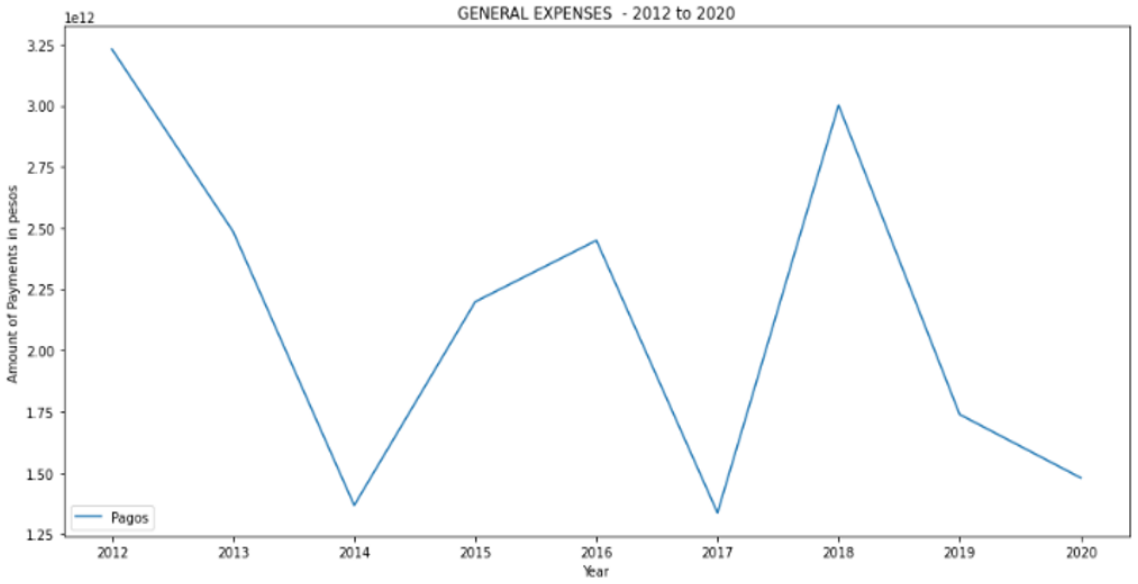


Figure 15: General expenses 2012 to 2020.

The last time series to be observed here, which among the three has less explanatory power, has a slightly different behaviour, maintaining a more reduced and constant oscillation range as of 2014, an interval in which it can be stated that there is a tendency to remain between 3 and 3.25 approximately. In addition, unlike the first two, there is a very noticeable maximum peak in 2013, indicating a possible outlier in that year.

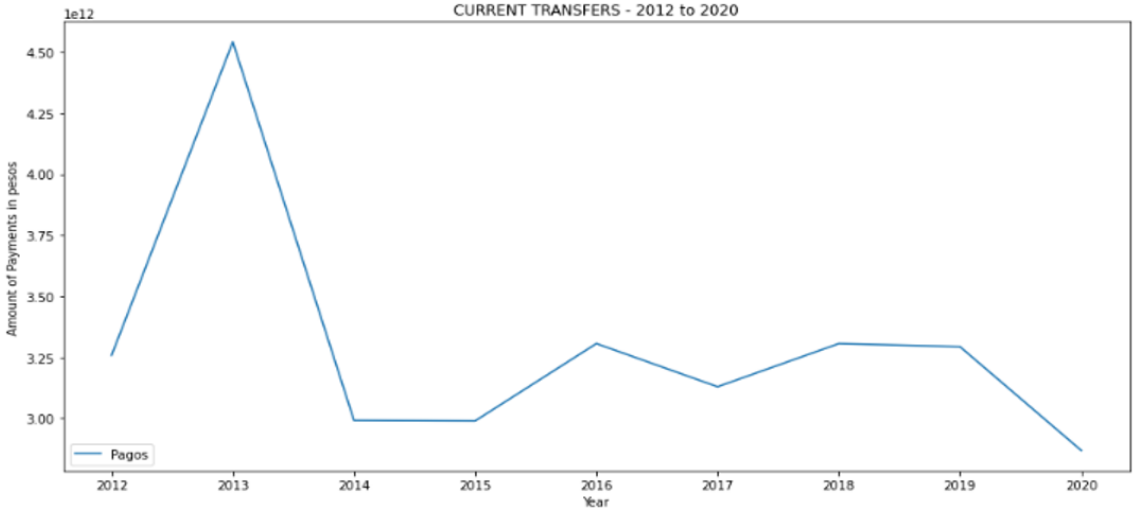


Figure 16: Current transfers 2012 to 2020.

Finally, it was decided to find a correlation matrix between the variables shown in the following table. The result is, as expected, a high correlation in such a way that an appropriate estimate of the budget made to pay all the respective obligations.

	Presupuesto Inicial	Presupuesto Definitivo	Compromisos	Obligaciones	Pagos
Presupuesto Inicial	1.000000	0.991654	0.990874	0.989790	0.988576
Presupuesto Definitivo	0.991654	1.000000	0.995852	0.993674	0.990938
Compromisos	0.990874	0.995852	1.000000	0.999510	0.998172
Obligaciones	0.989790	0.993674	0.999510	1.000000	0.998986
Pagos	0.988576	0.990938	0.998172	0.998986	1.000000

Figure 17: Expenses correlation matrix.

4 Data cleaning and feature engineering

4.1 Incomes

Since the incomes until the year 2016 are in miles of pesos and from 2017 in pesos, it is necessary to homologue the values. Additionally, we can observe that some numerical fields separated by characters like point, spaces, etc., and it is necessary to replace them. Also, for the dataset of the year 2021, point represents the decimal separator and require and individual cleaning. During the analysis, identified that the first municipalities with the highest historical revenues are Bogotá, Medellín, Cali, Barranquilla, and Cartagena. However, for the years from 2012 to 2016 an unusual behavior observed with some municipalities which shows incomes higher than the top five municipalities how is observed in the figure below. After plotting these cases separately, a similar pattern identified in that these municipalities reported their figures in pesos, when the format required reporting in thousands. Therefore, this data corrected.

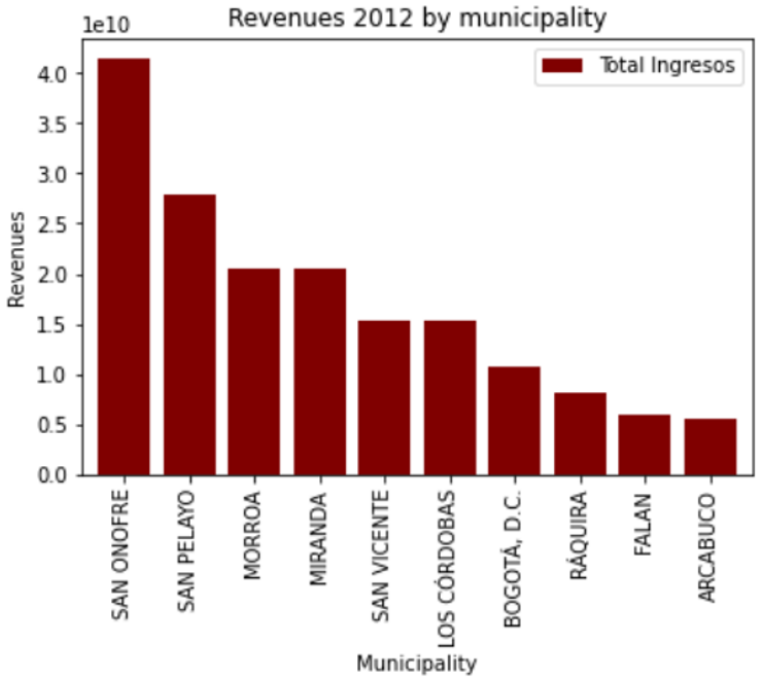


Figure 18: Revenues 2012 by municipalities.

For purposes of the analysis of fiscal performance, Law 617 of 2000 states that the current income of free destination for each municipality must be considered. Therefore, we only need these two fields in our final dataset, in addition to those that can identify each municipality and relate it to the expenditure dataset. On the other hand, unrestricted current income obtained after subtracting specific destination income from current income. In this sense, proceed as follows:

- Find the total income with a specific destination by municipality.
- Find the total current income by municipality.
- Subtract both values to find the total current income of free destination (ICLD)
- The total amount of payments made on financial commitments and obligations.

Once the previous procedure carried out, it identified that for some municipalities the value of the specific destination income exceeds the total income, obtaining negative free destination income. Given the unusual behavior, the following done to process this field:

- Add the category of each municipality to the dataframe.
- Separate the data in two groups, municipalities with positive free destination income and municipalities with negative free destination income.
- Find the media of specific destination incomes by category starting from the positive data
- Replace the negative values with the media found.

4.2 Operating Expenses

As with incomes, it was necessary to process the data for the operational expenses in such a way that they were standardized for all years and for all municipalities, therefore, for all data from 2012 to 2016, it was necessary to add 3 zeros, multiplying by 1000 to each of the financial columns **Presupuesto Inicial,Presupuesto Definitivo, Compromisos, Obligaciones, Pagos** prior to the conversion of these data from character strings to numerical data, since these data were in thousands of pesos.

Since the Nivel de concepto, i. the codigo de concepto, at the lowest levels 1,2,3..., aggregates the values of the other levels, it is necessary to split the dataset into multiple even smaller datasets that allow us to evaluate in a simpler way the financial ranges evaluated for each level. In addition to this and as seen previously, the Nivel de concepto column was generated to identify the length of the code of the concept to which each expense refers.

1. **Clean Outliers by Municipality**
   Becuase the outliers are mostly typing errors, it is necessary to adjust these data to match other data typed for the same Cód. DANE Municipio, for the same Código Concepto and for the same Código Unidad Ejecutora. In such a way that the detection of outliers is focused on a specific amount, so that this detection is not affected by outliers generated by different concepts, municipalities or executing units.

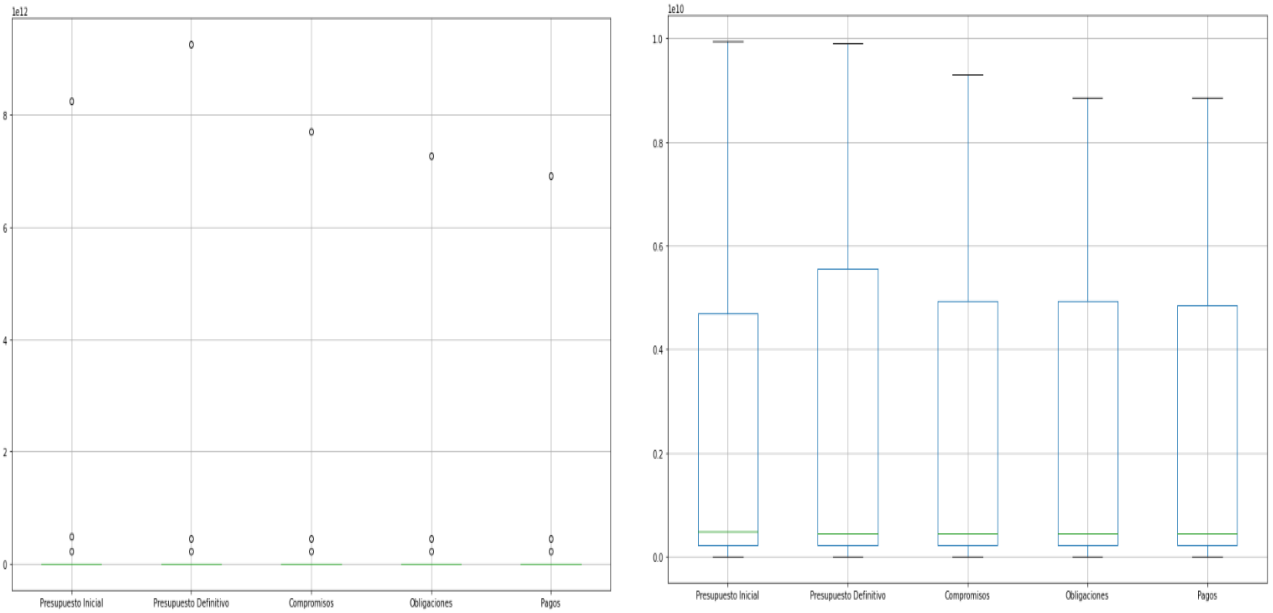
   To this end, the following function is generated, a series of for cycles in order to evaluate all the rows of the dataset, in this way we have that:

- The data is filtered by Cód. DANE Municipio, until all the municipalities are evaluated.
- Filter the previous result by codigo concepto, until all the concepts of that municipality are evaluated.
- The previous result is filtered by unidad ejecutora, until all the executing units of that concept code are evaluated.

For the identification of the outlier, the maximum given by the interquartile range is calculated.

$$maximun = Q_3 + IQR * 1.5$$

This means that all values higher than this value will be identified as outliers, however, since some outliers can be very close to the maximum and since the most common typing error involves typing more than one zero or one zero on the real value, only those values exceeding up to 9 times the value of the maximum will be rewritten. This is done by calculating how many times the value exceeds the maximum and dividing this value using a power of 10.



(a) Data previous the filter

(b) Data after the filter application

Figure 19: Data distribution comparison before and after filter by municipality application

2. **Clean Outliers by category**
   Although the results of the first filter are very effective in evaluating the outliers within each Cód. DANE Municipio, Código Concepto and Código Unidad Ejecutora. It is necessary to evaluate the outliers within each category to which each municipality belongs, since it is not logical that a category 4 municipality presents higher expenditures than the municipalities of the special categories or of the categories of a higher level.

   For this purpose and under the assumption that the higher level categories ESP, 1, 2, have a better control of the data, only the ouliers in categories 3, 4, 5 and 6 will be evaluated.

   To this end, the following function is generated, a series of for cycles in order to evaluate all the rows of the dataset, in this way we have that:

- The data is filtered by categoria, until all the categories are evaluated.
- Filter the previous result by Código Unidad Ejecutora, until all the unidades ejecutoras of that categories are evaluated.
- The previous result is filtered by Año, until all the years of that Unidad Ejecutora are evaluated.

For the identification of the outlier, the maximum given by the interquartile range is calculated.

$$maximun = Q_3 + IQR * 1.5$$

In this case, as with the previous filter, the maximums given by the IQR will be evaluated, with the difference that the value identified as outlier will be divided by 10 until its value is close to but not lower than the median. Without losing individuals values or that the rewrite values are affected by the outliers.

5 Frontend and backend infrastructure

The tool used to build the frontend is Google Data Studio, connected to CSV files that are related to a key value created with the year and code of the municipality. Also, the report embedded in a site web through an iframe. Both the HTML code and the CSS hosted on a server too. The report can be consulted in the following link: https://asocapitalesds4.julianorregoweb.com/

- **General view:** Shows the expenses by year, executing unit, and the behavior of the commitments, obligations, and payments by year.
- **Incomes vs expenses:** Shows current incomes vs operating expenses by year. Additionally, includes a filter to visualize the expenses classification by level.
- **Fiscal performance:** Shows the behavior of the Fiscal Performance Index by year and the categorization of the municipalities according to their level of performance.

Finally, the report includes filters by category, municipality, department, and year for each view.

Figure 20: Report structure.

6 Findings

In accordance with the needs identified in the entity and the exploratory analysis of the data, we proceed to describe some of the most relevant findings.

6.1 What types of operating expenses put the most pressure on revenues?

The types of expenses that most pressure revenues over the years in all municipalities tend to be the same, that is personnel expenses in the first place, followed by general expenses and, finally, current transfers. While for the last two there are increases and decreases over time, personnel expenses show a constant increase for the years analyzed. In the same sense, if the expense accounts analyzed at a higher level, the personal services associated with payroll, pension allowances, and the acquisition of services are the most representative accounts.

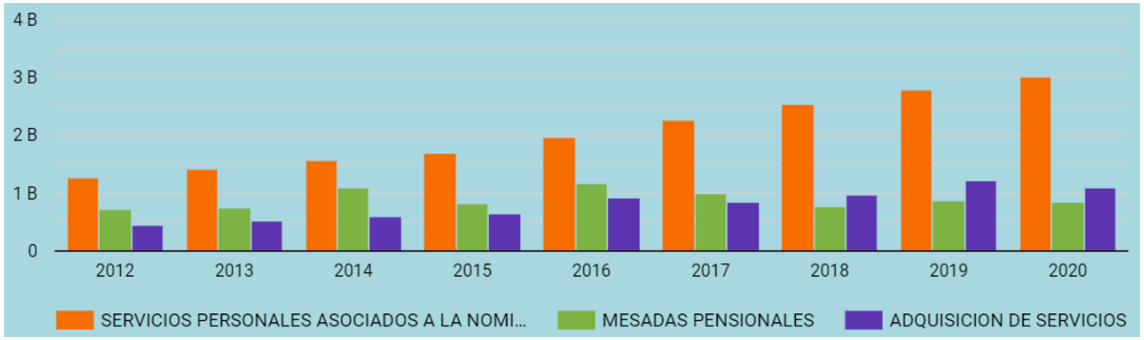


Figure 21: Pressure of operating expenses on revenues.

6.2 What categories of entities have more pressure on their operating expenses?

On average, the municipalities belonging to category 6 are the ones that have the most pressure on their operating expenses. If the behavior of the annual performance level analyzed, the entities belonging to this category located to a large extent in the vulnerable level, and to a lesser extent in the deterioration level (fiscal performance index below 40). The graph below shows the distribution of the municipalities in this category by the level of performance. Moreover, an analysis of the types of expenses of the entities belonging to this category makes it possible to identify that personnel expenses have a great impact on operating expenses, especially personal services associated with payroll and indirect personal services.

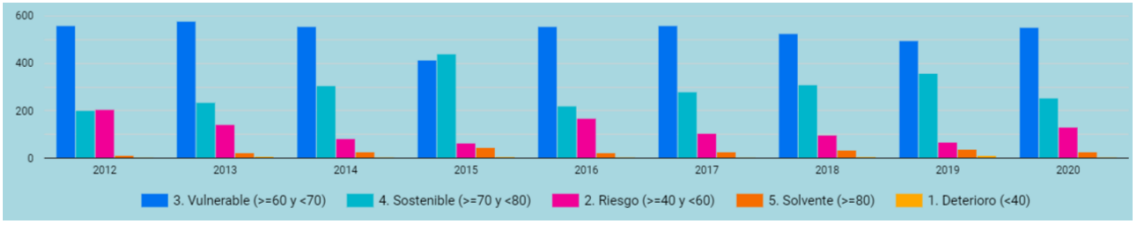


Figure 22: Level of performance category 6.

Likewise, it identified that the higher the category level, the lower the pressure exerted by operating expenses. That is, for the special category and category 1, a level of performance located in the sustainable and solvent ranges observed. This, in turn, has a possible relationship with the limits on operating expenses established in Law 617 of 2000, where the limit established for the aforementioned categories is stricter compared to categories 5 and 6.

6.3 Which expense units create more pressure on operating expenses? ¿What analysis can be done through categories?

In the first part, the higher the category, the greater the commitments that a municipality has. In the case of the commitments of the Central Administration, it is the unit with the most independent expenses of the selected category. Due to this, an analysis of the expenses will be carried out according to their category, ignoring the expenses of the Central Administration. In the entities belonging to category 1, the expenses of the municipal council exert greater pressure on income, followed by the Personería and the comptroller. Regarding the latter two, there is a change in the historical behavior in 2016. Thus, for the years prior to this date, the Comptroller's expenses were higher than those of the Personería and then became lower.

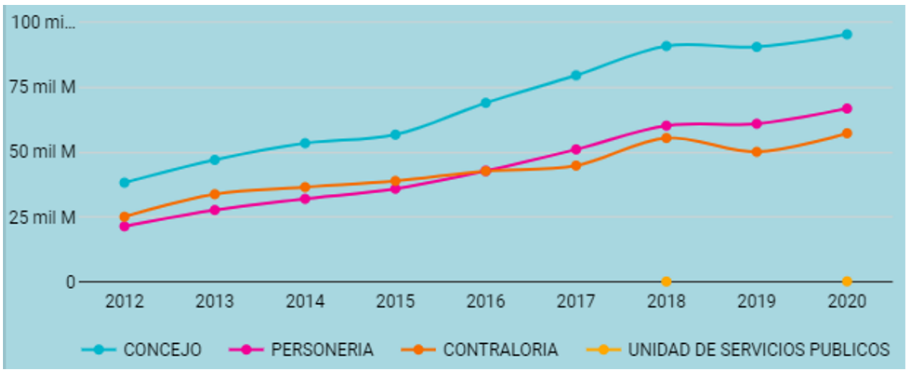


Figure 23: Expense Units by category 1.

In the entities belonging to category 2, a behavior similar to that of category 1 is observed, with greater pressure on expenses exerted by the Municipal Council, followed by the Personería, and finally the Comptroller.

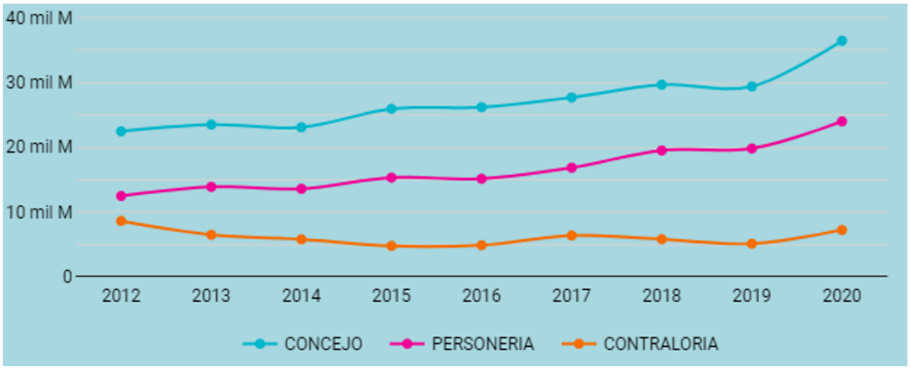


Figure 24: Expense Units by category 2.

In category 3, although the expenses of the Council continue to be the most relevant, an atypical behavior observed in the expenses of the Comptroller's Office, which were only presented for the year 2012.

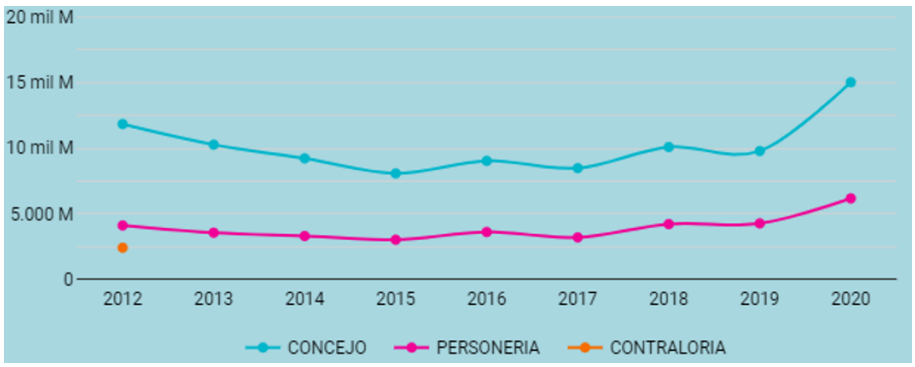


Figure 25: Expense Units by category 3.

For categories 4 and 5, the expenses of the Municipal Council continue to be the ones that exert the most pressure on income, being approximately double the expenses of the Personería. As for category 3, there are no expenses for the Comptroller's Office. However, a new Expenditure Unit presented, called the Public Services Unit, which has a significant behavior for entities in category 5, where there was an increase as of 2019, resulting in an expense similar to of the Personería for the year 2020.

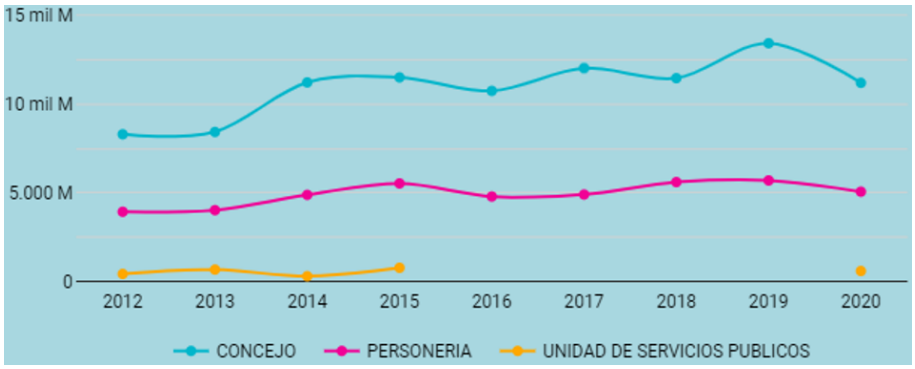


Figure 26: Expense Units by category 4.

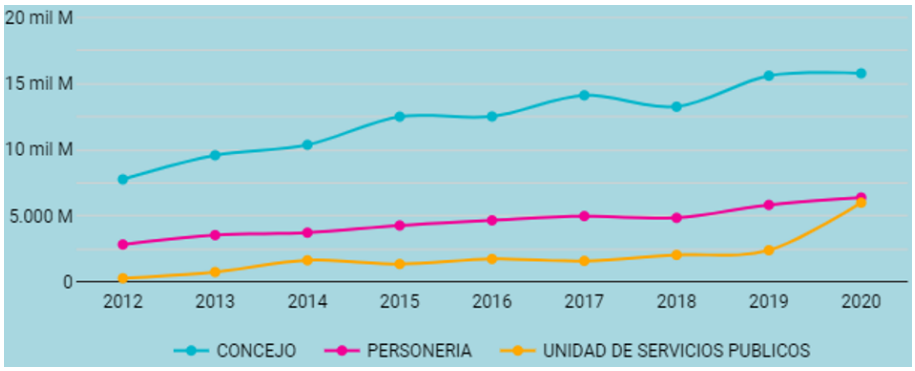


Figure 27: Expense Units by category 5.

For its part, category 6 presents a different behavior from the categories already analyzed. The expenses of the Comptroller's Office appear again, but they are lower than those incurred by the Public Service Units. However, the Council and the Personería continue to be the ones that exert the most pressure on income.

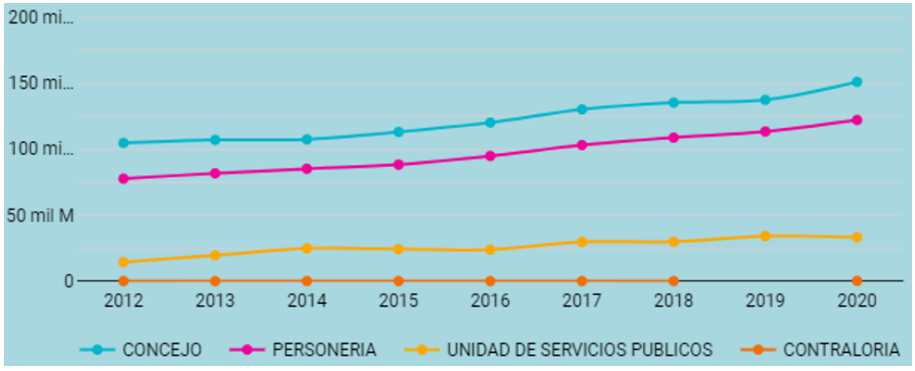


Figure 28: Expense Units by category 6.

Finally, contrary to the other categories, in category 6 the expenses of the Comptroller's Office take on greater relevance, being higher than those of the Personería and the Municipal Council. Likewise, a strong pressure exerted by the Public Services Unit is observed. However, this is only present in the city of Bogotá and is not representative of the behavior of the entire category.

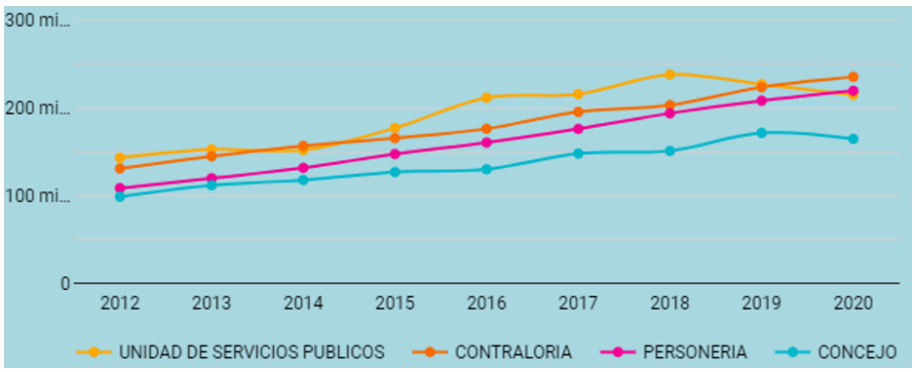


Figure 29: Expense Units by special category.

6.4 What types of operating expenses put the most pressure on revenues?

Doing a calculation of ratio values of expenses/(ICLD) "Current income of free destination", we can find that the years 2016 and 2017 were the two main years with the greatest ratio, meaning the more expensive in relation to the income.

Year	ICLD*	Expenses*	Expenses/ICLD (%)
2012	37.669	5.996	15.9
2013	44.163	6.743	15.2
2014	47.430	7.524	15.8
2015	51.248	7.907	15.4
2016	54.083	9.267	17.1
2017	59.703	10.044	16.8
2018	62.559	10.237	16.3
2019	71.051	11.569	16.2
2020	69.379	11.417	16.2

Table 1: Table of Expenses/ICLD

6.5 Fiscal performance index analysis per category

It's normal to have behavioral patterns as time goes by, for example, on the progression from the years 2019 and 2020 it's clear that 2020 was a decrease in all categories in comparison with 2019. Also, every single category has the IDF (fiscal performance index) decreased from 2019 to 2020 which makes sense considering that it was the pandemic lockdown beginning. There is a similar pattern from 2015 to 2016. And another observation completely unrelated to the preview's insight is the poor performance of the 6th category which has been constant with no significant improvement with a non-drastic outlier in the year 2015. It would be advised to take a second look at how the resources spent.

7 Model Selection and evaluation

7.1 Variable merging

During the model selection process, it was necessary to join all the variables under a single data frame, for this purpose a union between the data frame of income, expenditures, and the fiscal performance index made. The fiscal performance index generates the response variable. As can be seen, in order to gather all the data frames it was necessary to generate a Key column in which the DANE Code of the Municipality and the year concatenated.

	Nombre DANE Departamento	Cód. DANE Municipio	Nombre DANE Municipio	año	key	ICLD
0	AMAZONAS	91001.0	LETICIA	2012	91001_2012	24941077000
1	AMAZONAS	91001.0	LETICIA	2013	91001_2013	27789323000
2	AMAZONAS	91001.0	LETICIA	2014	91001_2014	29435176000
3	AMAZONAS	91001.0	LETICIA	2015	91001_2015	38667233000
4	AMAZONAS	91001.0	LETICIA	2016	91001_2016	36379531000

Figure 30: Incomes dataset.

Fuente Financiación	Presupuesto Inicial	Presupuesto Definitivo	Compromisos	Obligaciones	Pagos	Año	Nivel categoria	key	categoria
otros	4.836182e+10	6.225237e+10	5.668070e+10	5.288618e+10	5.288618e+10	2012	1.0	11001_2012	ESP
otros	5.018680e+10	5.417503e+10	4.224405e+10	3.508672e+10	3.508672e+10	2013	1.0	11001_2013	ESP
otros	7.426783e+10	7.426078e+10	3.985442e+10	3.471929e+10	3.471929e+10	2014	1.0	11001_2014	ESP
otros	7.395160e+10	7.499377e+10	5.240465e+10	4.651107e+10	4.651107e+10	2015	1.0	11001_2015	ESP
otros	7.872114e+10	7.813951e+10	5.460258e+10	4.716668e+10	4.716668e+10	2016	1.0	11001_2016	ESP

Figure 31: Operational expenses dataset.

	Año	Codigo DANE	Departamento	Municipio	indicador_fiscal	cod_desempeño	nivel_desempeño	key
0	2012	5	ANTIOQUIA	ADMINISTRACIÓN DEPARTAMENTAL	71.73	4	Sostenible	5_2012
1	2012	5001	ANTIOQUIA	MEDELLÍN	83.22	5	Solvente	5001_2012
2	2012	5002	ANTIOQUIA	ABEJORRAL	69.68	3	Vulnerable	5002_2012
3	2012	5004	ANTIOQUIA	ABRIAQUÍ	54.56	2	Riesgo	5004_2012
4	2012	5021	ANTIOQUIA	ALEJANDRÍA	59.52	2	Riesgo	5021_2012

Figure 32: Fiscal index dataset.

7.2 Correlation between variables

Since the data for the year 2021 are outliers, they are eliminated from the evaluation process, as well as the Initial Budget, Final Budget, Commitments, and Payments columns, since they have a correlation close to 1.

	Presupuesto Inicial	Presupuesto Definitivo	Compromisos	Obligaciones	Pagos
Presupuesto Inicial	1.000000	0.998412	0.997942	0.997397	0.997430
Presupuesto Definitivo	0.998412	1.000000	0.999469	0.999046	0.998809
Compromisos	0.997942	0.999469	1.000000	0.999810	0.999483
Obligaciones	0.997397	0.999046	0.999810	1.000000	0.999530
Pagos	0.997430	0.998809	0.999483	0.999530	1.000000

Figure 33: Correlation between expenses variables.

7.3 Generation of dummy variables

Because the index of individuals is constantly repeated, since each municipality has a concept code, i.e., it has more than one type of obligation per year, and since we do not want to lose this data by not evaluating them. A series of dummy variables generated for each concept code.

	1	2	4	5	6	7	8
0	0	0	0	0	0	1	0
1	0	0	0	0	0	1	0
2	0	0	0	0	0	1	0
3	0	0	0	0	0	1	0
4	0	0	0	0	0	1	0
5	0	0	0	0	0	1	0
6	0	0	0	0	0	1	0
7	0	0	0	0	0	1	0

Figure 34: Obligations categories dummy variables.

Subsequently, using matrix algebra, the value of each obligation per concept code multiplied by the dummy variables, in order to increase the dimensionality of the data frame, without the interaction of multiple concepts of the same individual generating an effect on the variance of a single variable.

key	categoria	obg_administracion	obg_consejo	obg_contraloria	obg_personeria	obg_educacion	obg_salud	obg_servicios
2012	ESP	0.0	0.0	0.0	0.0	0.0	5.288618e+10	0.0
2013	ESP	0.0	0.0	0.0	0.0	0.0	3.508672e+10	0.0
2014	ESP	0.0	0.0	0.0	0.0	0.0	3.471929e+10	0.0
2015	ESP	0.0	0.0	0.0	0.0	0.0	4.651107e+10	0.0
2016	ESP	0.0	0.0	0.0	0.0	0.0	4.716668e+10	0.0
...	...	...	...	...	...	...	...	...
2014	6	0.0	0.0	1000.0	0.0	0.0	0.000000e+00	0.0
2015	6	0.0	0.0	0.0	0.0	0.0	0.000000e+00	0.0
2015	6	0.0	0.0	0.0	0.0	0.0	0.000000e+00	0.0
2015	6	0.0	0.0	1000.0	0.0	0.0	0.000000e+00	0.0

Figure 35: Dimensionality Augmentation of obligation categories.

The dataframe is then grouped by municipality and year, and these results added together. This measure will not cause any effect on the data, the dataframe will only be affected in its dimensionality.

categoria	obg_administracion	obg_consejo	obg_contraloria	obg_personeria	obg_educacion	obg_salud	obg_servicios
ESP	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	6.200632e+09	0.0
ESP	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	9.337273e+09	0.000000e+00	0.0
ESP	0.000000e+00	0.000000e+00	0.000000e+00	1.232960e+10	0.000000e+00	0.000000e+00	0.0
ESP	0.000000e+00	0.000000e+00	2.559250e+10	0.000000e+00	0.000000e+00	0.000000e+00	0.0
ESP	0.000000e+00	1.500020e+10	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.0
ESP	3.825010e+11	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.0

Figure 36: Final expenses by concept 1 data frame.

A similar process followed for the creation of dummy variables for the historical category to which all municipalities belong.

7.4 Model selection

Given the small size of the final dataset, and the fact that we want to evaluate in some way the temporal change of municipalities' finances over time, without running a time series for more than 1200 individuals, we divide the dataset randomly using a 20% as the testing set.

```
from sklearn.model_selection import train_test_split
df_19, df_20= train_test_split(df, test_size=0.2)
```

Figure 37: Data splitting.

1. **Normalization**
   Although a classification model will be used, and these are not affected by the ranges of the variables, it was decided to normalize the data using min-max, in order to be able to carry out subsequent tests with different methods, such as neural networks, as well as being a good practice.

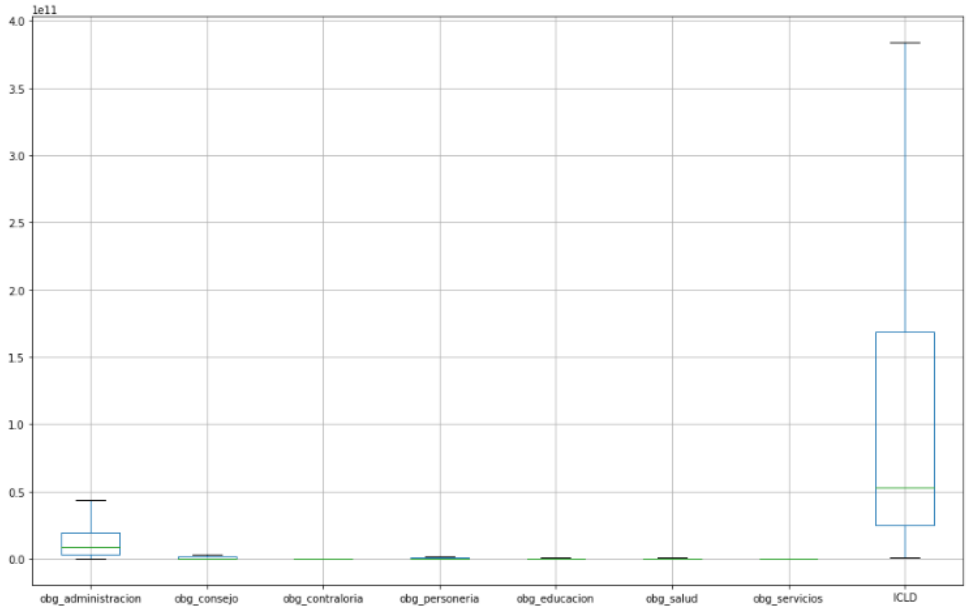


Figure 38: Boxplot numerical variables model dataframe.

Cód. DANE Municipio	nivel_desempeño	cat_ESP	obg_administracion	obg_consejo	obg_contraloria	obg_personeria	obg_educacion	obg_salud	obg_servicios	ICLD
5001	Solvente	1	0.139097	0.164817	0.180521	0.091983	0.088831	0.085331	0.000000	0.139837
5001	Solvente	1	0.131713	0.168302	0.193530	0.090952	0.081663	0.085190	0.000000	0.161060
5001	Solvente	1	0.144446	0.166945	0.207011	0.096210	0.087119	0.109147	0.000000	0.173391
5001	Solvente	1	0.162041	0.184329	0.214451	0.103206	0.093450	0.120695	0.000000	0.188366
5001	Solvente	1	0.177548	0.205624	0.229022	0.135944	0.107992	0.145619	0.000000	0.197023
...	...	...	...	...	...	...	...	...	...	...
99624	Riesgo	0	0.000285	0.000949	0.000000	0.000687	0.000000	0.000000	0.000000	0.000359
99624	Riesgo	0	0.000432	0.001056	0.000000	0.000852	0.000000	0.000000	0.000000	0.000428
99773	Riesgo	0	0.000884	0.001544	0.000000	0.000478	0.000000	0.000000	0.000000	0.001823
99773	Riesgo	0	0.000787	0.001566	0.000000	0.000597	0.000000	0.000000	0.000000	0.002238
99773	Riesgo	0	0.000887	0.001877	0.000000	0.000757	0.000000	0.000000	0.001708	0.003277

Figure 39: Model dataframe normalized.

2. **Dimensionality Reduction**
   Similarly, the categorical variables disappear, except for the one representing the special category, since, as shown below, they do not have a significant effect on the first two factorial planes, which explains 78.7% of the variance.

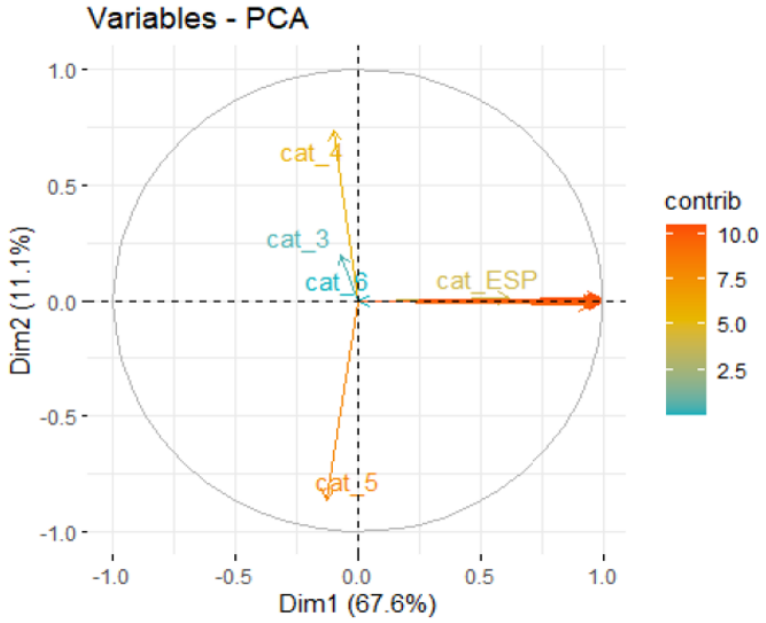


Figure 40: Contribution circle by variables DIM1 vs DIM2.

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5
Category 3	0.06	2.97	69.90	4.33	22.26
Category 4	0.12	40.51	24.78	4.54	29.24
Category 5	0.21	56.45	5.25	4.29	32.57
Category 6	0.00	0.00	0.00	0.00	0.00
Category ESP	4.70	0.01	0.02	70.82	14.95
Obg. Administracion	12.06	0.00	0.00	0.06	0.02
Obg. Consejo	12.21	0.00	0.00	0.13	0.31
Obg. Contraloria	12.09	0.01	0.01	0.10	0.12
Obg. Personeria	12.18	0.00	0.01	0.93	0.00
Obg. Educacion	11.61	0.00	0.01	2.99	0.51
Obg. Salud	11.65	0.00	0.00	1.66	0.01
Obg. servicios	11.09	0.03	0.03	9.62	0.02
ICLD	12.02	0.01	0.00	0.55	0.00

Table 2: Contribution by variables DIM 1 to DIM 5.

7.5 Model testing and results

1. **Classification models**
   To identify the best fitting model to predict the performance level variable, a Grid search used to quickly evaluate different **hyperparameters** specified in a dictionary on different models. The classification models evaluated were **Random Forest**, **KNN**, **Adaboost**, **Bagging Classifier**, since the data ere not balanced. A sampling is performed on category 6 and the following results obtained for the **accuracy**, **precision**, **recall**, and **F1** metrics. It can be clearly seen that the best results generated by a **KNN**, with $K = 20$

	Random Forest	Adaboost	Bagging	K-Means
Accuracy	0.76	0.62	0.62	0.73
Precision	0.75	0.45	0.45	0.71
Recall	0.76	0.62	0.62	0.73
F1	0.75	0.52	0.52	0.71

Table 3: Classification models metrics results.

2. **Deep Neural network model**
   Similarly, and thanks to the fact that the data were normalized, neural networks used, and the following result obtained, evaluating only the accuracy of the training data, using batches of 32 individuals.

```
Epoch 1/15
46/46 [==============================] - 1s 2ms/step - loss: 2.1665 - accuracy:
0.4534
Epoch 2/15
46/46 [==============================] - 0s 2ms/step - loss: 1.0982 - accuracy:
0.5837
Epoch 3/15
46/46 [==============================] - 0s 2ms/step - loss: 0.9493 - accuracy:
0.6022
Epoch 4/15
46/46 [==============================] - 0s 2ms/step - loss: 0.9352 - accuracy:
0.6145
Epoch 5/15
46/46 [==============================] - 0s 2ms/step - loss: 0.9138 - accuracy:
0.6104
Epoch 6/15
46/46 [==============================] - 0s 2ms/step - loss: 0.9091 - accuracy:
0.6084
Epoch 7/15
46/46 [==============================] - 0s 2ms/step - loss: 0.8965 - accuracy:
0.6187
Epoch 8/15
46/46 [==============================] - 0s 2ms/step - loss: 0.8895 - accuracy:
0.6180
Epoch 9/15
46/46 [==============================] - 0s 2ms/step - loss: 0.8837 - accuracy:
0.6139
Epoch 10/15
46/46 [==============================] - 0s 2ms/step - loss: 0.8649 - accuracy:
0.6235
Epoch 11/15
46/46 [==============================] - 0s 2ms/step - loss: 0.8576 - accuracy:
0.6358
Epoch 12/15
46/46 [==============================] - 0s 2ms/step - loss: 0.8466 - accuracy:
0.6385
Epoch 13/15
46/46 [==============================] - 0s 2ms/step - loss: 0.8368 - accuracy:
0.6502
Epoch 14/15
46/46 [==============================] - 0s 2ms/step - loss: 0.8260 - accuracy:
0.6543
Epoch 15/15
46/46 [==============================] - 0s 2ms/step - loss: 0.8200 - accuracy:
0.6584
```

Figure 41: Deep Neural network training results.

By using the following architecture, 65% accuracy is obtained:

```
model = tf.keras.Sequential([normalizer,
                             tf.keras.layers.Dense(120, activation = 'relu'),
                             tf.keras.layers.Dense(60, activation = 'relu'),
                             tf.keras.layers.Dense(30, activation = 'relu'),
                             tf.keras.layers.Dense(15, activation = 'softmax')
                             ])
```

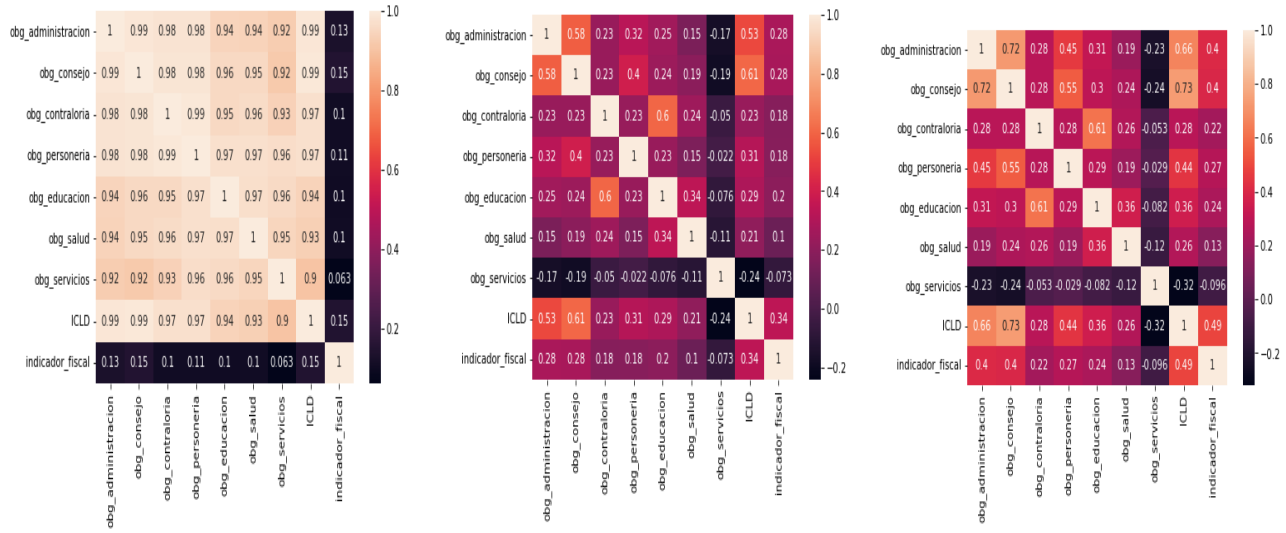
Figure 42: Deep Neural network architecture.

3. **Regression methods models**
   Analogously to the first one to the search of the classification models and using the normalized data. Using the Grid Search with different **hyperparameters**, but this time the fiscal performance index as a response variable. The following parameters obtained, on the **Random Forest**, **Decision Trees**, **daboost**, and **MLS** regression models. The low

	Random Forest	Decision Tree	Adaboost	Linear Reg
R^2	0.11	0.02	0.03	-0.18
MSE	0.03	0.03	0.03	0.04
MAE	0.14	0.15	0.15	0.16

Table 4: Regression models metric results.

R2 results of these models are mainly due to the low correlation between these variables, as will be seen below.



(a) Pearson correlation matrix (b) Kendall correlation matrix (c) Spearman corr. matrix

Figure 43: Correlation arrays between numerical variables ans response variable

It is observed that under the Pearson criterion there is a multicollinearity between all the variables and a low correlation with the variable response indicador fiscal, which indicates that a linear model is not the most appropriate one.

Similarly, the Kendall and Spearman correlation criteria based on the correlation between the range of the variables do not yield a significant correlation, which is an indication that other regression models will not yield significant results.

8 Model selection conclusion

In conclusion, from the model selection,taking as reference only the evaluated variables after the dimensionality reduction, it can be said that all the expenses(obligations) concepts and incomes (ICLD) represent a greater influence to classify the performance of a municipality, at the same time if those municipalities belong or not to the **ESPECIAL** category, the other categories have little or no influence.

It can be seen that the models that obtain the best results are the classification models, especially, **Random forest** and the **KNN**.For the case of K-means with $K = 20$, K represents the number of clusters in the data set, i.e. the data are grouped according to how similar they are to each other. For these, the distance between the individuals and the other items is calculated.

On the other hand if a classification model is generated using a **Random forest** sampling the category 6, it is possible to obtain results close to 80%, for the metrics of accuracy, precision, recall and F1.

Metric	Score
Accuracy	0.76
Precision	0.75
Recall	0.76
F1	0.75

Table 5: Random Forest results without category 6.

With the following hyperparameters:

Hyperparameter	Value	Description
n estimators	177	Number of trees to be used
min samples split	6	Number of samples required to split
min samples leaf	1	Min. samples required to be at a leaf node.
max features	8	The number of features to consider the split.
max depth	10	The maximum depth of the tree.
criterion	Gini	The function to measure the quality of a split.
Bootstrap	True	Whether samples are drawn with replacement

Table 6: Random Forest hyperparameters

9 References

1. Information on operating expenses of territorial entities 2012-2021 (Formato Unico Territorial - Contaduria Nacional), More information here.

2. Information on revenues of territorial entities 2012 - 2012 (Formato Unica Territorial - Contaduria Nacional)More information here.

3. Fiscal Performance Index (DNP) : It establishes the evaluation of public policies, the follow-up of development plan goals and the strengthening of results-oriented management at the national and territorial levels.More information here.

4. Information on categories Law 617 of 2000 (CGR): Defines norms tending to strengthen decentralization are and enacting rules for the rationalization of national public expenditure.More information here.

5. The Territorial Fiscal Gap in Colombia: Presents an estimation of expenditure needs and fiscal capacity of Colombia's municipalities and departments.More information here.

6. Índice de Capacidad Estadística Territorial (ICET): Is a multidimensional and systemic indicator that measures territorial statistical capacity.More information here.

7. Projections and back projections of the municipal population: Presents the municipal population for the period 1985-2017 and 2018-2035 based on the 2018 National Census.More information here.

8. Information current incomes: These are permanent revenues generated from the normal activity of the entity and are classified as tax and non-tax. Tax revenues come from the collection of taxes, while non-tax revenues come from sales of goods, provision of services, fines, among others.More information here

9. History of the category assigned to the territories: The category of a territory is determined by population size and current income, this is determined each year and therefore the budget of each territory.

10. Github Code Repository.