Estimation modelling of project development for Skit Consulting LTDA

Bogotá - Team 70

Lina Calvo | Juan David Durán | Javier Pinilla | Christian Pinzón | Jody Riaño |
Wilson Sandoval | Julio Valencia

Abstract

In Colombia, the software and IT services are one of the industries with the biggest grow, and their applications are mainly for the financial and accounting sectors, thus, the demand for software projects is increasing as well. However, there is always a problem with projects and that is an accurate estimate of the resources needed to complete these projects effectively, which results in delays, cost overruns, and cancelation of projects. Considering this, in the following project we aim to use data science in order to create models to predict the resources needed for new projects as well as giving the keys for a successful endeavor, based on a company's history and the statistics they keep for their completed tasks, to be able to develop an application useful for a company to understand the statistics of their past projects, predict the resources needed for new ones and assess their viability so that the company becomes more efficient and productive.

Content

1 Business Problem

During the last 10 years in Colombia, the IT market has grown at a rate of 18%; the software sector grew 19.1% and IT services grew 15.4%. According to the International Data Corporation in Colombia, the industry has doubled its sales in the last 7 years and, in 2017, it reached 9.5 billion dollars, divided into: hardware (56.5%), IT services (32.2%) and software (11, 4%). Colombia has a strong and growing domestic demand, being the sectors with the highest IT spends: the industrial sector, the government, the financial sector, and the agricultural sector. (Velneo, 2019)

In the capital and its area of influence, financial software is mainly developed for the banking industry. The biggest bet in Bogotá is to offer financial software and IT services, given the large number of banks that have concentrated in this city. Large-scale software projects are being developed due, in part, to Bogotá's abundant human capital, dominated by the most relevant corporate software platforms such as SAP, ORACLE, Microsoft, or IBM. (Velneo, 2019)

On the other hand, it should be noted that the software industry brings together more than 400 companies and that it has more than 30 years of work, making it a powerful productive sector, since "according to the latest figures from the IT Observatory, the software sector sold COP $13, 5 trillion in 2016 and it is employing 109,000 people in the IT industry, software and other related services ", Paola Restrepo, president of Fedesoft, told the newspaper La República. (Colombia Bring It On, 2020)

As for companies of this type registered, according to Fedesof, 82% are micro-enterprises, 13% small, 3% medium and only 2% large, which nevertheless contribute 74% of the revenue. (Portafolio, 2018)

As time has passed the need to evolve in the processes that companies have has increased, since some of them consume a large volume of the employees' time and they are activities that can be automatic with the implementation of software. Which would allow employees who have been developing these activities, to be oriented to carry out different activities that are aligned with the company's strategic objective.

Skit Consulting is a software company with 15 years of experience in the market. Its main clients are companies in the financial and accounting sector, to whom it offers products for reconciliation, operational risk, payment button and IT consulting services, flexible plant and software factory. Skit is also in charge of offering assistance in any other problem by developing custom software if the client requires it.

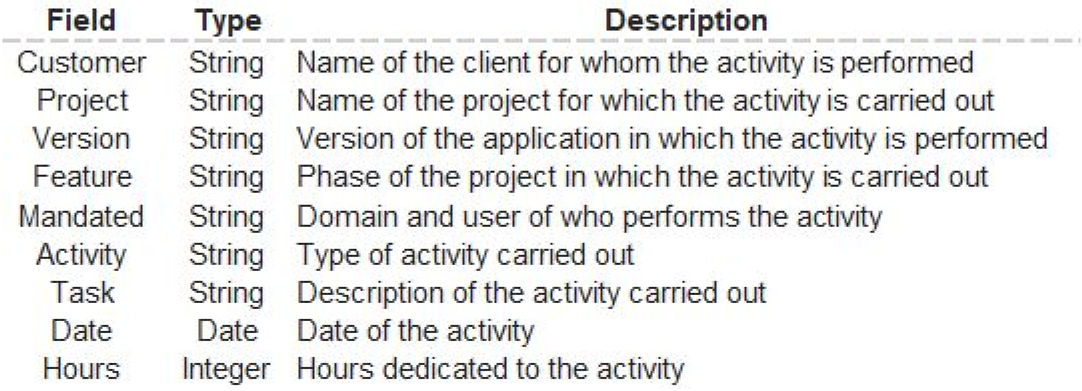
However, in this type of software development services, there is always a constant problem, and that is that the contract is made at the beginning of the project with certain product, time and price specifications, which are sometimes difficult to estimate, and at the same time In the end, many more resources can be spent than was initially thought and the price given may be below what the project actually ended up costing.

The company has a repository of historical information, which allows it to evaluate the time invested in each of the projects developed. This data is also used for basic historical analysis, but not for the identification of trends or prospective analysis. It is expected that this information will be used to make estimates of different types of projects, identifying which resources and time (cash and calendar) we require for their development.

It is necessary to identify trends or prospective analysis of history that the entity has of the projects that have ended (complete closed cycle), in order to make estimates of the different types of projects, in which the resources and time needed for their development are identified. This, in order to improve the productivity, performance and quality of their projects, with their respective effect on economic efficiency, in addition to identifying what made the difference in successful projects to replicate these practices in the future. This, along with the statistics presented above, in which is shown that this industry is one of the most important in Colombia could represent

1.1 Data Description

An information repository is available with data from 2012, starting with the data from the *SKIT Activities Application* and other company information sources. The SKIT Activities data is digitized by each collaborator of the organization, through a web application, where it is registered Date, Project, Version, Characteristic, Type of activity, Description and Hours invested for each activity developed.



Field	Type	Description
Customer	String	Name of the client for whom the activity is performed
Project	String	Name of the project for which the activity is carried out
Version	String	Version of the application in which the activity is performed
Feature	String	Phase of the project in which the activity is carried out
Mandated	String	Domain and user of who performs the activity
Activity	String	Type of activity carried out
Task	String	Description of the activity carried out
Date	Date	Date of the activity
Hours	Integer	Hours dedicated to the activity

Table 1. Variables

1.2 Methods

Sometimes when it is decided to implement analytical models it is believed that it is possible to carry them out with a large volume of information, however different stages must be addressed in order to generate a positive impact with the developed model and always have the clear objective of what you want to improve with the proposed model. The stages are:

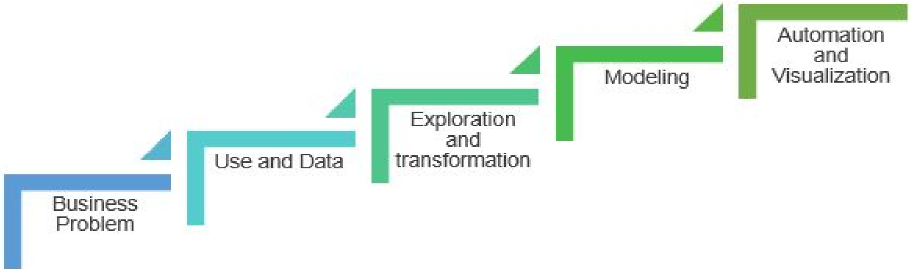


Figure 1. Project stages

The first stage is already addressed at the beginning of the document and is the business problem, where it seeks to be clear which is the objective of the project. In the second stage, the usability of the model is analyzed (viability) and the data available in order to determine whether the information is sufficient to solve the problem posed. The third stage is where the available data are known, and an exploratory analysis is made to see the basic statistics and identify outliers. We also make the cleaning of the available data applying some transformations to leave the data available for modeling. The fourth stage of modeling we have the creation of additional variables that can be built with the information available, for example: average number of people responsible for a specific client, average number of versions, etc.

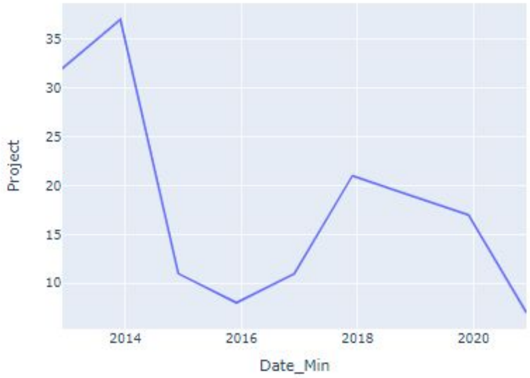
In this stage we also carry out the construction of the analytical model to be used, which with the information available at the time, we consider feasible to implement an supervised learning algorithm. Although we have a target variable to predict, we still do not have labeled information that may be useful for this purpose. However, we believe that a clustering algorithm as the k-means could provide useful information for the entity's purposes. And the last stage is automatization and visualization that will be shown later.

2 Exploratory Data Analysis and Cleaning Data

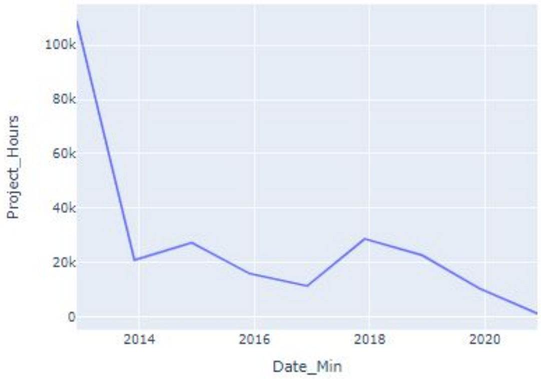
2.1 Description Analysis

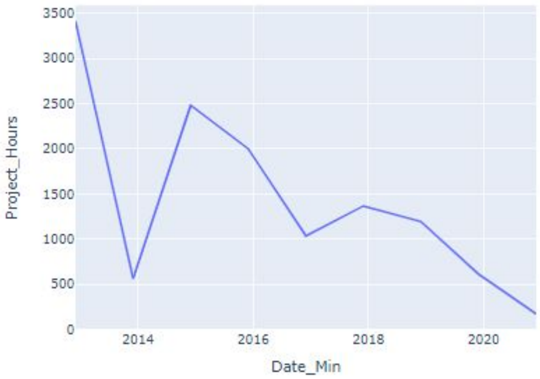
The original base has a total of 70.132 observations and 15 variables, however of those 15 variables, 5 variables are duplicated information (Year, Month, Day, Person in charge, Percentage), 2 are key variables (client name, project) and we have 8 variables for analysis. One of the most important findings of the analysis is the high concentration in 0's of the variable "planned hours", when reviewing these data with the entity, it was concluded that 0 means that it does not have information available, so it could not be used as an element for the construction of the project's target variable, then it was defined to use as the project's target variable "hours executed". On the other hand, the review of the percentage of missing variables of the base is made, and it is found that all have a completeness of 100%.

- Total projects per year started



- Total and average hours executed per year started



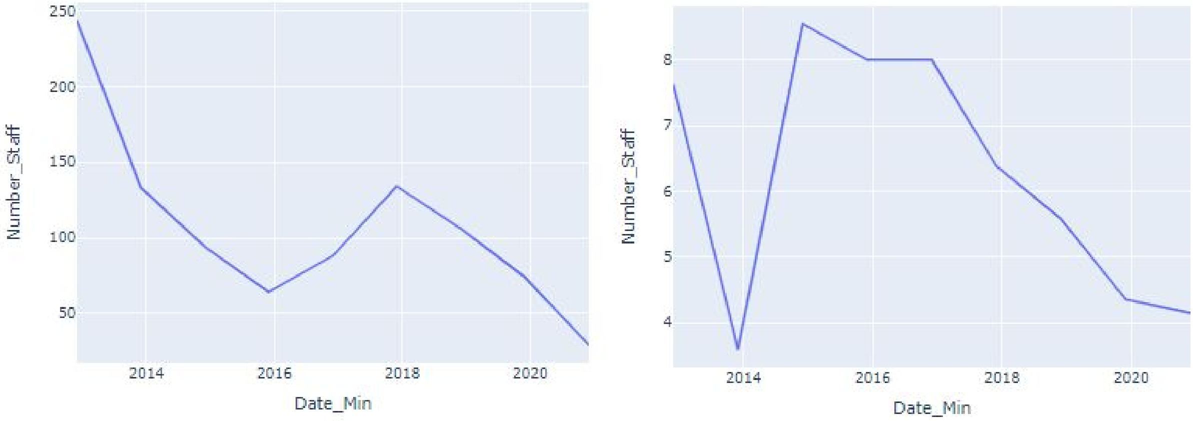


The maximum registered hours that the company has are in 2012. From there they begin to decrease until 2017, which has a slight increase from there, they continue to decrease slightly until 2020.

This may be due to the fact that with time and experience, the development of projects takes less time, but seeing the same information counting the projects per year we see that the peaks are for the number of projects.

There is a very high average use per year in 2014 and 2015, for all projects of that year

- Total an average number of employees per year



The number of employees used in the project is in line with the number of projects managed each year. However, when looking at the average number of employees per year the same question arises in 2014 and 2015, why so many officials for so few projects in that year.

Taking into account what was observed in the previous graphs, that the total and average number of hours per project has been decreasing, especially in 2020, allowed us to realize that we had to consider that some projects have not been closed, ie, have not fulfilled their life cycle, and should be excluded from the analysis, therefore included a filter to remove projects that have no record in the last 3 months, finding that 39 projects should be removed from the analysis.

On the other hand, it is known that some activities had to be excluded that are not specific to the project but are transversal, so we found 4 cases that do not have activities associated with a project but are transversal, so they should also be excluded. And finally, the exclusion was made for the modeling of some projects that presented an atypical behavior as we observed in the previous graphs, since they were very old projects and were affecting the adjustment of the models and left us a total of 91 projects for the model.

2.2 Feature Engineering

In this part, the construction of the base for modeling will be carried out, given that the original base described above is a log of the activities carried out in the projects, and what we want is to predict the hours executed in the projects, thus going from a base of 70.132 observations, to a base of 91 observations. Which is a big problem for the development of this project, because it becomes a great limitation since there would be very little data to fit a good model.

On the other hand, the construction of different variables that can be useful in the modeling is carried out and they are created with the information available in the original base, the variables constructed are:

- Minimum Date (Source for construction other variables)
- Maximum Date (Source for construction other variables)
- Months of project execution
- Days of project execution
- Number of staff in the project
- Number of staff per activity (20 activities)
- Ratio of hours per activity (20 activities)
- Number of versions
- Stage count

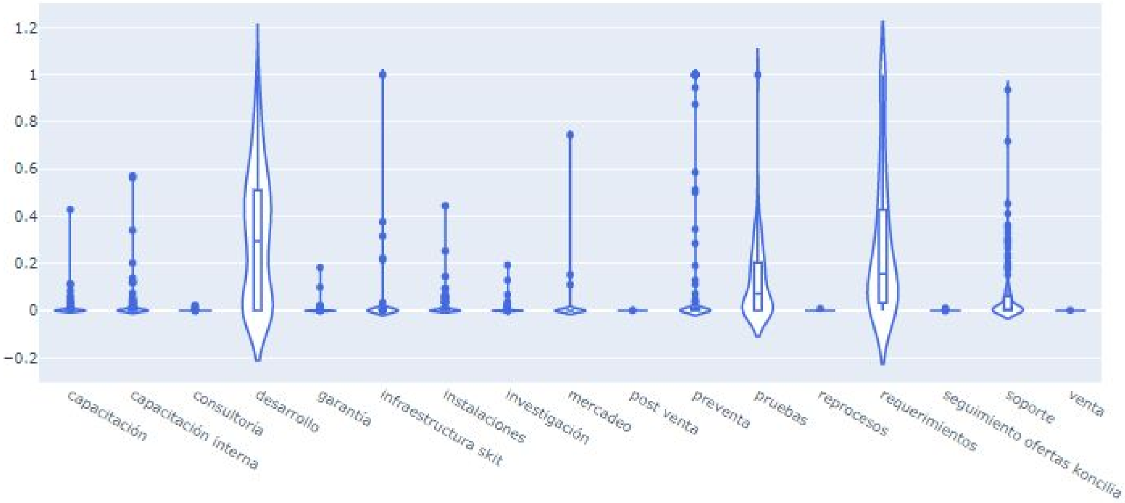
With the base built with the variables of interest, the descriptive and exploratory analysis is resumed, where it is expected to identify if there are outliers in the variables, therefore boxplots are made for some variables of interest.

2.2.1 Descriptive Analysis on the database built.

- Number of activities

The median number of activities per project is 6, we can also see there are projects with only one activity while there were in which 16 activities were necessary.

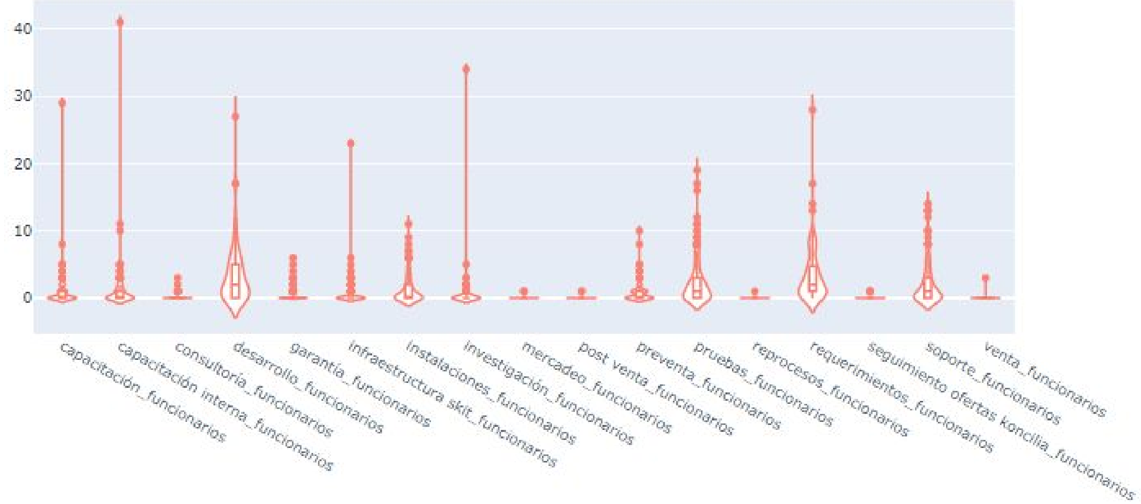
- Ratio of hours per activity



Previously, the construction of the proportion variables by activity was mentioned, and when analyzing these variables, we found that the activities that have a greater participation, with respect to the hours executed in the projects, are Development with 33% and Requirements with 16.5% and Tests with 8.1%. The other activities have very little participation within the projects.

It is important to mention that in order to calculate the proportion of hours executed per activity per project, the filter was made of some activities that are transversal and not specific to the project, such as Administrative and Financial Tasks, Human Resources, and Project Management. In addition, the support activity was consolidated, which was made up of the union of Corrective Support, Business Support, Night Support, and Weekend Support.

- Number of staff per activity



It is observed that the activities with greater representativeness in the projects are those that have a greater average value of number of personnel than the other activities.

2.2.2 Exploratory Analysis on the database built.

In order to complement the previous descriptive analysis, it was decided to see some graphs of dispersion of some variables, in order to identify trends among them.

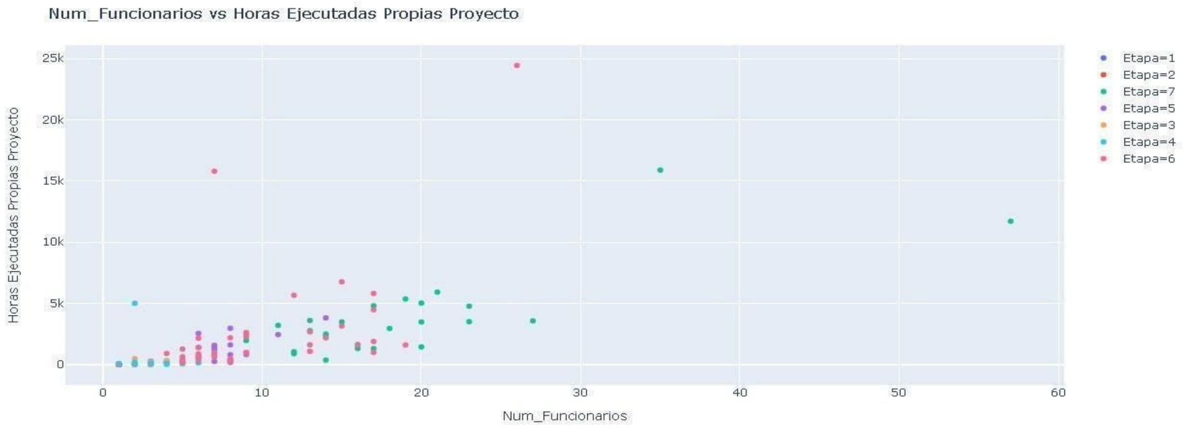


Figure 2. Number of employees Vs. Number of hours needed by the project

The previous graph shows the relationship between the number of staff associated with the project and the hours of execution. It was possible to identify some projects with atypical behavior because they have a large number of staff associated with the project and quite a few hours of execution.

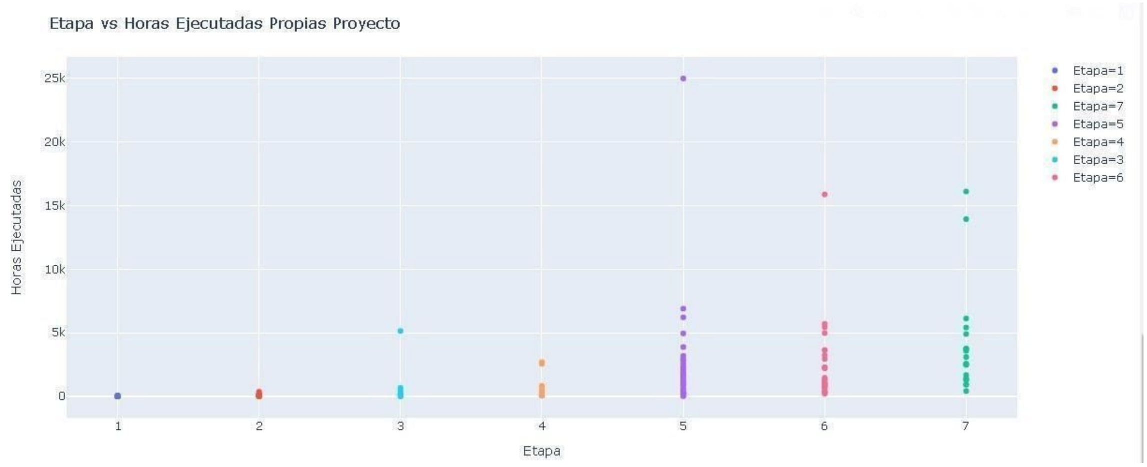


Figure 3. Number of hours per stage of the project

In this graph we find a relationship between the stages and the number of hours executed, that is, the greater the number of stages of the project, the greater the number of hours executed.

Finally, some atypical data were found, so the imputation of these by the median is made in order not to affect the distribution of the variables. In addition, a review of the data is considered after the mentioned transformations.

2.3 Univariate Analysis

This univariate analysis is done in order to review the data after the cleaning and to be able to filter those variables whose variability is 0%, since these will not contribute anything in the modeling.

After the review we found that 17 variables are constant at 0, and this is an effect of the elimination of some cases that could provide information on these variables but did not meet the other criteria for project completion or record of activities of the project.

- consulting
- guarantee
- skit infrastructure
- research
- marketing
- post sale
- reprocessing
- follow-up of offers koncilia
- sale
- staff consulting
- guarantee functions
- skit functionality infrastructure
- staff marketing
- post sale officials
- reprocessing functions
- follow up offers koncilia funcionarios
- staff sales

2.4 Bivariate Analysis

In this bivariate analysis, the relationship of each of the independent variables and the dependent variable is reviewed to find out if any variable has unstable behavior and should not be considered in the modeling. For this purpose, the Population Stability Index (PSI) is used, in which it was found that the value is quite small (inf) in all the variables, that is, that no variable exceeds the 0.15 threshold of instability, therefore, all the variables that were defined in the univariate analysis are maintained and the multivariate analysis is carried out.

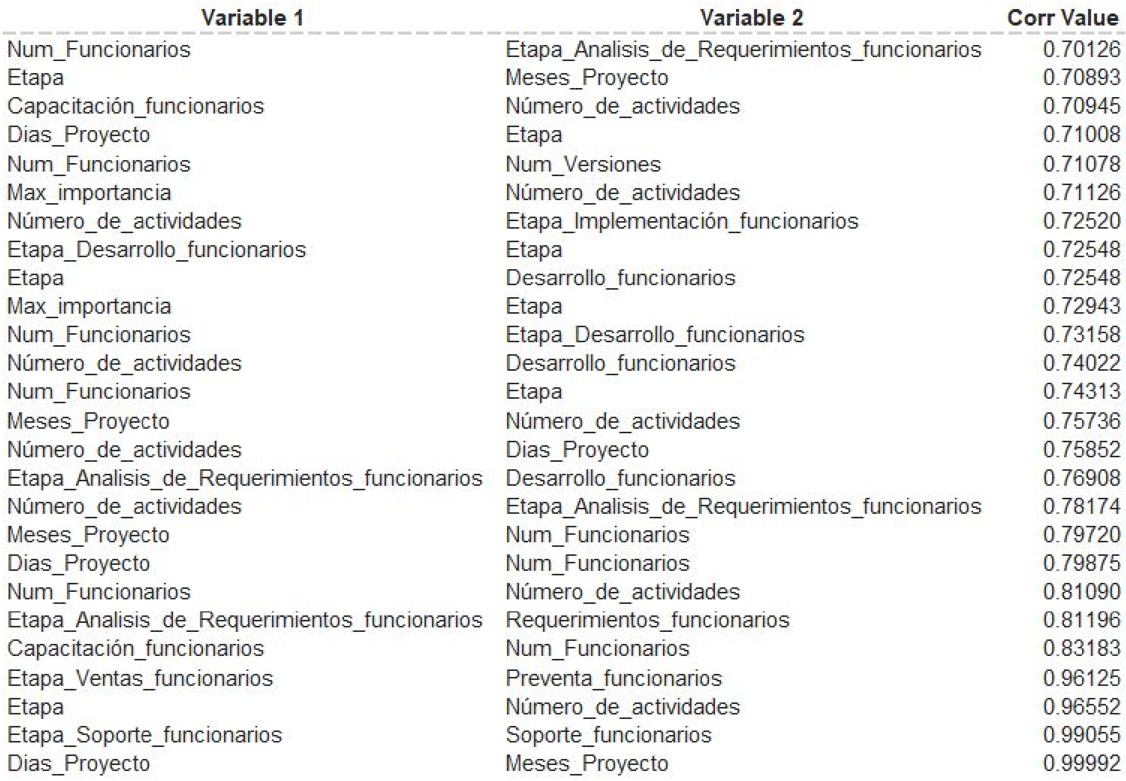
2.5 Multivariate Analysis

Multivariate analysis methods are used to explain the relationships between the different variables that may be associated with these data. The objective is to detect a structure on one hand and verify the data of the structures on the other. We use the correlation matrix, which measures the degree of linear relationship between each pair of items or variables.

- Correlation Matrix

When the proposal to build variables with the available information was made, it was known that some of them could be correlated, so the correlation graph was made in order to identify which are these variables and take them into account in the modeling to avoid multicollinearity problems.

A large number of variables were intentionally constructed from original variables, so it was perfectly understandable that there was a high correlation between several of them, for this reason the correlations were calculated and variables in which there was a correlation greater than 0.7 and less to -0.7 were eliminated. The pairs of variables that according to the previous criterion had a high correlation between them are shown below:



Variable 1	Variable 2	Corr Value
Num_Funcionarios	Etapa_Analisis_de_Requerimientos_funcionarios	0.70126
Etapa	Meses_Proyecto	0.70893
Capacitación_funcionarios	Número_de_actividades	0.70945
Dias_Proyecto	Etapa	0.71008
Num_Funcionarios	Num_Versiones	0.71078
Max_importancia	Número_de_actividades	0.71126
Número_de_actividades	Etapa_Implementación_funcionarios	0.72520
Etapa_Desarrollo_funcionarios	Etapa	0.72548
Etapa	Desarrollo_funcionarios	0.72548
Max_importancia	Etapa	0.72943
Num_Funcionarios	Etapa_Desarrollo_funcionarios	0.73158
Número_de_actividades	Desarrollo_funcionarios	0.74022
Num_Funcionarios	Etapa	0.74313
Meses_Proyecto	Número_de_actividades	0.75736
Número_de_actividades	Dias_Proyecto	0.75852
Etapa_Analisis_de_Requerimientos_funcionarios	Desarrollo_funcionarios	0.76908
Número_de_actividades	Etapa_Analisis_de_Requerimientos_funcionarios	0.78174
Meses_Proyecto	Num_Funcionarios	0.79720
Dias_Proyecto	Num_Funcionarios	0.79875
Num_Funcionarios	Número_de_actividades	0.81090
Etapa_Analisis_de_Requerimientos_funcionarios	Requerimientos_funcionarios	0.81196
Capacitación_funcionarios	Num_Funcionarios	0.83183
Etapa_Ventas_funcionarios	Preventa_funcionarios	0.96125
Etapa	Número_de_actividades	0.96552
Etapa_Soporte_funcionarios	Soporte_funcionarios	0.99055
Dias_Proyecto	Meses_Proyecto	0.99992

Table 2. Correlation between variables

It is important to clarify the criteria that was possessed when making the debugging, that was a function constructed where if they were two highly correlated variables, the one that had a stronger relationship with the variable response was kept and the other was eliminated.

Finally, after cleaning and debugging, we have 20 variables left, which are the ones we will start the modeling with.

3 Models and Results

3.1 Train and Test Data Base

The division of the modeling base between train and test is a technique that allows us to evaluate the performance of the models and is also used in classification and regression problems. With the train set we hope to develop the model and with the test set the objective is to estimate the performance of the model on new data, data not used to train the model. In this case it was decided that the data was going to be split in 75% for the train set and 25% for the test set.

3.2 Multiple Linear Regression

The first model proposed, is a multiple linear regression model, since our objective variable to model, is a continuous variable (hours that can take a project), and we seek to be able to make the estimate with variables that allow to give an approximate result and thus can be used by the entity at the beginning of any project.

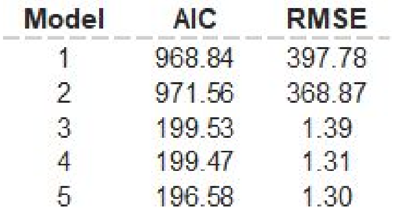
In this methodology it is known that it is quite sensitive to multicollinearity, so the previous step of the previous step of the debugging was necessary at the time of modeling. In addition, different model tests were made (they can be found in the project's notebook) as they are:

- Model with all variables (Model 1).
- Model removing the highest p-value variable. (Model 2)

However, it was found that several variables have a p-value greater than 0.05, so it is decided to perform the box cox test in which the value of 0.15 was obtained, indicating that the best transformation to perform is the logarithmic one. After performing the transformation, the following models were made:

- Model with transformed response variable. (Model 3)
- Model with transformed response variable and some variables with transformations as well. (Model 4)
- Model with transformed response variable, eliminating variables that due to regularization (lasso-ridge) were suggested not to be considered. (Model 5)

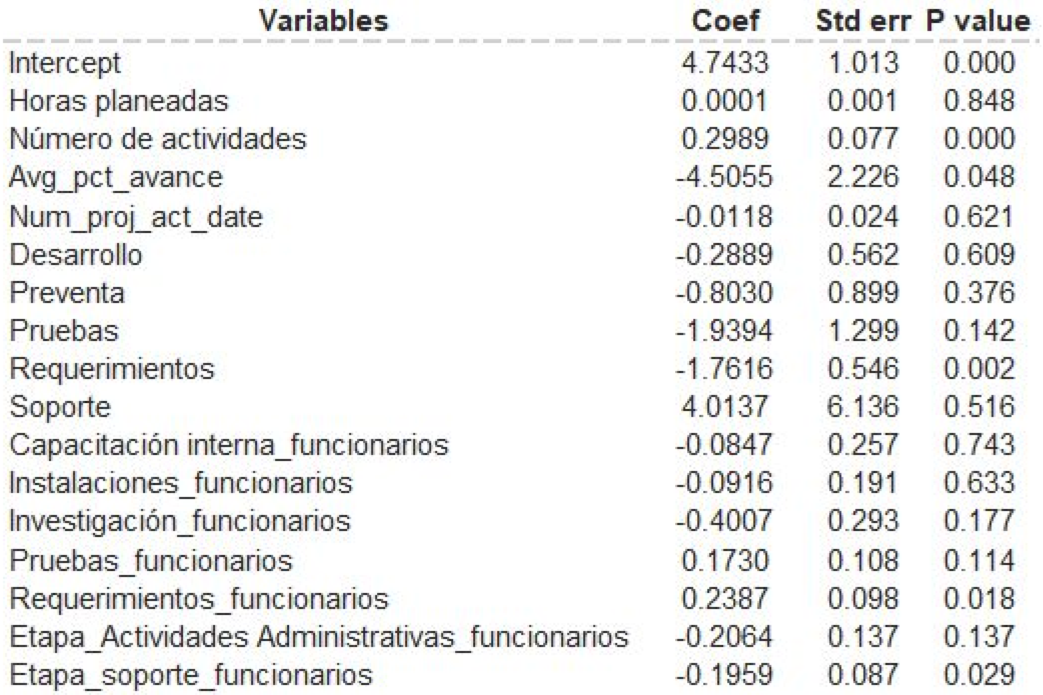
All the models were made the graphs of analysis of residuals, and finally the comparison between them is made by means of the AIC.



Model	AIC	RMSE
1	968.84	397.78
2	971.56	368.87
3	199.53	1.39
4	199.47	1.31
5	196.58	1.30

Table 3. AIC and RMSE from the Multiple Linear Regression models

As can be seen, Model 5 is the one with the lowest RMSE and there is a great improvement compared to Model 1. In addition, the square R of Model 1 is 0.59, and that of Model 5 is 0.74. However, reviewing the summary of the model, several variables are still presented that have a value greater than 0.05, so it is expected to be able to make more tests until the variables that remain in the model are significant.



Variables	Coef	Std err	P value
Intercept	4.7433	1.013	0.000
Horas planeadas	0.0001	0.001	0.848
Número de actividades	0.2989	0.077	0.000
Avg_pct_avance	-4.5055	2.226	0.048
Num_proj_act_date	-0.0118	0.024	0.621
Desarrollo	-0.2889	0.562	0.609
Preventa	-0.8030	0.899	0.376
Pruebas	-1.9394	1.299	0.142
Requerimientos	-1.7616	0.546	0.002
Soporte	4.0137	6.136	0.516
Capacitación interna_funcionarios	-0.0847	0.257	0.743
Instalaciones_funcionarios	-0.0916	0.191	0.633
Investigación_funcionarios	-0.4007	0.293	0.177
Pruebas_funcionarios	0.1730	0.108	0.114
Requerimientos_funcionarios	0.2387	0.098	0.018
Etapa_Actividades Administrativas_funcionarios	-0.2064	0.137	0.137
Etapa_soporte_funcionarios	-0.1959	0.087	0.029

Table 4. Output Model 5

3.2.1 Regularization

The regularization is a sort of "penalty" for having too many useless variables in a model. It aims to improve out-of-sample model predictions by reducing model complexity and preventing overfitting.

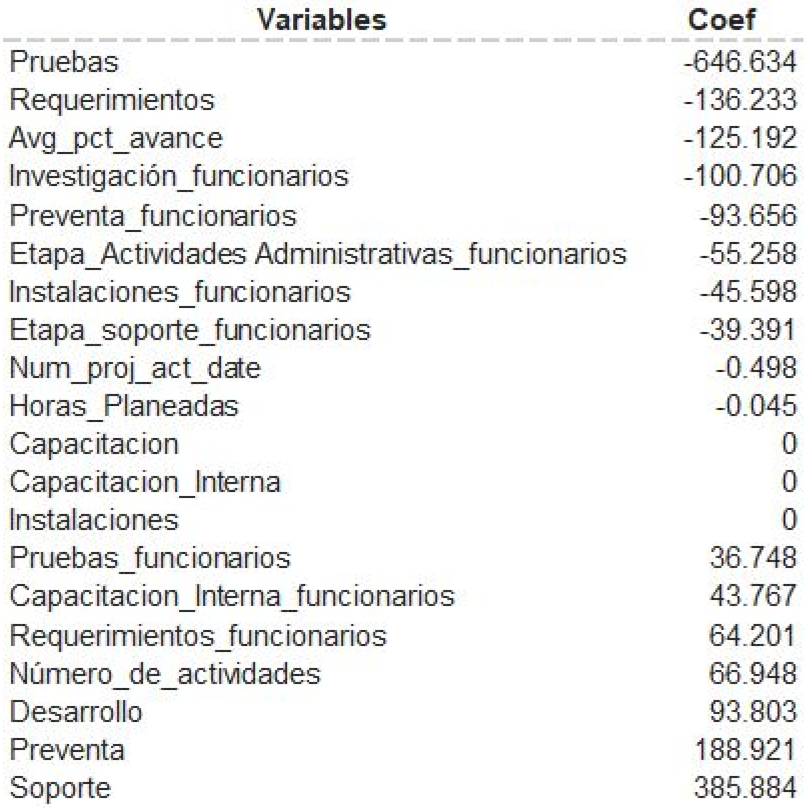
Due to the number of variables in the regression that are not significant, we are going to reduce the possibility of overfitting and eliminate some of the variables that might be less useful for the model.

For this we will use two regularization tools, the Lasso Regression and the Ridge Regression.

- **Lasso regression**

Also called **L1 regularization**, a LASSO (Least Absolute Shrinkage and Selection Operator) regression model shrinks the coefficients toward zero, meaning that if any of the coefficients is set to zero with this regularization, it is very likely that it is not significant for the predictions of the project hours.

After applying the Lasso Regression with a lambda of 0.5, the following coefficients were obtained for the variables:



Variables	Coef
Pruebas	-646.634
Requerimientos	-136.233
Avg_pct_avance	-125.192
Investigación_funcionarios	-100.706
Preventa_funcionarios	-93.656
Etapa_Actividades Administrativas_funcionarios	-55.258
Instalaciones_funcionarios	-45.598
Etapa_soporte_funcionarios	-39.391
Num_proj_act_date	-0.498
Horas_Planeadas	-0.045
Capacitacion	0
Capacitacion_Interna	0
Instalaciones	0
Pruebas_funcionarios	36.748
Capacitacion_Interna_funcionarios	43.767
Requerimientos_funcionarios	64.201
Número_de_actividades	66.948
Desarrollo	93.803
Preventa	188.921
Soporte	385.884

Table 5. Lasso Regression coefficients

The result of L1 regularization shows the coefficients of '**capacitacion'**, '**capacitacion interna'** and '**instalaciones'** to be zero. Let us remove these three variables going forward and refit a multiple linear regression model (This we saw in the multiple linear regression 4 and 5).

Lasso regression can be affected when data shows multicollinearity. This is why we also use L2 regularization, or Ridge Regression.

- **Ridge regression**

L2 regularization also reduces the coefficients towards zero, however, the effects of shrinkage are usually smaller than LASSO regression.

In our case, we did not find variables whose coefficients became zero, to a greater extent because before making the models we had already eliminated in the multivariate analysis those variables that were presenting high correlation indices.

3.3 Random Forest

A Random Forest model is made up of a set of individual decision trees, each trained with a slightly different sample of the training data generated by bootstrapping. The prediction of a new observation is obtained by adding the predictions of all the individual trees that make up the model.

We adjusted the model using "Horas Ejecutadas Propias Proyecto" as the response variable and all the other available variables as predictors. We use RandomForestRegressor from the sklearn.ensemble module to train the random forest model for regression problems. The most important parameters and hyperparameters are:

n_estimators: number of trees included in the model.

max_depth: maximum depth that trees can reach.

max_features: number of predictors considered a in each division.

min_samples_split: minimum number of observations that a node must have before it can be divided.

min_samples_leaf: minimum number of observations that each child node must have for the division to occur.

max_leaf_nodes: maximum number of terminal nodes that trees can have.

- Grid search

We use cross-validation based grid search to analyze various combinations of hyperparameters and get the most suitable values and find the next values for the hyperparameters

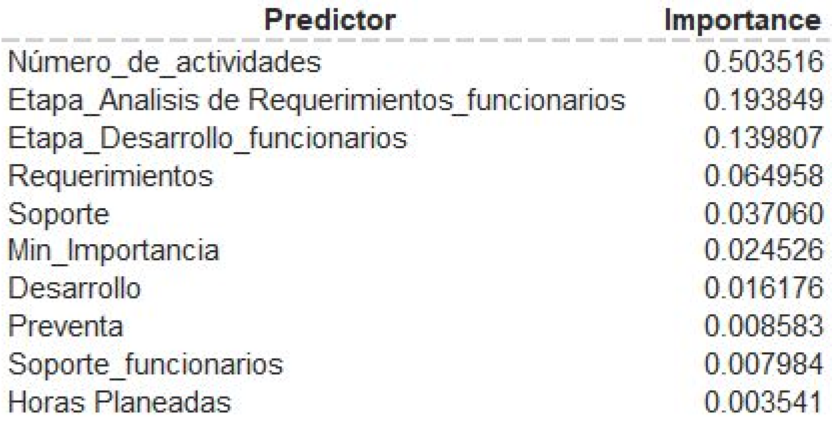
'max_depth': 3,

'max_features': 5,

'n_estimators': 70

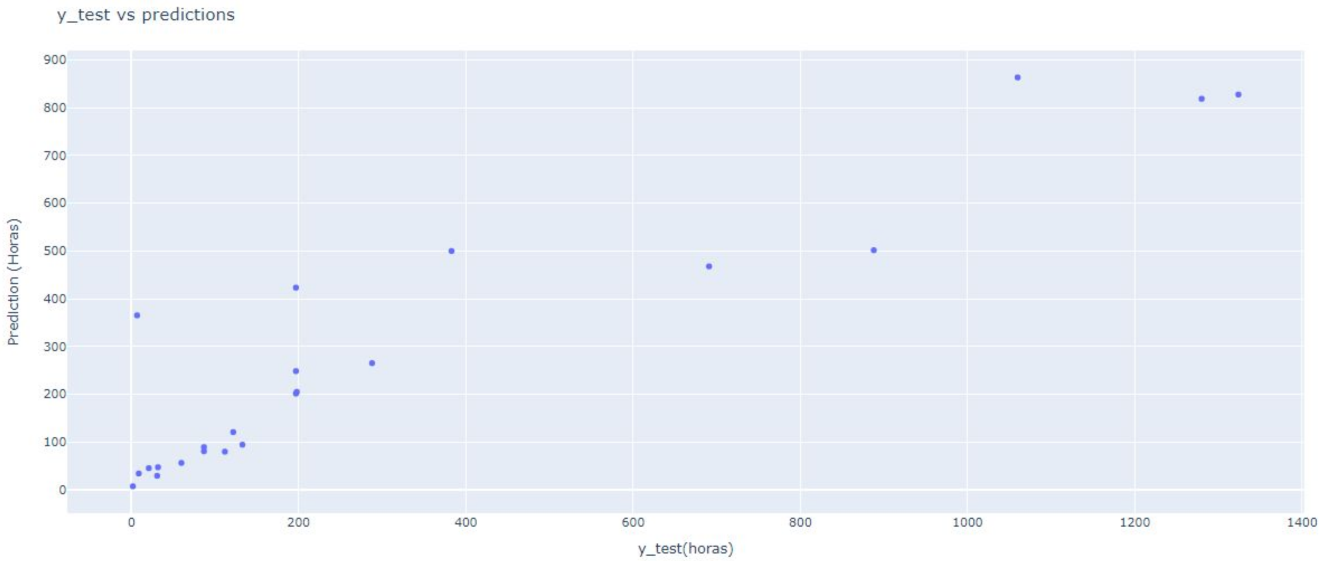
Once the best hyperparameters have been identified, the model is retrained, indicating the optimal values in its arguments.

After optimizing the hyperparameters, the rmse error of the model is 1.0165. Then we find the importance of the predictors, in this case two strategies were used: Importance by purity of nodes and Importance by permutation. where it is found that the most influential predictors are: Stage_Development_officials, tests_officials, Num_officials, internal training.



Predictor	Importance
Número_de_actividades	0.503516
Etapa_Analisis de Requerimientos_funcionarios	0.193849
Etapa_Desarrollo_funcionarios	0.139807
Requerimientos	0.064958
Soporte	0.037060
Min_Importancia	0.024526
Desarrollo	0.016176
Preventa	0.008583
Soporte_funcionarios	0.007984
Horas Planeadas	0.003541

Table 6. Importance of Random Forest predictors



In the previous graph, the y_test vs the predictions on the y-axis are shown on the x-axis, it is observed that there are some points very far from the line y = x.

3.4 XGBoost

XGBoost is one of the most used algorithms in the world mainly due to its good performance, additionally it can be used for regression problems, binary classification and multiclass classification and it does not require that the pre-processing of the data is very exhaustive.

This algorithm as many of the new machine learning methodologies fall into the trade-off between prediction and explainability, if the objective is to characterize the relationships between variables or perhaps do research on a specific topic is recommended to take much simpler statistical strategies such as regressions, instead, if the objective is to predict a phenomenon with the highest accuracy is recommended to use XGBoost. In the following graph you can see some of the advantages of this methodology.

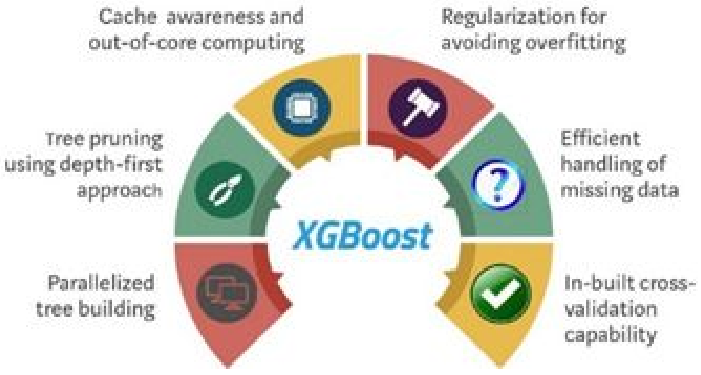
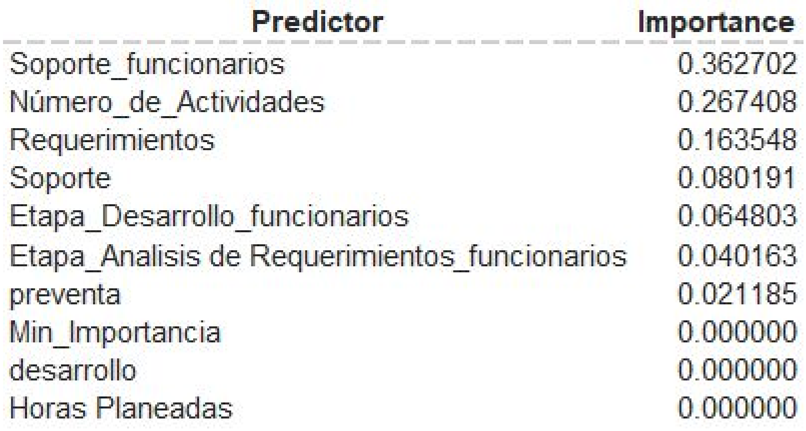

Cache awareness and out-of-core computing
Regularization for avoiding overfitting
Tree pruning using depth-first approach
Efficient handling of missing data
XGBoost
Parallelized tree building
In-built cross-validation capability

Figure 4. Advantages XGBoost

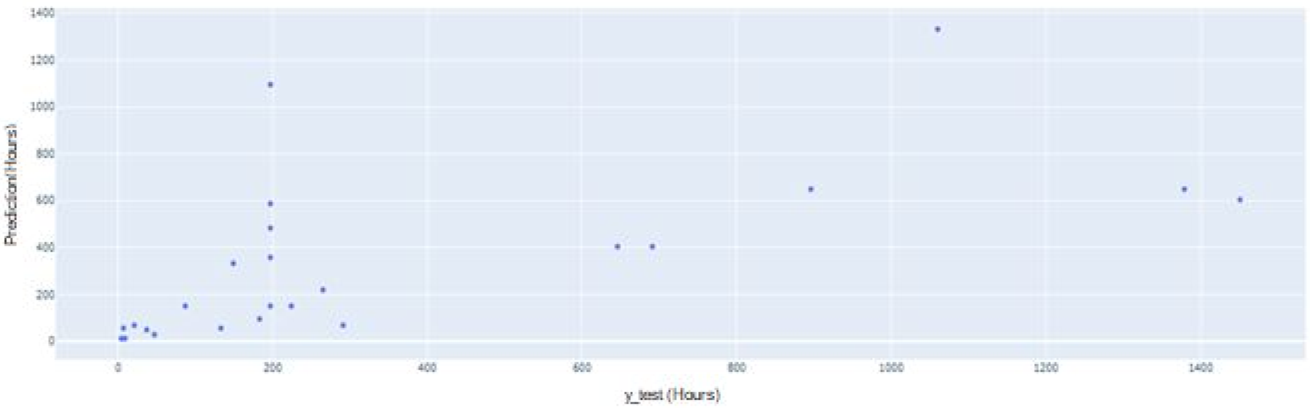
Taking into account that the objective of the project is to seek precision in the estimation of the hours that can take the development of a project, it was decided to carry out several tests as the previous methodologies, the best result being the one presented below, since in comparison to the first test of the same methodology, there was a positive impact on the errors, since it was reduced by 50% , when using the response variable with the box cox transformation.

It is also important to mention that as in Random Forest, Grid Search was used (described above), to find the best parameters for the modeling, and obtaining as a result an rmse error of 0.8948, which is the smallest error obtained from all the models tested. Bellow, you can see the importance of the variables:



Predictor	Importance
Soporte_funcionarios	0.362702
Número_de_Actividades	0.267408
Requerimientos	0.163548
Soporte	0.080191
Etapa_Desarrollo_funcionarios	0.064803
Etapa_Analisis de Requerimientos_funcionarios	0.040163
preventa	0.021185
Min_Importancia	0.000000
desarrollo	0.000000
Horas Planeadas	0.000000

Table 7. Importance of XGBoost predictors



Finally, in this last graph is observed a more linear behavior between the test data and the prediction obtained with the model, in addition to highlighting that the concentration of the projects is less than 400 hours, and in that quadrant the errors are lower, in cases where the error is a little higher, is in those who still remain on one side of the cloud of points with a number of hours higher than the average of current projects.

4 Application Overview

The main goal of the application is to allow the user to effectively predict the amount of resources (time and money) a project needs depending on its type. As well as showing statistics and the history of past projects developed by the company.

The application counts with the following components:

1. A page to input the data for the model to predict.
2. A view of the resources predicted to be used in the new project.
3. A report section where the user can view the information of the project's execution and make comparisons.

5 Data Engineering

5.1 Front-end

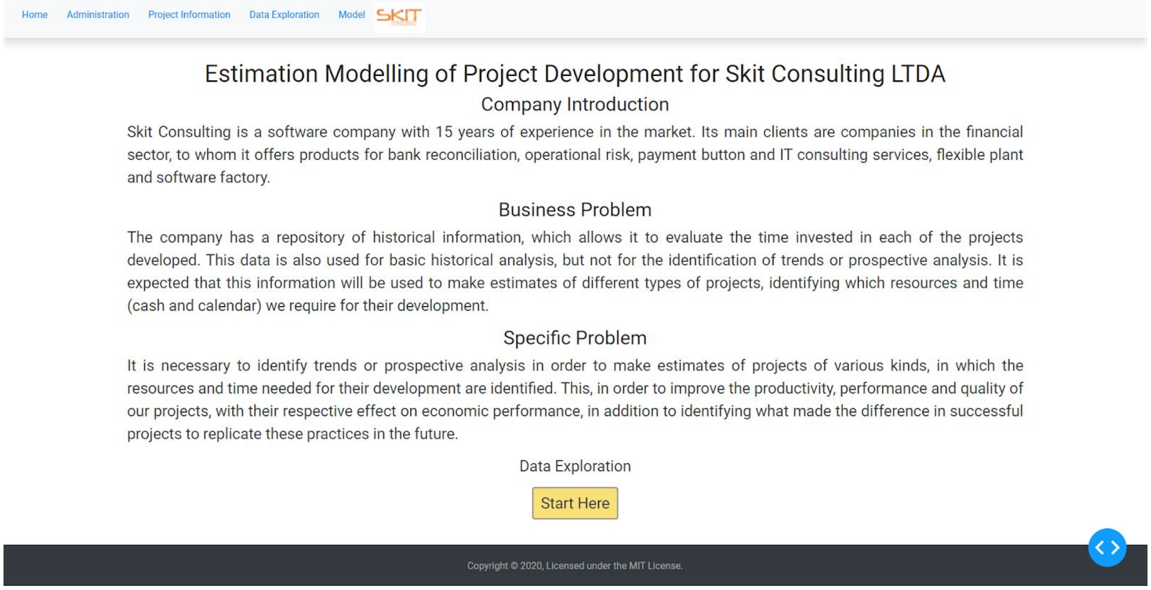
The link to access the application is: http://13.59.51.181:8080/

The application will have a Login that will be in charge of authenticating the user, this in order to avoid intruders. In addition, it will identify the type of user that entered ('Administrator', 'Leader', 'Developer').

In addition to this, the application has 5 modules, which contain information on the business problem, graphs on the projects, activities and stages of the same, and finally the model with which the estimated duration in hours of new projects is predicted.

Module "Home"

In this module, you will enter the relevant variables that will be the input for the prediction model, which calculates the estimated hours for the execution of a project.



Estimation Modelling of Project Development for Skit Consulting LTDA

Company Introduction

Skit Consulting is a software company with 15 years of experience in the market. Its main clients are companies in the financial sector, to whom it offers products for bank reconciliation, operational risk, payment button and IT consulting services, flexible plant and software factory.

Business Problem

The company has a repository of historical information, which allows it to evaluate the time invested in each of the projects developed. This data is also used for basic historical analysis, but not for the identification of trends or prospective analysis. It is expected that this information will be used to make estimates of different types of projects, identifying which resources and time (cash and calendar) we require for their development.

Specific Problem

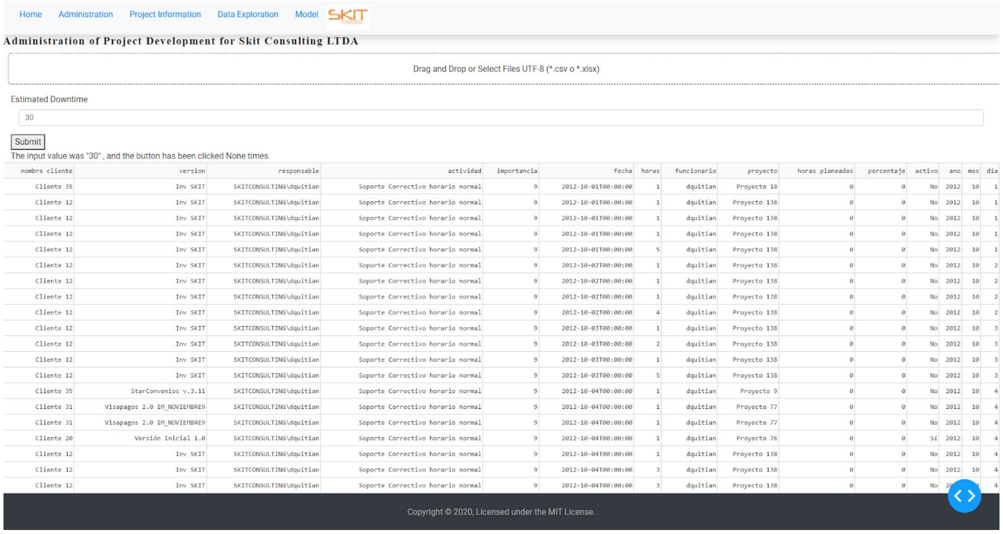
It is necessary to identify trends or prospective analysis in order to make estimates of projects of various kinds, in which the resources and time needed for their development are identified. This, in order to improve the productivity, performance and quality of our projects, with their respective effect on economic performance, in addition to identifying what made the difference in successful projects to replicate these practices in the future.

Data Exploration

Copyright © 2020, Licensed under the MIT License.

Figure 5. Home module

Module “Administration”



Administration of Project Development for Skit Consulting LTDA

Drag and Drop or Select Files UTF-8 (*.csv o *.xlsx)

Estimated Downtime

30

Submit

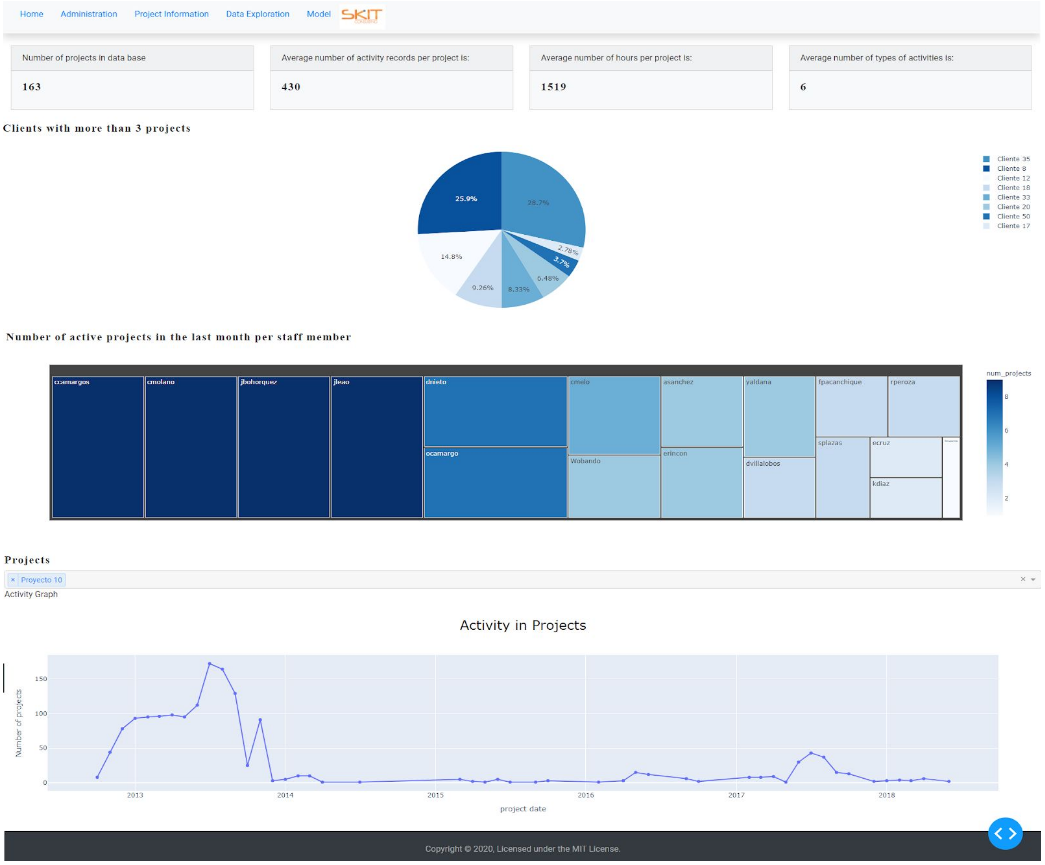
The input value was "30" , and the button has been clicked None times

nombre cliente	version	responsable	actividad	importancia	fecha	horas	funcionario	proyecto	horas planeadas	porcentaje	activo	ano	mes	dia
Cliente 35	Inv SKIT	SKITCONSULTING\dquitian	Soporte Correctivo horario normal	9	2012-10-01T00:00:00	1	dquitian	Proyecto 18	0	0	No	2012	10	1
Cliente 12	Inv SKIT	SKITCONSULTING\dquitian	Soporte Correctivo horario normal	9	2012-10-01T00:00:00	1	dquitian	Proyecto 138	0	0	No	2012	10	1
Cliente 12	Inv SKIT	SKITCONSULTING\dquitian	Soporte Correctivo horario normal	9	2012-10-01T00:00:00	1	dquitian	Proyecto 138	0	0	No	2012	10	1
Cliente 12	Inv SKIT	SKITCONSULTING\dquitian	Soporte Correctivo horario normal	9	2012-10-01T00:00:00	1	dquitian	Proyecto 138	0	0	No	2012	10	1
Cliente 12	Inv SKIT	SKITCONSULTING\dquitian	Soporte Correctivo horario normal	9	2012-10-01T00:00:00	5	dquitian	Proyecto 138	0	0	No	2012	10	1
Cliente 12	Inv SKIT	SKITCONSULTING\dquitian	Soporte Correctivo horario normal	9	2012-10-02T00:00:00	1	dquitian	Proyecto 138	0	0	No	2012	10	2
Cliente 12	Inv SKIT	SKITCONSULTING\dquitian	Soporte Correctivo horario normal	9	2012-10-02T00:00:00	1	dquitian	Proyecto 138	0	0	No	2012	10	2
Cliente 12	Inv SKIT	SKITCONSULTING\dquitian	Soporte Correctivo horario normal	9	2012-10-02T00:00:00	1	dquitian	Proyecto 138	0	0	No	2012	10	2
Cliente 12	Inv SKIT	SKITCONSULTING\dquitian	Soporte Correctivo horario normal	9	2012-10-02T00:00:00	4	dquitian	Proyecto 138	0	0	No	2012	10	2
Cliente 12	Inv SKIT	SKITCONSULTING\dquitian	Soporte Correctivo horario normal	9	2012-10-03T00:00:00	1	dquitian	Proyecto 138	0	0	No	2012	10	3
Cliente 12	Inv SKIT	SKITCONSULTING\dquitian	Soporte Correctivo horario normal	9	2012-10-03T00:00:00	2	dquitian	Proyecto 138	0	0	No	2012	10	3
Cliente 12	Inv SKIT	SKITCONSULTING\dquitian	Soporte Correctivo horario normal	9	2012-10-03T00:00:00	1	dquitian	Proyecto 138	0	0	No	2012	10	3
Cliente 12	Inv SKIT	SKITCONSULTING\dquitian	Soporte Correctivo horario normal	9	2012-10-03T00:00:00	5	dquitian	Proyecto 138	0	0	No	2012	10	3
Cliente 35	StarConvenios v.3.11	SKITCONSULTING\dquitian	Soporte Correctivo horario normal	9	2012-10-04T00:00:00	1	dquitian	Proyecto 9	0	0	No	2012	10	4
Cliente 31	Visapagos 2.0 IM_NOVIEMBRE9	SKITCONSULTING\dquitian	Soporte Correctivo horario normal	9	2012-10-04T00:00:00	1	dquitian	Proyecto 77	0	0	No	2012	10	4
Cliente 31	Visapagos 2.0 IM_NOVIEMBRE9	SKITCONSULTING\dquitian	Soporte Correctivo horario normal	9	2012-10-04T00:00:00	1	dquitian	Proyecto 77	0	0	No	2012	10	4
Cliente 20	Versión inicial 1.0	SKITCONSULTING\dquitian	Soporte Correctivo horario normal	9	2012-10-04T00:00:00	1	dquitian	Proyecto 76	0	0	Sí	2012	10	4
Cliente 12	Inv SKIT	SKITCONSULTING\dquitian	Soporte Correctivo horario normal	9	2012-10-04T00:00:00	1	dquitian	Proyecto 138	0	0	No	2012	10	4
Cliente 12	Inv SKIT	SKITCONSULTING\dquitian	Soporte Correctivo horario normal	9	2012-10-04T00:00:00	3	dquitian	Proyecto 138	0	0	No	2012	10	4
Cliente 12	Inv SKIT	SKITCONSULTING\dquitian	Soporte Correctivo horario normal	9	2012-10-04T00:00:00	3	dquitian	Proyecto 138	0	0	No	20	0	4

Figure 6. Administration module

Module “Project Information”

This module is in charge of visualizing and adjusting the model; allowing to parameterize certain variables in order to fit the model.



Number of projects in data base	Average number of activity records per project is:	Average number of hours per project is:	Average number of types of activities is:
163	430	1519	6

Clients with more than 3 projects

Number of active projects in the last month per staff member

Projects

× Proyecto 10

Activity Graph

Activity in Projects

Figure 7. Project information module

Module "Data Exploration"

This module is the information of the execution of the projects, making comparisons with the different variables that incur in the development of the same.

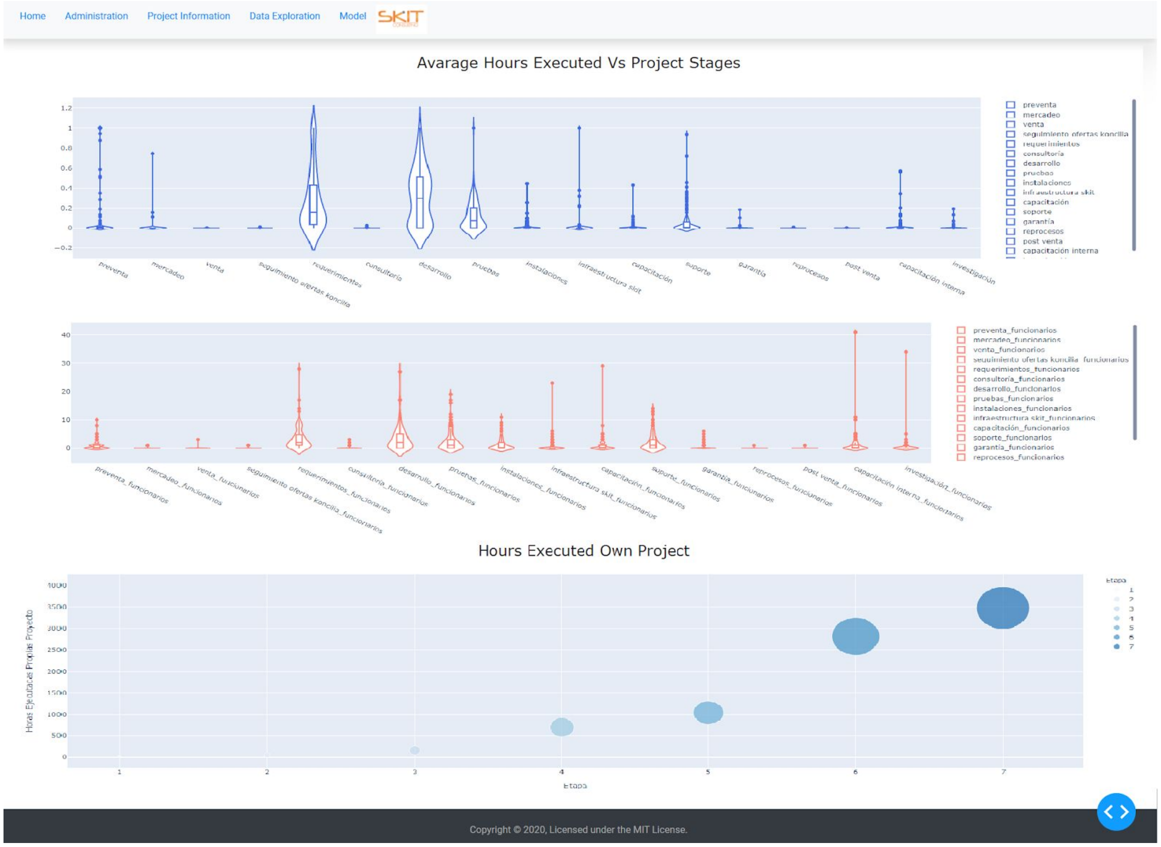
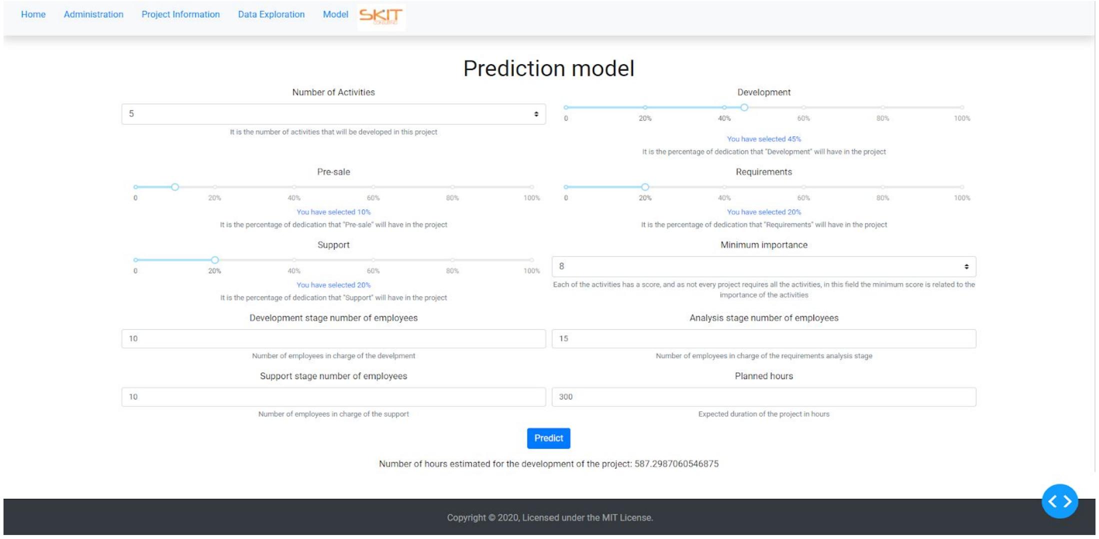


Figure 6. Data exploration module

Module "Model"



Prediction model

Number of Activities

5

It is the number of activities that will be developed in this project

Development

0 20% 40% 60% 80% 100%

You have selected 45%

It is the percentage of dedication that "Development" will have in the project

Pre-sale

0 20% 40% 60% 80% 100%

You have selected 10%

It is the percentage of dedication that "Pre-sale" will have in the project

Requirements

0 20% 40% 60% 80% 100%

You have selected 20%

It is the percentage of dedication that "Requirements" will have in the project

Support

0 20% 40% 60% 80% 100%

You have selected 20%

It is the percentage of dedication that "Support" will have in the project

Minimum importance

8

Each of the activities has a score, and as not every project requires all the activities, in this field the minimum score is related to the importance of the activities

Development stage number of employees

10

Number of employees in charge of the develpment

Analysis stage number of employees

15

Number of employees in charge of the requirements analysis stage

Support stage number of employees

10

Number of employees in charge of the support

Planned hours

300

Expected duration of the project in hours

Number of hours estimated for the development of the project: 587.2987060546875

Figure 7. Model module

5.2 Architecture

To achieve the goal written in this document, we needed to design an architecture suitable for the different components to interact. To do this, we use an amazon infrastructure with an Amazon RDS, Amazon EC2 and a web server. All this to be able to interact with the company's data.

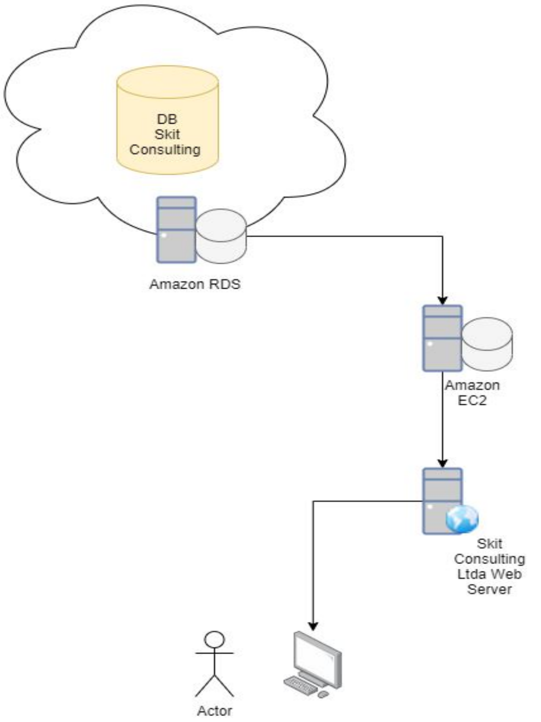


Figure 8. Flowchart data interaction

- **Amazon RDS:**

Here we display the different relational databases for the project. The project has one relational bases databases: TSKIT_FINAL.

The database administrator that will be implemented is POSTGRES, since it is open source software, and it is more friendly with the virtual machine.

This component will interact directly with the back-end of the application implemented in EC2.

Note: The database is already connected to the modeling.

- **Amazon EC2:**

This virtual machine is responsible for the execution of the main model. It receives requests from the web server of Skit Consulting Ltda and returns the filtered and processed data according to the request.

It also Interacts with Amazon RDS (original data sets) (resulting data sets). This will constitute the back-end of the architecture.

- **Skit Consulting Ltda Web:**

In this web server we will deploy the front-end of the application. Here, the company Skit Consulting Ltda, will be able to interact with the application. Through the API, requests are made to Amazon EC2 and then the data is filtered and processed.

They will be able to consult the model, and the status of the projects that they have loaded at the moment.

6 Conclusions

Several conclusions were found, the first are some about the analysis of the data, and to close we will present those of the project in general.

- (Analysis) In the exploratory analysis it is important to highlight that even though we have 163 projects, we must always take into account projects that meet the conditions of what we want to model, and this is the reason why projects that are still active were excluded, because they could generate noise in the modeling.

- (Analysis) It is also important to take into account the optimization of the company's processes, so the ideal is to take recent projects that have completed their full life cycle, as these will generate a greater accuracy in the model's fit.

- (Analysis) Finally, it is recommended to calibrate this model with more information, because as it is evident in the course of the document, the population of modeling was quite small, and that has an effect on the errors obtained in the different models tested.

- (Project) With the developed solution it is possible to have a greater precision on the estimated time of the development of a project, which will allow the employees of the company to have more real times of development depending on the complexity and thus to improve the productivity, the yield and the quality of the projects, distributing the time of work in a suitable way.

- (Project) On the other hand, by carrying out projects in the agreed times and with high quality, the companies that use SKIT will increase their confidence in the company, its products and services.

- (Project) Also by having satisfied user companies, they will make good comments on the project developed and the company, so it can lead to the increase of customers who were not there before.

7 References

Colombia Bring It On. (2020, October 28). *Colombiabringiton.co.* Retrieved from https://www.colombiabringiton.co/es/estas-son-las-fortalezas-de-la-industria-de-software-colombiana-la-mejor-opcion-en

Portafolio. (2018, May 21). *portafolio.co.* Retrieved from https://www.portafolio.co/negocios/industria-del-software-creceria-19-en-el-2018-517332

Velneo. (2019, January 16). *Velneo Blog.* Retrieved from https://velneo.es/el-desarrollo-de-software-de-gestion-en-colombia-en-2019/