ISSN 1870-4069

Predicting the Demand for Services at a Government Institute of Health in Mexico

Abraham Barroso, Noé Méndez,
Hiram Ponce

Universidad Panamericana,
Facultad de Ingeniería,
Mexico

{0264915, 0264134, hponce}@up.edu.mx

Abstract. Medical care is one of the issues that afflict the public health Mexican institutes' right holders on a daily basis, due to lack of personnel or waiting for care for long periods of time, so this paper seeks to support the government agency with the use of data and new technologies for better decision-making, through the use of machine learning and cloud computing technologies. For this reason, we have used linear regression models for our prediction tasks and compared their predictive power, in order for the institution to make a first approach and see the advantage of using new technologies and make more intensive use of them. Our results show that it is necessary to contemplate a greater number of data for more precise predictions, but it is something that the institution is not contemplated in the short time.

Keywords: Linear regression, medical services, helth institute, poor attention.

1 Introduction

Health system in Mexico suffers from limit capacity, lack of health professionals, lack of supplies and medicines, as well as some poorly implemented policies, which make trying to be attended a bad experience and in many cases discouraging due to these deficiencies.

Since the Mexican Government has different institutions that provide health services to the population, it is necessary to select one to start with, so the case study will focus on the Instituto de Seguridad y Servicios Sociales de los Trabajadores del Estado (ISSSTE), which is a parastatal entity of the Mexican Government that provides health and social security services to State workers and their families; providing services to 13.5 million beneficiaries in the country, which represents approximately 9% of the total population entitled to them in Mexico.

Public health institutes in Mexico have an important area of opportunity to make use of the large amounts of data that are generated every day in the different hospitals and services (from prescription refills, inventory management, hospital admissions and discharges, absenteeism, hiring, dismissals and retirements of medical and administrative personnel, etc.) in order to make better decisions and implement policies that lead to an improvement in the care and services provided to the population [4, 5].

Table 1. Datasets used in the experiment.

Description	Type
Federal Entity	string
Medical Unit Code	string
Name of the Medical Unit	string
Type of Medical Unit	string
Level of Care	string
Service/Specialty	string
Number of Consultations	integer
Service	string
Type of Consultation	string

The above through the use of technologies such as artificial intelligence and cloud computing services [1, 8], allowing to predict and anticipate the demand for these services to improve the care provided to beneficiaries, seek the correct allocation and recruitment of medical staff and implement measures and policies to mitigate the social discontent caused by poor or no care and have gained relevance in recent years.

Therefore, the interest and objective, as the first scope of this work, is to focus on the prediction of the demand for services and patient care, because it has become a controversial issue in the health institutions of the Mexican Government, especially in those that have to do with care specialties or involving specialized and complex procedures, with the need for the use of devices and high medical technology and the participation of a multidisciplinary team in some cases. This work aims to apply machine learning models for predicting the number of services required in the medical units of ISSSTE.

The aforementioned is sought to be achieved through the use of information obtained from the year 2022 from the institute, the use of machine learning models and an architecture based on cloud computing, which will allow us, at some point, to replicate it to other health institutions. The rest of the paper is as follows. Section 2 describes the dataset. Section 3 presents the proposal of the work. Section 4 includes the experimental design and Section 5 shows the results and discussion. Finally, Section 6 concludes the work.

2 Description of the Dataset

The dataset contains [3] information of number of consult for service an medical unit, as shown in Table 1 and Table 2. We have 35 federal entities (includes subdivisions by region for Mexico City), 10 types of medical units, 112 medical units and 3 levels of care. It is very important that we study the distribution of the response variable, since, at the end of the day, this is what we are interested in predicting. Its distribution is visualized in Fig. 1, and we will apply a logarithmic and square root transformation to see its distribution from different perspectives.

Table 2. Catalog of types of medical units.

ID	Description	Level	Number of Medical Units
CMCT	Medical Office in the Workplace	1	47
CAF	Family Care Office	1	391
UMF	Family Medicine Unit	1	402
CMF	Family Medicine Clinic	1	91
CMFEQ	Family Medicine Clinic with Specialty and Operating Room	1,2	16
CE	Specialty Clinic	1,2	6
CEQ	Specialty Clinic with Operating Room	2	5
CH	Hospital Clinic	2	72
HG	General Hospital	2	26
HR/HAE	Regional Hospital / High Specialty Hospital	2,3	14
CMN	National Medical Center	3	1

Table 3. Datasets statistics for variable "Consultas".

Variable	count	mean	Std	min	25%	50%	75%	max
Consultations	9408.00	1598.31	4719.47	1.00	182.00	551.00	1364.00	96506.00

With the above and with the support of Python we can evaluate which distribution fits our data, because some of the machine learning models need a specific distribution and in our case chi-square is the one that best fits our data. In Table 3, we show the statistical data for our numerical predictor variable. We generated a chart to show the distribution of the number of medical consultations by "Federal Entity", as depicted in Fig. 2. In the Table 4 we show the statistical data for our categorical variables.

3 Description of the Proposal

We adopt the general workflow of machine learning for tackling the problem, as summarized in Fig. 3. The details are described below.

Dataset Selection. As a first step, only the values were taken from the information provided by ISSSTE.

Cleaning and Adjustment. Subsequently, the headings were constructed as follows: We noticed that each specialty has 4 columns: First Time, Subsequent, Visit, Total. When we had this situation where what we needed was to distinguish each column in each specialty, a concatenation was made between that column and the specialty. For example, the column "First Time" appears in both Allergology and Anesthesiology, so they were as follows: Allergology_First_Time and Anesthesiology_First_Time.

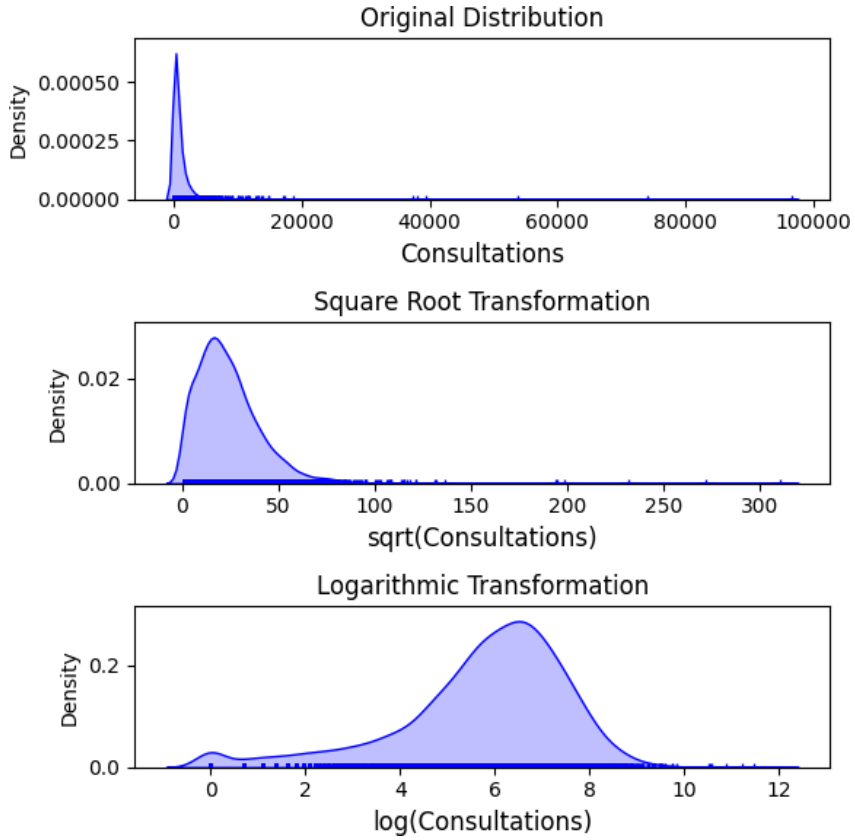

Fig. 1. Distributions.

Once this process was completed, a filter was implemented to eliminate the totals that were presented between the same information by Delegation. In this way, we obtained an optimal dataset to work with and begin to perform exploratory analysis.

Normalization Process. It is a method in which the values in a numeric column change so that the data set has a common scale, without distorting the differences in the ranges of values or losing information, and this may be a necessary activity for use in certain algorithms. For our data normalization we will use StandardScaler, a class that standardizes the data by removing the mean and scaling the data so that its variance is equal to 1.

Null Values. To handle null values we will use Sklearn's SimpleImputer function, which allows us to substitute null values for other values according to various strategies available in it.

Categorical Variables. For our work and to code our variables we will use OneHotScaler, whose strategy is based on creating a binary column (with values 0 or 1) for each single categorical value and places a 1 in the corresponding column where a value is present, leaving the rest of the columns with value 0.

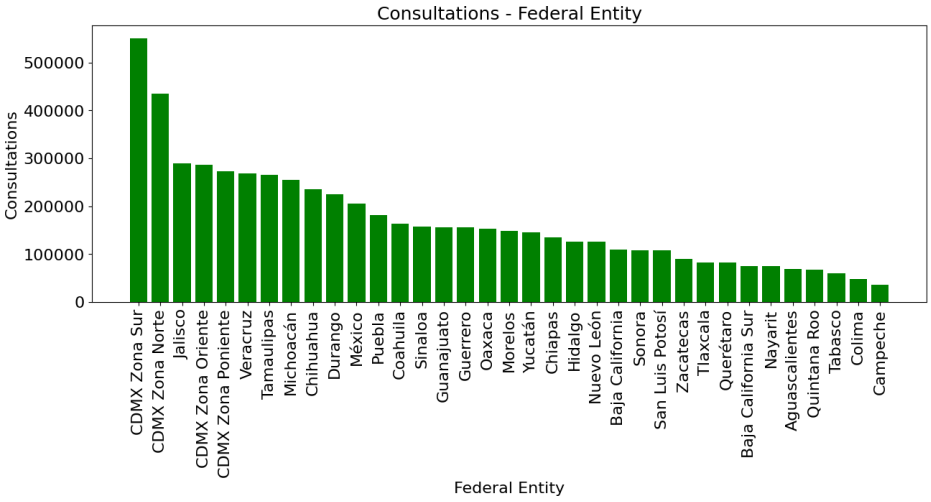

Fig. 2. Number of medical consultations.

Table 4. Datasets statistics for categorical variables.

Variable	Count	Unique	Top	Freq
Federal Entity	9408.00	35	Veracruz	21600.00
Key	9408.00	1071.00	001-204-00	270.00
Name	94080.00	1021.00	Tuxpan	1080.00
Type	9408.00	11.00	UMF	108540.00
Level	9408.00	3.00	1st	251370.00
Service	9408.00	90.00	Continuous.Admission.Adults	3213.00
Type of Consultation	9408.00	3.00	First.time	96390.00

Solution Arquitecture. As you can see in Fig. 4, the architecture includes components in the Microsoft Azure cloud, with the idea of a much more agile deployment, easy to scale and start with small scopes. Given that the information with which the predictions will be generated is non-sensitive, it is possible to take advantage of having an architecture in the cloud with security.

Models. We select four well-known machine learning models for the study:

– **Multiple linear regression [6]:** allows us to generate a linear model in which the value of our dependent variable (also known as response (Y)) and which is determined from a set of variables, known as independent or predictors (X1, X2, X3...). This is a variation or extension of simple linear regression.

In our case we will use it to predict the value of the dependent variable, but it is not the only thing that can be done with this model as it can allow us to see how the response variable is influenced by the independent variables.

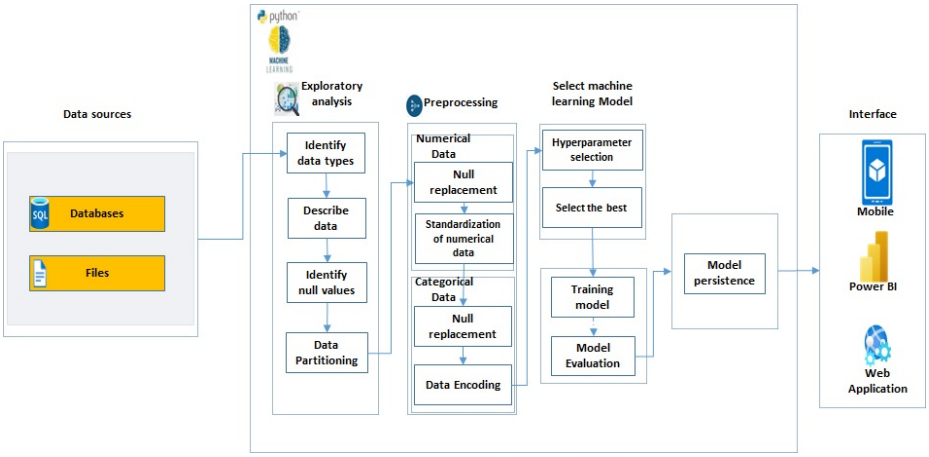

Fig. 3. General flow of our machine learning process.

The multiple linear model has the following equation (1):

$$Y_i = (\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_n X_{ni}) + e_i. \tag{1}$$

– **Lasso regression [6]:** (Least Absolute Shrinkage and Selection Operator) is a linear model that penalizes the coefficient vector by adding its L1 norm to the cost function:

$$\text{Minimize } \theta : \sum_{i=1}^{N} [y_i - f(x_i, \theta)]^2 + \lambda \sum_{j=1}^{M} |\theta_j|. \tag{2}$$

This model has the characteristic of generating "sparse coefficients": which are vectors of coefficients in which most of them take the value zero. So the model considers ignoring some of the predictive features, which can be considered a type of automatic feature selection.

The model by performing feature exclusion seeks to generate a model that is simpler to interpret and exposes the most important features of our data set. If there is a correlation gap between the predictive features, the Lasso model will tend to choose one of them at random.

– **Ridge regression [2]:** Also known as contracted regression or Tikhonov regularization, aims to regularize the resulting model and imposes penalties on the size of the coefficients of the linear relationship between the predicted characteristics and the target variable. The coefficients that are calculated in the model seek to minimize the sum of the squares of the residuals by penalizing them by adding the square of the L2 norm of the vector formed by the coefficients:

$$\text{Minimize } \theta : \sum_{i=1}^{N} [y_i - f(x_i, \theta)]^2 + \lambda \sum_{j=1}^{M} \theta_j^2. \tag{3}$$

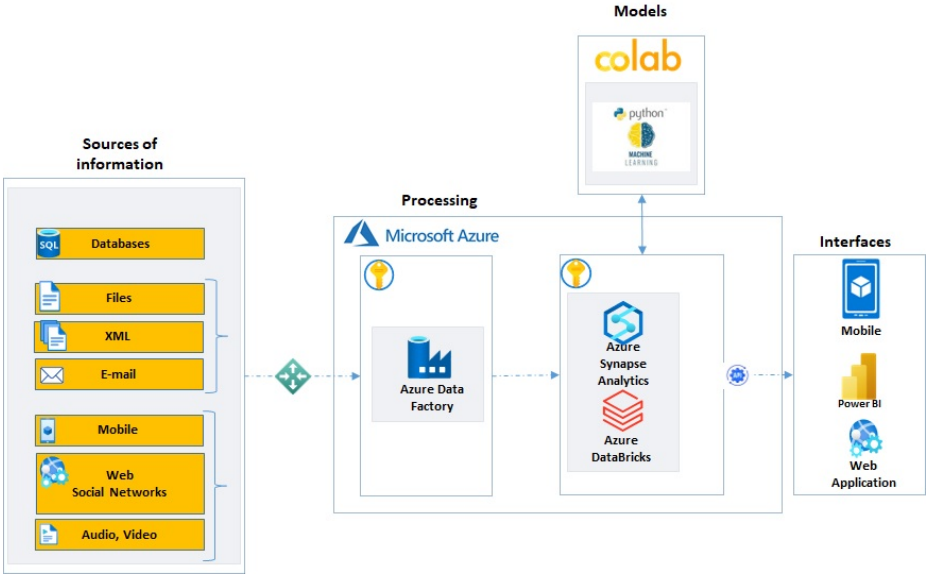

Fig. 4. Solution architecture.

In the formula of the model we have that λ is a parameter that controls the degree of penalization: the higher the value of λ , the lower the coefficients will be resulting more robust to collinearity.

– **Support Vector Regression (SVR) [7]:** this model is born from a variant of the Support Vector Machine (SVM) analysis model which is to perform classification tasks, however, the SVR model makes some minor changes in its definition.

For its use in regression cases, a tolerance margin (ϵ) is established near the vector and its purpose is to try to minimize the error, taking into account that part of that error is tolerated.

Equation for Linear SVR:

$$y = \sum_{i=1}^{N}(\alpha_i - \alpha_i^*)(x_i,\, x) + b. \tag{4}$$

Equation for Non-linear SVR:

$$y = \sum_{i=1}^{N}(\alpha_i - \alpha_i^*)(\varphi(x_i),\, \varphi(x)) + b, \tag{5}$$

$$y = \sum_{i=1}^{N}(\alpha_i - \alpha_i^*)K(x_i,\, x) + b. \tag{6}$$

For hyper-parameter optimization we use grid search which performs an exhaustive search by evaluating all parameter combinations.

Table 5. Results of evaluation metrics.

ID	Model	Model	Model	SVR
	Linnear Regression	**Regression Lasso**	**Regression Ridge**	**SVR**
MSE	3.620 e+42	3693492.38	3563258.70	3907206.78
MAE	59.67 e+48	798.64	819.77	1976.66
$\mathbf{R}^2$	-9.53 e+37	0.02895	0.06319	-0.02723

This strategy has the disadvantage of requiring high consumption of computational resources when the number of data becomes too large, as well as the evaluation of regions that may be of little interest before evaluating more combinations, it uses cross-validation techniques for its operation.

4 Experimentation

For this work it was decided to use cross-validation, which consists of randomly dividing the observations obtained into k groups of equal size. One of the k groups is used as the validation set, while the remaining $k-1$ groups are used to train the model. The mean squared error (MSE) is calculated on the $k-1$ groups excluded from the model, this validation process is repeated k times because each group is used as a validator. So in the end we obtain k estimates of the mean squared error and calculate the overall estimate by averaging the k values of our linear regression using (7):

$$CV_k = \frac{1}{k}\sum_{i=1}^{k} \mathrm{MSE}_i \,. \tag{7}$$

For our experiments we use 80% of the data for training and 20% for validation of our models. As evaluation metrics, we consider the following ones (8)-(10):

Mean Absolute Error (MAE):

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} \mid y_i - x_i \mid}{n}. \tag{8}$$

Mean Squared Error (MSE):

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2. \tag{9}$$

R Squared Value (R^2):

$$R^2 = \frac{\mathrm{SSR}}{\mathrm{SST}} = 1 - \frac{\mathrm{SSR}}{\mathrm{SST}} = 1 - \frac{\sum_{i}^{n}(y_i - \hat{y_i})^2}{\sum_{i}^{n}(y_i - \bar{y_i})^2}. \tag{10}$$

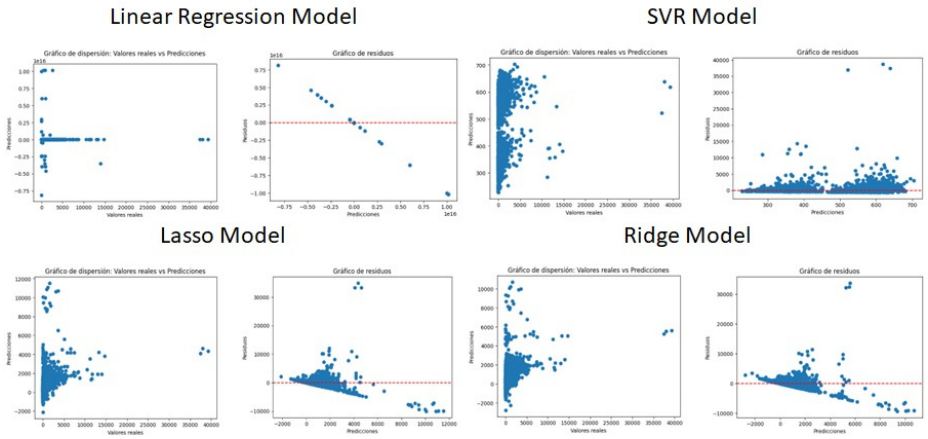

Fig. 5. Scatter plots and residual plots of the models used.

5 Results and Discussion

As part of our results we reviewed the coefficient of determination, also known as R^2 (R-squared), as a metric used to assess the quality of our regression models. This provides us with a measure of how well the model predictions fit the actual values of the target variable. The comparison of our models with respect to the value of the coefficient of determination ranges from 0 to 1, as shown below in Table 5.

As we can see our Lasso regression and Ridge regression models are close to 0 in our R^2 evaluation and indicate that the model is not able to explain the variability in the data and that the predictions are similar to simply using the mean value of the target variable. This suggests that the model is not adequate to represent the relationship between the predictor variables and the target variable.

In the case of SVR the R^2 value is negative, so our model performs very poorly in making predictions. Our model behaves in extremely poor ways and the errors are large compared to the variability of the data. Analyzing the Mean Squared Error (MSE) used to assess the quality of a regression model. Consider that the MSE is a measure of the variance or dispersion of the errors of our model and for our model the lowest MSE we have in the Ridge regression model is the one with the best fit to the data, since the errors are smaller and closer to zero.

On the other hand, the SVR and linear regression models have a higher MSE indicating that the models have a worse fit, this because the errors are larger and farther from zero. Finally we analyze the MAE which, as mentioned, is calculated by taking the absolute difference between the values predicted by the model and the actual values, and then calculating the average of these differences.

When interpreting the MAE for each of our models, we must take into consideration that this value is representing the average magnitude of the model errors under the same scale as our original data (number of queries) and as a difference from the MSE, the MAE does not square the errors, which allows us to keep the metric unaffected by outliers or large errors.

In our models, the Lasso Regression and Ridge model values have the lowest MAE metrics and indicate that the models have a better fit to the data, since the errors are smaller on average, while the linear regression and SVR model have high values that indicate that our models have a worse fit, due to the fact that the errors are larger on average. Lastly, Fig. 5 shows the scatter and residual plots of the models. We notice that the SVR model for the part of the residuals is behaving with homoscedasticity, while the rest of the models are showing heteroscedasticity.

6 Conclusions

This work aimed to study the performance of four machine learning models for predicting the number of services required in the medical units of ISSSTE. We adopted the general workflow of machine learning to approach our goal. By performing the evaluation of our models using the metrics proposed in our project (MAE,R^2, MSE), we have concluded that the results obtained by these currently have discrepancies to perform the prediction of some of the data and the metrics indicates a very low value in their prediction process. Given the above, it is necessary to:

– Evaluate and increase the number of data to improve the developed models as a first action, to subsequently evaluate the use of more features, provided that these do not introduce noise or that they are not significant and may affect their performance.

– To evaluate in a more exhaustive way the hyper-parameters that each of the models use and that were carried out in the development of the present project.

– Evaluate the use of other models, currently the machine learning area has a wide range of models (some with improvements of the base versions) that can support a better evaluation of data prediction and support better decision making.

When experimenting using grid search techniques for optimal hyper-parameters that maximize the performance of our models, we found that it can be a time-consuming and costly process in terms of computational processing, so as a continuation of this project, additional hyper-parameter search methods (e.g., Bayesian sampling) should be evaluated.

References

1. Cohen, T. A., Patel, V. L., Shortliffe, E. H.: Intelligent systems in medicine and health: The role of AI. Springer International Publishing (2022) doi: 10.1007/978-3-031-09108-7
2. Ehsanes Saleh, A. K. M., Arashi, M., Golam-Kibria, B. M.: Theory of ridge regression estimation with applications. John Wiley and Sons, Inc (2019) doi: 10.1002/9781118644478
3. Instituto de Seguridad y Servicios Sociales de los Trabajadores del Estado: Estadística anuarios (2022) www.issste.gob.mx/datosabiertos/anuarios/anuarios2022.html
4. King, Z., Farrington, J., Utley, M., Kung, E., Elkhodair, S., Harris, S., Sekula, R., Gillham, J., Li, K., Crowe, S.: Machine learning for real-time aggregated prediction of hospital admission for emergency patients. npj Digital Medicine, vol. 5, no. 1, pp. 104 (2022) doi: 10.1038/s41746-022-00649-y

5. Liu, Y., Qin, S.: An interpretable machine learning approach for predicting hospital length of stay and readmission. In: Proceedings of the 17th International Conference on Advanced Data Mining and Applications, vol. 13087, pp. 73–85 (2022) doi: 10.1007/978-3-030-95405-5_6
6. Montgomery, D. C., Peck, E. A., Vining, G. G.: Introduction to linear regression analysis. Wiley (2021)
7. Ozer, M. E., Sarica, P. O., Arga, K. Y.: New machine learning applications to accelerate personalized medicine in breast cancer: Rise of the support vector machines. OMICS: A Journal of Integrative Biology, vol. 24, no. 5, pp. 241–246 (2020) doi: 10.1089/omi.2020.0001
8. van-Houten, H.: El poder de la predicción: Cómo la IA puede ayudar a los hospitales a prever y gestionar el flujo de pacientes. Philips: Centro de noticias España (2021) www.philips.es/a-w/about/news/archive/standard/news/blogs/2021/20210906-the-power-of-prediction-how-ai-can-help-hospitals-forecast-and-manage-patient-flow.html