Journal of Systemics, Cybernetics and Informatics (2023) 21(1), 1-8
https://doi.org/10.54808/JSCI.21.01.1

Data science application for creation of maternal morbidity and mortality predictive software

Rúsbel DOMÍNGUEZ-DOMÍNGUEZ
Faculty of Engineering and Technology, Department of Informática, Montemorelos University,
Montemorelos, Nuevo León, 67510, México.

Germán-H. ALFÉREZ
Faculty of Engineering and Technology, Department of Informática, Montemorelos University,
Montemorelos, Nuevo León, 67510, México.

Verenice GONZÁLEZ-MEJIA
Medical School, Department of Research Support in Health Sciences, Montemorelos University,
Montemorelos, Nuevo León, 67510, México.

Norbet DONÍAS
Medical School, Department of Research Support in Health Sciences, Montemorelos University,
Montemorelos, Nuevo León, 67510, México.

ABSTRACT

In Mexico, the estimated Maternal Mortality Ratio is 34.6 deaths per 100,000 estimated births. Consequently, healthcare facilities and services have given precedence to prenatal care, childbirth services, and postpartum care.

In Mexico, the Ministry of Health maintains an open database concerning maternal deaths, encompassing 58 variables. Among these variables is the CIE (International Statistical Classification of Diseases and Related Health Problems), which covers a total of 248 diseases linked to maternal deaths.

Currently, there is no software that classifies women undergoing pregnancy check-ups (according to their socio-clinical risk of mortality), using variables selected with data science.

This project is rooted in the methodology advanced by International Business Machines (IBM) for the implementation of data science.

The software's utilized model was constructed through the Naïve Bayes supervised learning algorithm, yielding an accuracy of 0.7236. The overall precision stood at 0.75, with an overall recall of 0.74, and an overall F1-score of 0.71. For the eclampsia during labor class, precision reached 0.71, recall was 0.94, and the F1- score attained 0.81. As for secondary or late postpartum hemorrhage, precision scored 0.81, recall measured 0.43, and the F1-score was 0.56.

Keywords: Data science, software creation, maternal mortality.

1. INTRODUCTION

The maternal mortality rate is a basic health indicator, reflecting the economic, social, educational and health development of a country [1]. Maternal mortality is characterized as the passing of a woman during pregnancy or within 42 days after the conclusion of a pregnancy [2]. This definition includes maternal deaths that have a direct or indirect cause in pregnancy.

In Mexico, the main causes of maternal death are the following: obstetric hemorrhage, preeclampsia, eclampsia, abortion complications, and puerperal sepsis, which together represent 68% of all maternal deaths [1].

Women in the Americas who live in conditions of poverty, in remote places, lower level of education, indigenous and from other ethnic or racial groups and, in a situation of gender violence, have an overrepresentation in the maternal death rate\. It is recognized that 95% of maternal mortality could be avoided with the knowledge and technology already widely available in Mexico [3].

In recent years, Mexico has placed emphasis on social variables such as the economy and social status of patients. It has been found that the lack of formal education is associated with a six-fold increase in the risk of suffering a complication during pregnancy [4].

What makes the difference in the stability of pregnant women is that the faster a health problem can be diagnosed, the less likely a patient may develop a complication that could end in death [5].

2. THEORETICAL ASPECTS

Having tools that facilitate health management in patients, based on their particular risks, could help reduce maternal mortality, as stated by the World Health Organization [6].

Data science is an emerging multidisciplinary discipline that ranges from software development, artificial intelligence, data management and statistics [7]. Data science projects are based on identifying correlations, causal relationships, classifying and predicting events, identifying patterns and anomalies, and inferring probabilities. Machine learning is a branch of artificial intelligence that employs technologies to evolve into a valuable instrument for constructing systems that autonomously learn through existing knowledge [8], [9].

Having a greater availability of patient data could improve the training of machine learning algorithms and would allow addressing the problem of disease identification through the use of artificial intelligence [10], [11].

Machine learning is divided into two types: supervised learning and unsupervised learning. Unsupervised learning through principal component analysis, is a filter that allows the transformation of the data and that uses a search method, which in a reduction of dimensions is sought by choosing the characteristics that give a percentage of variance in the data and which filters attribute noise by removing features. It is mainly used in data science analysis and to build predictive models [12], [13].

Supervised learning consists of creating a function capable of predicting the value corresponding to any valid input object after having seen a series of examples called training data [10], [14]. To do this, the function has to generalize from the data presented to previously unseen situations; that is, it becomes feasible to uncover the connections among independent variables (also known as attributes or traits) and to generate forecasts that hold value in the process of making decisions.

3. MATERIAL AND METHODS.

The overall aim of this endeavor was to develop software that utilizes data science and machine learning to classify the risk of maternal mortality in Mexico.

The methods to achieve the specific objectives were:

1) Utilize data science and machine learning (retrospectively) on an open Ministry of Health database to reduce the number of risk variables associated with maternal mortality.
2) Create predictive models for the classification of diseases that can lead to maternal death using machine learning.
3) Evaluate the generated classification models.
4) Create software that uses the predictive model with the best results to classify pregnant women according to socio-clinical mortality risks.

4. DATA SCIENCE METHODOLOGY

This project is based on the methodology that IBM proposes for the application of data sciences [15]. This methodology is organized into ten stages that represent an interactive process: problem understanding, analytical approach, data requirements, data collection, data understanding, data preparation, modeling, evaluation, deployment, and feedback (see figure 1).

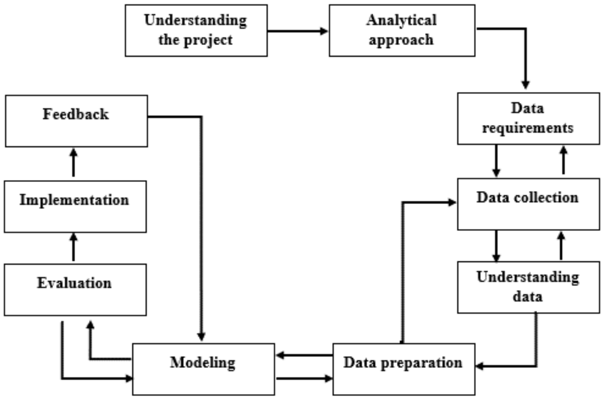


Figure_1. Fundamental methodology for IBM data science [15].

The generation of the predictive model for the diseases found was through training carried out with unsupervised machine learning and supervised.

Supervised learning algorithms were used: Naïve Bayes, Support Vector Machine (SVM), K Nearest Neighbors (KNN), Logistic Regression, and Unsupervised learning through Principal Component Analysis (PCA).

Several models were evaluated in order to create software that used the most accurate predictive model:

1) A software prototype was developed for data entry of pregnant patients that returns the classification of the patient according to risk factors that may cause death.
2) The Python programming language and the Scikit Learn library were used to build the predictive models.
3) The models were evaluated using cross-validation mechanisms.

Python is an interpretive, interactive, and object-oriented programming language. It provides high-level data structures such as lists and associative arrays (called dictionaries), dynamic syntax, dynamic linking, modules, classes, exceptions, automatic memory management, among others [16].

Scikit Learn is a Python module that integrates a wide range of state-of-the-art machine learning algorithms to solve medium-scale supervised and unsupervised learning problems [17].

5. DESCRIPTION OF THE STEPS FOLLOWED.

The ten steps followed in this research, based on the IBM data science methodology:

1) **Understanding the project:** The Ministry of Health of Mexico has an open database with data from each state, from 2002 to 2013 on maternal death. This database is organized into 58 variables. Within these variables is the variable " International Statistical Classification of Diseases and Related Health Problems " which contains a total of 248 diseases that are related to maternal deaths. The values within the CIE variable were employed as categories for the progression of this project. In essence, the CIE variable acts as the dependent variable.

2) **Analytical approach:** In order to analyze the maternal death data, it was decided to use supervised learning algorithms because there is a need to classify women in pregnancy control according to their mortality risk based on the data. open from the Ministry of Health of Mexico.

3) **Data requirement:** Each of the 14,308 records or instances in the data set obtained from the Ministry of Health is described by a set of characteristics or independent variables. These characteristics are descriptive (String data type) or numeric (Integer data type). However, in this investigation only numerical characteristics were used.

4) **Data collection:** The data set has 14,308 records of maternal deaths from the year 2002 to the year 2013 and with a total of 58 variables.

5) **Data Understanding:** Microsoft Excel was used in this step to facilitate visualization and understanding of the data set. Likewise, the set of data obtained from the Secretary's Office was compared with the public data from INEGI.

6) **Data preparation:** The techniques used in this investigation for data preparation are the following: Data errors and validation of empty or null data, Validation of data types and values (there are records with values outliers with the value 998). In total, 29 records with this problem were removed, leaving a total of 14,279 in the data set.

7) **Modeling:** In this stage, predictive models were generated using machine learning that could be used to predict high-risk diseases in pregnant women. At the beginning of the project, the use of the free license software "Weka" was contemplated as it is a popular program for data analysis and processing [18].

   However, when performing classification experiments, it was found that in version 3.6 of Weka for Microsoft Windows 10 Pro, the SVM and Logistic Regression algorithms are not enabled. To solve these problems, the "Libsvm-3.21" library was downloaded in order to be able to correctly access SVM from Weka, but no positive results were achieved.

   Thus, it was decided to venture into the Python programming language in which classification models could be developed with SVM, KNN, Naïve Bayes, and Logistic Regression. These algorithms were chosen because they are widely used for data analysis using supervised learning [19].

   In this instance, the CIE was treated as the category for each entry within the dataset. Similarly, in this investigation, the potential utilization of Principal Component Analysis (PCA) was taken into account to attempt the reduction of the quantity of attributes to be employed in the experiments.

8) **Evaluation:** During this phase, the assessments were conducted on the experiments carried out using the set of supervised learning algorithms selected in the preceding stage: Naïve Bayes, KNN, SVM, and Logistic Regression. In order to find the best predictive model when executing these algorithms, the accuracy, precision, recall and F1 of each generated model were calculated. To do this, the data set was divided into two: 70% of these data were used to carry out the training while 30% were used for the evaluation of the generated model.

   Training and evaluation were performed in Python version 2.7 running on Windows 10 pro (64-bit), on an HP Laptop with an AMD Athlon™ II P320 Dual-Core 2.10 GHz processor and 4 MB RAM.

9) **Deployment:** In this step, software was built to be able to predict the risk that a pregnant patient has of presenting a disease that could lead to death. For this purpose, the predictive model with the best results from the evaluation of the previous step was chosen in order to use it to classify the patient. The backend of this computer program is designed, the organization receives information about the performance of the model and the way in which it affects its environment is observed. Analysis of this information shows the data scientist whether to refine the model. This increases its precision and thus its utility was developed in Python. The frontend was written in PHP and HTML.

10) **Feedback:** By collecting the results of the implemented model. They are described in a more detailed section on results.

6. SOFTWARE METHODOLOGY

Once the predictive characteristics for maternal mortality from the most common causes found were obtained, a traditional cascade software development methodology was used. With the approach already described, and analysis of the data found, the design, programming and the corresponding tests were carried out to obtain the prototype product.

Prototyping was completed with the medical members of the team and with personnel from the local Health Jurisdiction for feedback. Prototyping is based on building a software prototype that is built quickly so that users can test it and provide feedback. Thus, you can fix what is wrong and include other requirements that may arise. It is an iterative model that is based on the trial and error method to understand the specifics of the product.

7. SEQUENCE DIAGRAM FOR THE EXECUTION OF THE PRODUCT

Figure 2 shows the UML sequence diagram of the modules present in the software product. The medical actor opens a form in which he will enter the patient's data (form.html). Once the questionnaire is completed, the form sends the data to costom.js which assigns numerical values to each entered value in order to be analyzed by the classification model.

proccess.php starts python.exe to run the ExecuteModel.py module. ExecutaModelo.py performs a classification with the new values, contained in an array, and returns the name of the class in which the new data was classified. The result is sent to form.html, which displays the patient's classification to the clinician via an alert message.

Figure 3 shows the interface prototype of the software developed for the capture of information from pregnant patients. This interface displays the questionnaire, which is made up of thirteen questions. Most of these questions have immediately below each of them the option to select the required information.

With the exception of the day of birth and the completed age, since these two fields are manually entered and have a limited range. The month of birth is from day 1 to day 31. The completed age ranges from 12 to 55 years. In the latter case, it was limited to these ages since they are the values with which the algorithm was trained.

The information is organized using the JSON format to enable presentation through form.html. The Python code ExecuteModel.py is responsible for taking the input data provided by medical professionals and converting it from textual representation (String) to numerical values (Integer), a process referred to as data discretization.

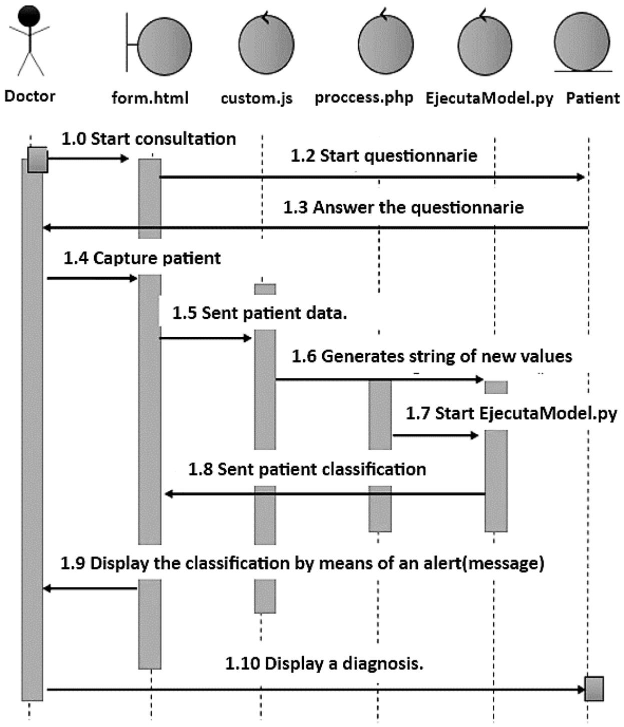
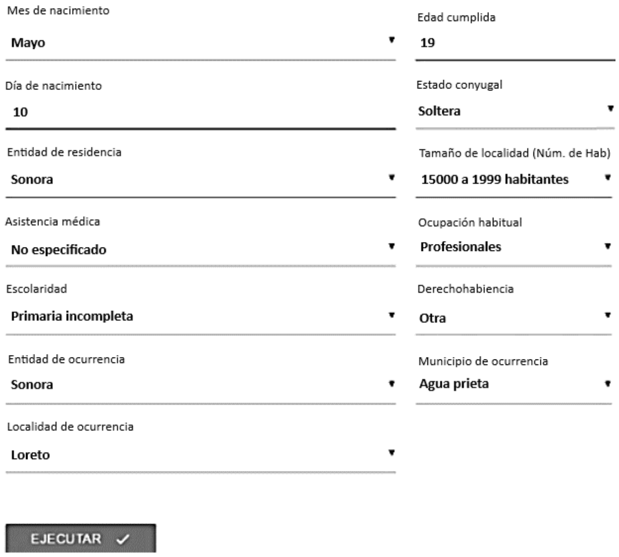

Doctor
form.html
custom.js
proccess.php
EjecutaModel.py
Patient
1.0 Start consultation
1.2 Start questionnarie
1.3 Answer the questionnarie
1.4 Capture patient
1.5 Sent patient data.
1.6 Generates string of new values
1.7 Start EjecutaModel.py
1.8 Sent patient classification
1.9 Display the classification by means of an alert(message)
1.10 Display a diagnosis.

Figure_2. Diagram of sequence



Mes de nacimiento
Mayo ▾

Edad cumplida
19

Día de nacimiento
10

Estado conyugal
Soltera ▾

Entidad de residencia
Sonora ▾

Tamaño de localidad (Núm. de Hab)
15000 a 1999 habitantes ▾

Asistencia médica
No especificado ▾

Ocupación habitual
Profesionales ▾

Escolaridad
Primaria incompleta ▾

Derechohabiencia
Otra ▾

Entidad de ocurrencia
Sonora ▾

Municipio de ocurrencia
Agua prieta ▾

Localidad de ocurrencia
Loreto ▾

Figure_3. Interface of the software product for the classification of pregnant women

This conversion facilitates analysis by the algorithms. The response message includes the name of the potential ailment that the patient might experience throughout pregnancy and in the subsequent 42 days following childbirth. This message is displayed through an alert message: Example in figure 4: “Risk factors present for eclampsia in labor”.

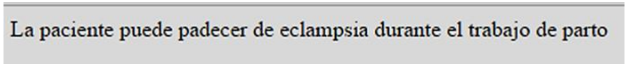

La paciente puede padecer de eclampsia durante el trabajo de parto

Figure 4. Alert Message showing the classification result.

8. RESULT

Experiments were carried out with four supervised learning algorithms, with the outstanding characteristic that they are classifiers: SVM, KNN, Logistic Regression and Naïve Bayes. In addition, PCA was applied in order to reduce the number of features that could be used for training.

Thus, apart from generating models with the application of individual algorithms, the following were also executed in conjunction with PCA: SVM + PCA, KNN + PCA, Logistic Regression + PCA, and Naïve Bayes + PCA. In all the experiments, 70% of the data set was used to carry out the training and 30% of them to carry out the evaluations of the trained model, Figure 5 shows the sequence of experiments performed.

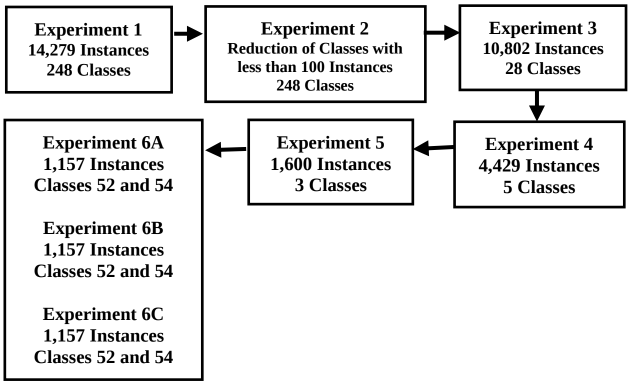


Figure_5. Diagram of the sequence of the experiments carried out.

Experiment_1.
The maternal death data set with 248 classes (diseases causing death) and 14,279 instances (number of patients who died from one of the classes -disease-) was used. Table 1 shows the comparison of the results obtained with each of the algorithms executed. The results show the average of the results for all classes with respect to precision, recall, f1-score, score and accuracy.

In this case, SVM is the algorithm that shows the highest score and accuracy, with the values closest to 1. It was found that, of the 248 initial classes, 51 of at least 100 expected correspond to a single instance. For this reason the algorithms could not be properly trained.

Table_1. Evaluation of eight distinct algorithms for analyzing maternal mortality data

Classifier algorithm	Accuracy	Precision	Recall	F1-Score
KNN	0.7182	0.06	0.06	0.06
PCA + KNN	0.7163	0.06	0.06	0.06
Naïve Bayes	0.0153	0.05	0.01	0.01
PCA + Naïve Bayes	0.0130	0.04	0.00	0.00
Regression Log.	0.1170	0.05	0.11	0.06
PCA + Regression Log.	0.1250	0.05	0.11	0.05
SVM	0.8948	0.04	0.08	0.02
PCA + SVM	0.6619	0.04	0.08	0.02

Experiment_2.
Classes that did not meet the requirements for a good classification were eliminated, according to Beleites et al. (2013). Of a total of 248 classes, 28 classes have one hundred or more instances. This allowed us to reduce the number of instances from 14,279 to 10,802. See table 2.

Table_2. Classes containing one hundred or more instances and the number of instances in each class.

Class	Number of instances
4	221
7	141
50	101
52	1219
53	134
54	1264
56	308
57	367
78	162
104	146
117	393
120	204
154	167
155	109
163	1055
164	1006
165	135
184	322
195	345
212	415
213	167
214	112
224	155
225	595
226	446
227	305
229	781
238	241

Experiment_3.
In this experiment the results of experiment 2 with instances belonging to 28 classes were used (see table 3).

Table_3. Experimenting with instances in 28 classes.

Classifier algorithm	Accuracy	Precision	Recall	F1-Score
KNN	0.9994	0.08	0.08	0.08
PCA + KNN	0.9996	0.09	0.09	0.09
Naïve Bayes	0.1524	0.12	0.14	0.11
PCA + Naïve Bayes	0.1524	0.11	0.11	0.09
Regression Log.	0.1558	0.14	0.15	0.09
PCA + Regression Log.	0.1623	0.10	0.17	0.10
SVM	0.9409	0.05	0.12	0.05
PCA + SVM	0.9409	0.11	0.12	0.06

Table 4 shows the five classes with a better classification. The column named "class" indicates the number of the class that obtained the best classification in terms of accuracy in the execution of the eight algorithms that contained a total of 28 classes. The column "Number of algorithms" indicates the number of times that each class was well classified within the group of eight algorithms. The "Best Accuracy" column indicates the algorithm with which the highest accuracy value was obtained when performing the experiments.

Emphasizing: The four classes obtained are the following: 52 eclampsia during labor, 54 eclampsia, in unspecified period, 163 secondary or late postpartum hemorrhage and 229 mild mental and behavioral disorders, associated with the puerperium, not elsewhere classified.

Table_4. Selection of classes by tendency to a better classification.

Class	Number algorithm	Better accuracy
52	7	KNN
54	7	KNN
163	6	KNN
229	3	KNN
155	1	SVM

Experiment_4.
According to the results of Experiment 3, in the present experiment the group of algorithms was applied once more on the data set with instances corresponding to classes 52, 54, 155, 163 and 229. (see table 4).

The results of this experiment are shown in table 5.

Table_5. Application of the group of algorithms to the set that contains five classes of experiment 3.

Classifier algorithm	Accuracy	Precision	Recall	F1-Score
KNN	1.00	0.34	0.34	0.34
PCA + KNN	1.00	0.33	0.33	0.33
Naïve Bayes	0.3769	0.38	0.38	0.37
PCA + Naïve Bayes	0.3769	0.40	0.37	0.35
Regression Log.	0.3827	0.35	0.37	0.34
PCA + Regression Log.	0.3884	0.34	0.36	0.35
SVM	0.9887	0.28	0.28	0.19
PCA + SVM	0.9887	0.31	0.30	0.20

In this case, the KNN and KNN+PCA classifiers present the highest accuracies. However, the results for precision and recall are low. This shows that the simple evaluation of accuracy is insufficient in this type of evaluation.

Experiment_5.
The three best classified classes by each algorithm were taken, applying the same method used in experiment 3. The classes that tended to be best classified are c**lasses 52 (eclampsia during labor), 54 (eclampsia, in period unspecified) and 163 (secondary or late postpartum hemorrhage).** With these three classes, the group of algorithms was applied again. In this case, neither result was optimal in terms of precision and recall (table 6).

Table_6. Group of algorithms to the data set that contains the classes 52, 54 and 163.

Classifier algorithm	Accuracy	Precision	Recall	F1-Score
KNN	1.00	0.40	0.40	0.40
PCA + KNN	1.00	0.41	0.41	0.41
Naïve Bayes	0.4943	0.47	0.45	0.44
PCA + Naïve Bayes	0.4943	0.54	0.48	0.45
Regression Log.	0.4968	0.49	0.48	0.48
PCA + Regression Log.	0.5093	0.55	0.52	0.52
SVM	1.00	0.27	0.35	0.23
PCA + SVM	1.00	0.31	0.36	0.23

Experiment_6.
Since the results in Experiment 5 were not satisfactory, in this experiment three combinations were made between **classes 52 (eclampsia during labor), 54 (eclampsia, in unspecified period) and 163 (secondary or late postpartum hemorrhage**) with a total of 1,157 instances. These experiments are named 6A, 6B and 6C and are presented below.
Experiment_6A (combination of classes 52 and 54) See table 7.

Table 7. Group of algorithms to the data set that contains classes 52 and 54.

Classifier algorithm	Accuracy	Precision	Recall	F1-Score
KNN	1.00	0.52	0.52	0.52
PCA + KNN	1.00	0.50	0.50	0.50
Naïve Bayes	0.5695	0.56	0.51	0.47
PCA + Naïve Bayes	0.5695	0.56	0.56	0.53
Regression Log.	0.5911	0.55	0.55	0.55
PCA + regression Log.	0.5980	0.59	0.59	0.59
SVM	1.00	0.49	0.50	0.46
PCA + SVM	1.00	0.42	0.47	0.34

Experiment_6B (combination of classes 54 and 163).
Table 8 shows the results when using the instances of **classes 54 (eclampsia, in unspecified period) and 163 (secondary or late postpartum hemorrhage)** with a total of 1,025 instances.

Experiment_6C (combination of Classes 52 and 163)
Table 9, Presents outcomes utilizing instances from categories 52 (eclampsia during labor) and 163 (secondary or late postpartum hemorrhage), totaling 1,018 instances.

In this table, an increase can be observed, both in accuracy, as well as in precision and recall of several algorithms. These results exceed those found in experiments 6A and 6B.

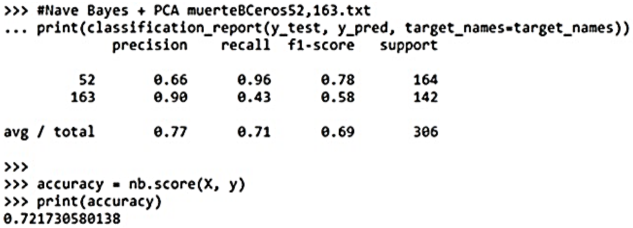
Figures_6, 7 and 8 belong to the three algorithms that best classify in table 9: Naïve Bayes, PCA + Naïve Bayes (see figure 6), Logistic Regression (see figure 7), PCA + Logistic Regression (figure_8).

Table_8. Group of algorithms to the data set that contains classes 54 and 163.

Classifier algorithm	Accuracy	Precision	Recall	F1-Score
KNN	1.0	0.45	0.55	0.39
PCA + KNN	1.0	0.60	0.61	0.60
Naïve Bayes	0.6985	0.74	0.72	0.70
PCA + Naïve Bayes	0.6985	0.75	0.71	0.69
Regression Log.	0.6858	0.70	0.70	0.70
PCA + regression Log.	0.6936	0.71	0.71	0.70
SVM	1.0	0.32	0.56	0.40
PCA + SVM	1.0	0.34	0.58	0.43

Table_9. Group of algorithms to the data set that contains classes 52 y 163.

Classifier algorithm	Precision	Recall	F1-Score	Accuracy
KNN	0.60	0.59	0.60	1.00
PCA + KNN	0.68	0.68	0.68	1.00
Naïve Bayes	0.79	0.75	0.73	0.7236
PCA + Naïve Bayes	0.76	0.72	0.70	0.7222
regression Log.	0.77	0.77	0.76	0.7364
PCA + regression Log.	0.70	0.71	0.70	0.7374
SVM	0.54	0.57	0.42	1.00
PCA + SVM	0.31	0.56	0.40	1.00



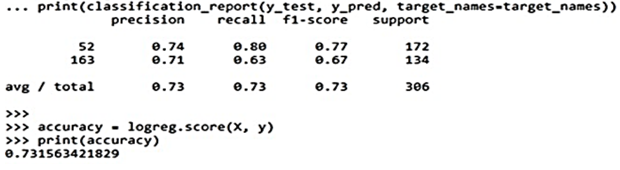
```
>>> #Nave Bayes + PCA muerteBCeros52,163.txt
... print(classification_report(y_test, y_pred, target_names=target_names))
             precision    recall  f1-score   support

         52       0.66      0.96      0.78       164
        163       0.90      0.43      0.58       142

avg / total       0.77      0.71      0.69       306

>>>
>>> accuracy = nb.score(X, y)
>>> print(accuracy)
0.721730580138
```

Figure_6. Classes 52 and 163 with the Naïve Bayes + PCA algorithm.



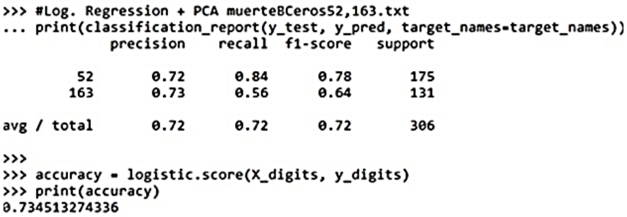
```
>>> #Log. Regression muerteBCeros52,163.txt
... print(classification_report(y_test, y_pred, target_names=target_names))
             precision    recall  f1-score   support

         52       0.74      0.80      0.77       172
        163       0.71      0.63      0.67       134

avg / total       0.73      0.73      0.73       306

>>>
>>> accuracy = logreg.score(X, y)
>>> print(accuracy)
0.731563421829
```

Figure_7. Classes 52 and 163 with the Logistic Regression algorithm.



```
>>> #Log. Regression + PCA muerteBCeros52,163.txt
... print(classification_report(y_test, y_pred, target_names=target_names))
             precision    recall  f1-score   support

         52       0.72      0.84      0.78       175
        163       0.73      0.56      0.64       131

avg / total       0.72      0.72      0.72       306

>>>
>>> accuracy = logistic.score(X_digits, y_digits)
>>> print(accuracy)
0.734513274336
```

Figure_8. Classes 52 and 163 with the Logistic Regression + PCA algorithm.

9. DISCUSSION AND CONCLUSIONS

In order to build a software, in this work the variables that can be used to carry out a classification of maternal death in Mexico were discovered, through the IBM data science methodology and machine learning (in a retrospective way) to classify pregnant patients due to risk of mortality. The model trained by Naïve Bayes with the instances of classes 52 and 163 is used to predict if any mother can be classified as one of the two most prevalent possible causes of death: **eclampsia and hemorrhage.** To this end, the medical personnel enter the following data: **month of birth, day of birth, completed age, marital status, entity of residence, size of town, habitual occupation, schooling, entitlement, entity, municipality, town, and medical assistance.** These data are used to show the results of the inference based on the trained Naïve Bayes model.

Precisely, the supervised learning algorithm that demonstrated the most favorable outcomes in categorizing maternal mortality risk was Naïve Bayes, achieving an accuracy of 0.7236, a general precision 0.79, a recall 0.75, an f1-score 0.73, and for the classes eclampsia during labor (precision 0.72, recall 0.96 and f1-score 0.83) and secondary or late postpartum hemorrhage (precision of 0.88, recall of 0.44 and f1-score of 0.59).

Unique Contributions of this Research
The particular contributions of this research are:

1) The number of risk variables for maternal death was reduced through the use of data science and machine learning, obtaining sociodemographic and epidemiological variables that can predict a clinical situation that can lead to maternal death. And that if social care is considered, maternal mortality in Mexico could be reduced.
2) Predictive models for the classification of diseases that can lead to maternal death were created using machine learning.
3) The generated classification models were evaluated.
4) A software prototype was created that uses the predictive model with the best results to classify

pregnant women according to maternal mortality risks in Mexico [20].

Future Perspectives

In the future, a series of investigations are anticipated, considering the following aspects:

1) It is anticipated that this research can be expanded by utilizing a more up-to-date dataset from the General Directorate of Health Information (2022). This will enable the inclusion of a larger number of instances, potentially leading to improved results.
2) The intention is to deploy the developed software prototype in real-world environments to obtain additional insights. This includes comparing the tool's outcomes against patient classification based on medical diagnosis.
3) The plan is to extend the program's functionality beyond merely classifying new patients, to also store their information. This expansion of the dataset will enhance the potential for creating future predictive models.

Acknowledgments: Saulo Hernández, Ana L. Santos Chong, Martha A. Olivas Dyk, Review & editing.

10. REFERENCES

[1] Pan American Health Organization and World Health Organization**, “Portal de Indicadores Básicos,”** May 21, 2023. https://opendata.paho.org/es/indicadores-basicos (accessed May 20, 2023).

[2] World Health Organization, **“The global health observatory,”** May 21, 2023. https://www.who.int/data/gho/indicator-metadata-registry/imr-details/4622 (accessed May 20, 2023).

[3] Pan American Health Organization, **“Situación de Salud en las Américas: Indicadores Básicos 2014,”** May 21, 2023. https://iris.paho.org/handle/10665.2/31074 (accessed May 20, 2023).

[4] M. T. LaFleur and J.Vélez, **“Determinantes de la salud materna e infantil y de los objetivos de desarrollo del milenio en Honduras,”** 2014. Accessed: May 20, 2023. [Online]. Available: https://www.un.org/en/development/desa/policy/capacity/presentations/honduras/Determinantes-de-MIyMM-en-Honduras.pdf

[5] A. Segura, **“Prevención, iatrogenia y salud pública,”** *Gac Sanit*, vol. 28, no. 3, pp. 181–182, May 2014, doi: 10.1016/j.gaceta.2014.02.002.

[6] World Health Organization, **“Maternal mortality,”** 2023. Accessed: May 20, 2023. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/maternal-mortality

[7] V. Dhar, **“Data science and prediction,”** *Commun ACM*, vol. 56, no. 12, pp. 64–73, Dec. 2013, doi: 10.1145/2500499.

[8] N. Kühl, M. Schemmer, M. Goutier, and G. Satzger, **“Artificial intelligence and machine learning,”** *Electronic Markets*, vol. 32, no. 4, pp. 2235–2244, Dec. 2022, doi: 10.1007/s12525-022-00598-0.

[9] P. Rattan, D._D. Penrice, and D_A. Simonetto, **“Artificial Intelligence and Machine Learning: What You Always Wanted to Know but Were Afraid to Ask,”** *Gastro Hep Advances*, vol. 1, no. 1, pp. 70–78, 2022, doi: 10.1016/j.gastha.2021.11.001.

[10] S. Kurdi *et al.*, **“Proof-of-concept Study of Using Supervised Machine Learning Algorithms to Predict Self-care and Glycemic Control in Type 1 Diabetes Patients on Insulin Pump Therapy,”** *Endocrine Practice*, Mar. 2023, doi: 10.1016/j.eprac.2023.03.002.

[11] D. Liu, K. Fox, G. Weber, and T. Miller, **“Confederated learning in healthcare: Training machine learning models using disconnected data separated by individual, data type and identity for Large-Scale health system Intelligence,”** *J Biomed Inform*, vol. 134, p. 104151, Oct. 2022, doi: 10.1016/j.jbi.2022.104151.

[12] J. Shlens, **“A Tutorial on Principal Component Analysis,”** Apr. 2014.

[13] C. Liang, L. Wang, L. Liu, H. Zhang, and F. Guo, **“Multi-view unsupervised feature selection with tensor robust principal component analysis and consensus graph learning,”** *Pattern Recognit*, vol. 141, p. 109632, Sep. 2023, doi: 10.1016/j.patcog.2023.109632.

[14] F. A. González, **“Modelos de aprendizaje computacional en reumatología,”** *Revista Colombiana de Reumatología*, vol. 22, no. 2, pp. 77–78, Jun. 2015, doi: 10.1016/j.rcreu.2015.06.001.

[15] J. B. Rollins, **“Metodología Fundamental para la Ciencia de Datos,”** 2015. Accessed: May 20, 2023. [Online]. Available: https://www.ibm.com/downloads/cas/WKK9DX51

[16] B. Zohuri, F. Mossavar-Rahmani, and F. Behgounia, **“Python programming–driven artificial intelligence,”** in *Knowledge is Power in Four Dimensions: Models to Forecast Future Paradigm*, Elsevier, 2022, pp. 827–836. doi: 10.1016/B978-0-323-95112-8.00026-X.

[17] A. Pajankar and A. **Joshi, Hands-on Machine Learning with Python. Berkeley,** CA: Apress_2022. doi: 10.1007/978-1-4842-7921-2.

[18] D.Merlini and M. Rossini, **“Text categorization with WEKA: A survey,”** *Machine Learning with Applications*, vol. 4, p. 100033, Jun. 2021, doi: 10.1016/j.mlwa.2021.100033.

[19] A. M. Sequeira, D. Lousa, and M. Rocha, **“ProPythia: A Python package for protein classification based on machine and deep learning,”** *Neurocomputing*, vol. 484, pp. 172–182, May 2022, doi: 10.1016/j.neucom.2021.07.102.

[20] R. D. Domínguez, **"Aplicación De Ciencia De Datos Para La Creación De Software Predictivo De Morbimortalidad Materna En México."** Order No. 29448469, Universidad de Montemorelos (Mexico), Mexico, 2017.