http://journalppw.com
ISSN 2587-0130

Algorithm for predicting the most frequent causes of mortality by analyzing age and gender variables.

Fredys A. Simanca H.[1]*, Jairo Cortés Méndez[2], Jaime Paez Paez[3], Alexandra Abuchar Porras[4], Jairo Palacios Rozo[5], Fabian Blanco Garrido[6]

[1]Universidad Cooperativa de Colombia; fredys.simanca@campusucc.edu.co
[2]Universidad Cooperativa de Colombia; jairo.cortes@campusucc.edu.co
[3]Universidad Cooperativa de Colombia; jaime.paez@campusucc.edu.co
[5]Universidad Distrital Francisco José de Caldas; aabucharp@udistrital.edu.co
[4]Universidad Colegio Mayor de Cundinamarca; jjpalacios@unicolmayor.edu.co
[6]Universidad Cooperativa de Colombia; fabian.blanco@campusucc.edu.co
**Correspondence: fredys.simanca@campusucc.edu.co*;

Abstract:

High density of populations in cities, complexity of risk factors that influence health and the impact of inequalities in sanitary outcomes, call for the adoption of decisive measures to improve health, and thus avoid injuries that trigger pathological events directly leading to the death of people. The above applies to Colombia and especially to Bogota D.C.; after the massive health crisis due to the pandemic. Consequently, it was proposed to implement a Prediction Algorithm based on a database directly taken from Salud Data and Salud Capital, which registered 31,720 deaths in Bogota in 2016, representing a rate of 397.5 deaths per 100,000 inhabitants, leading the list of the top ten causes: ischemic heart disease, with a rate of 65.8%, chronic respiratory tract diseases, with a rate of 26.4%, and cerebrovascular diseases with a rate of 25.7% per 100,000 inhabitants. The above data have shown the need to find a death prediction system, since it was difficult to predict the number of deaths that the pandemic was going to cause. It should be understood that the causes of mortality maintain a direct relationship with the medical study, as it evolves and develops according to the processing of the data obtained from the causes of mortality. Obtaining a good prediction system based on the data obtained greatly helps the medical area to centralize more efforts to counteract diseases with a higher rate, seeking to reduce the most significant causes of mortality. The algorithm designed analyzed two variables age and gender to predict the probability of death of a person with a percentage of 94.66% accuracy.

Keywords: health, algorithm, prediction, mortality

1. Introduction

The complexity of the risk factors that influence health, due to the predominance and high density of populations in cities, as well as the impact of inequalities in health outcomes, argue for the adoption of measures on the determinants of health to improve it and thus manage to avoid injuries that trigger pathological events that lead directly to the death of users. More than one billion people were added to urban areas between 2000 and 2014, so the United Nations estimates that more than 90% of future urban population growth will occur in low- and middle-income countries, which will generate an increase in risks not only economic but also health risks, difficult to control unless predictive measures are taken to reduce these risks in the population [1].

The above mentioned, has set in motion the increase of population control supported by technological progress, which in recent years of the Big Data era, has managed to generate machine learning tools, based on computational techniques such as logistic regression algorithms, neural networks, Python programming language among others, capable of modeling complex interactions with data variables, minimizing the margin of statistical error in different socio-economic activities, having outstanding field of action in predicting future risks, where high degrees of accuracy have been demonstrated [2] [3]. Medical science has been one of the areas with the greatest use of these technologies, as the use in standard statistical methods, which produce clinical risk prediction algorithms, for example, predicting the future risk of cardiovascular disease (CVD) among others [4].

Many of these algorithms demonstrate a high degree of accuracy, verified and replicated with several studies, being machine learning a way to explore results of even greater complexity and multifactorial causality, such as premature death. According to the above, it was intended to implement a prediction algorithm system of the most frequent causes of mortality in the city of Bogota, through the Python programming language, supporting such development, in favor of health and new technologies in the city, also addressing the proposed objectives to address problems exposed throughout this paper [5].

Mortality, according to current research approaches, is seen as the number of people who die in a given place and time in relation to the total population; this number is converted into data that are classified into their respective causes of death. According to the International Classification of Diseases, cause of death is defined as "the disease or injury that triggered the succession of pathological events leading directly to death, or the circumstances of the accident or act of violence that produced the fatal injury" [6]. Therefore, similar causes of mortality have been found throughout history, which have allowed them to be classified according to these similarities. Likewise, it should be understood that the causes of mortality maintain a direct relationship with the medical study, since this evolves and develops according to the treatment given to the data obtained in this regard. The objective of the analytical work is to statistically classify the death data, seeking to reduce the indexes with highest significance. Accordingly, obtaining a good prediction system based on data mining of this topic is greatly helpful for the medical area, centralizing more efforts to counteract diseases with higher rates.

In the most developed countries, the leading cause of death is heart disease, followed by Alzheimer's; in the least developed countries, many children do not even grow up, with the leading cause of death being neonatal conditions, followed by respiratory infections. In addition to the above, there is the COVID-19 pandemic, which could alter the list this year, according to experts from the OECD-UN health agency [7]. On the other hand, in Colombia, according to the DANE report for the years 1998 to 2010, one of the major diseases causing mortality at the general level was coronary heart disease, being the main cause of death, with a crude mortality rate of 64.59 per 100,000 inhabitants, followed by interpersonal violence with 40.51 per 100,000 and cerebrovascular disease with 26.92 [8]. Additionally, the database directly taken from Salud Data, Salud Capital shows for example registered rates for the year 2016 in Bogotá of 31,720 deaths, which represents a rate of 397.5 deaths per 100,000 inhabitants.

Topping the list of the top ten causes are: ischemic heart disease, with a rate of 65.8, chronic respiratory diseases, with a rate of 26.4, and cerebrovascular diseases with a rate of 25.7 per 100,000 inhabitants. Homicides ranked fourth in the city (16.1/100,000 inhabitants), followed by chronic diseases such as hypertension, diabetes and cancer. Of the registered deaths, 52% occurred in men and 48% in women. Seventy-one percent of the deaths occurred in the group of people over 60 years of age. The age group with the lowest number of deaths was the group aged 1 to 4 years [9]. However, according to the report updated to 2019, it records a gross mortality rate in the Colombian capital, stable between the years 2007 and 2019 [10].

The aforementioned, added to the pandemic that has aggravated the health crisis in an alarming way, point that it is necessary to find predictive tools that can prevent future diseases [11]. The above data have led to the need to find a probability system about the causes of mortality, since with the emergence of the unexpected pandemic it was difficult to predict the number of deaths that this was going to cause, for this reason, understanding and preventing diseases is a wise decision, if one takes into account that during a certain period of age a person can be affected by different causes of mortality and if that person has reason or knowledge of these, he/she can have greater probability and control to avoid them.

Likewise, the analytical studies of these data are essential for the control and development of the medical areas, since they will have a more direct approach based on the highest mortality rates. In this way, having a more direct vision of the problem, leads to a faster and more controlled reduction of health problems, thus avoiding unnecessary efforts trying to control non-existent problems. Similarly, the state is obliged to control the most frequent or probable causes in order to avoid social and health collapses [12], so that more budget should be invested in technologies and scientific areas.

It is worth taking as an example the pandemic that has affected countries, which have had to take several measures and invest large budgets to control the cases of death. Therefore, it is important to use the latest technology based on artificial intelligence to predict the causes of death in the city, using easy-to-use tools that would greatly help the health system to be more efficient at this health emergency situation that is happening globally [7]. For the above, the algorithmic application systems to be used have been analyzed, finding two types of predictive models: Classification and Regression models [13], being the classification models binary and a way to predict class belonging, for example, if one tries to classify among customers who are more likely to cancel the service. A classification model can be applied to our project, for example, classify by age which diseases have the highest mortality or classify which disease is more frequent in women and vice versa.

Regression models enable to predict a value, for example, what is the estimated profit we will obtain from a certain client in the coming months or provide results on the estimation of the sales forecast or, for example, what is the mortality rate for people suffering from leukemia [14] [15]. Taking into account the aforementioned, the methodology was analyzed and proposed to help prevent health cases that could be nonexistent, by developing a system for predicting causes of mortality, in order to determine the most frequent variable according to the age and gender of the person.

Consequently, it was intended to develop a prediction algorithm of the most frequent causes of mortality in Bogotá D.C.; using Python programming language to predict future events or results, through an interactive process, where the predictive model is developed according to the validation of data through automatic learning or Machine Learning, through a detailed understanding of the information processing algorithms by the computer, achieving valid tests to determine its accuracy. The above will allow an advance in achieving future results that will help the health sector to generate technological and practical tools to detect diseases in users in advance, achieving efficiency with health programs capable of predicting future diseases.

2. Materials and Methods

For the development of the algorithm, we worked with the Cause of Mortality database for the year 2018 in the city of Bogota (https://saludata.saludcapital.gov.co/osb/index.php/datos-de-salud/demografia/causasmortalidad2018/). The details of the data are specified in Table 1.

Table 1. Totals by cause of death year 2018 city of Bogotá.

Row Labels	Men	Women	Total
Motor transport accidents	40	26	66
Accidents that obstruct breathing	1	1	2
Assaults (homicides) and sequels	45	21	66
Accidental drowning and submersion	6	3	9

Anemias: nutritional, hemolytic, aplastic and others	4	1	5
Aortic aneurysm	2	2	4
Falls	24	3	27
Setbacks during medical care and sequels		2	2
Coagulation defects, purpura, purpura and other hemorrhagic and blood disorders and disorders affecting immunity	3	7	10
Nutritional Deficiencies	5	2	7
Diabetes mellitus	34	31	65
Pregnancy, childbirth and postpartum period		10	10
Cardiopulmonary disease and diseases of the pulmonary circulation.	1	4	5
HIV disease (AIDS)	36	17	53
Cerebrovascular diseases	54	51	105
Chronic diseases of the respiratory	26	25	51
Skin and subcutaneous cellular tissue diseases		3	3
Diseases of the blood vessels and other diseases of the circulatory system.	1	2	3
Diseases of the appendix, hernia and intestinal obstruction	5	4	9
Diseases of the esophagus and other diseases of the stomach and duodenum.		1	1
Liver diseases	18	6	24
Diseases of the peritoneum and all other diseases of the digestive system		1	1
Diseases of the lung due to external agents	4	3	7
Diseases of the musculoskeletal system and tissue.		1	1
Musculoskeletal system and connective tissue diseases	3	14	17
Glomerular and tubule-interstitial diseases	1	1	2
Hypertensive diseases	25	25	50
Intestinal infectious diseases	3	3	6
Ischemic heart diseases	56	43	99
Enteritis, non-infectious colitis and other diseases of the intestines	4	1	5
Accidental poisoning by and exposure to harmful substances	1	2	3
Epilepsy and other episodic and paroxysmal disorders		1	1
Epilepsy and other episodic and paroxysmal disorders	14	12	26

Exposure to electric current, radiation, and extreme ambient air temperature and pressure	1		1
Exposure to smoke, fire and flame		1	1
Fetus or newborn affected by certain maternal conditions	11	9	20
Fetus or newborn affected by obstetric complications and birth trauma	17	13	30
Gastrointestinal hemorrhage	1	1	2
Viral hepatitis and sequels	1		1
Specific infections of the perinatal period	17	14	31
Heart failure	1	2	3
Renal failure	8	9	17
Intentionally self-inflicted injuries (suicide)		1	1
Intentionally self-inflicted injuries (suicides) and sequels	32	17	49
Leukemia	27	31	58
Congenital malformations of the circulatory system	28	20	48
Congenital malformations, deformities and congenital anomalies	30	32	62
Melanoma and other malignant skin tumors	4	2	6
Meningitis and other inflammatory diseases of the central nervous system	6	3	9
Pneumonia	61	49	110
Other causes	129	105	234
Fetal growth retardation, fetal malnutrition, short gestation, and low birth weight.	6	3	9
Sepsis, except neonatal	1	4	5
Syphilis and other venereal diseases		2	2
Disorders of the gallbladder, biliary tract and pancreas	2	5	7
Hemorrhagic and hematologic disorders of the fetus and neonate	11	7	18
Mental and behavioral disorders		1	1
Respiratory disorders specific to the perinatal period	20	18	38
Tuberculosis and sequels	7	7	14
Malignant tumor of the breast in women		48	48
Malignant tumor of the prostate gland	19		19
Malignant tumor of the thyroid and other endocrine glands	4	2	6
Malignant tumor of the trachea, bronchi and lung	33	29	62

Malignant tumor of urinary bladder		1	1
Malignant tumor of the gallbladder and biliary tract		3	3
Malignant tumor of bones and articular cartilage	5	8	13
Malignant tumor of other parts of the uterus		3	3
Malignant tumor of the colon, junction and anus	34	45	79
Malignant tumor of the colon, rectosigmoid junction	2	1	3
Malignant tumor of the colon, rectosigmoid junction, rectum and anus	1		1
Malignant tumor of the cervix uteri		37	37
Malignant tumor of the brain, eye and other parts of the central nervous system	30	21	51
Malignant tumor of the stomach	47	41	88
Malignant esophageal tumor	3	1	4
Malignant liver tumor	2	7	9
Malignant ovarian tumor		20	20
Malignant pancreatic tumor	12	13	25
Secondary and poorly defined malignant tumors.	2	1	3
Malignant tumors of the lip, oral cavity and pharynx. ill-defined	3	3	6
Ulcer	2		2
Grand Total	**1036**	**969**	**2005**

The Machine Learning algorithm used to predict the probability of death was the Support Vector Machine (SVM). The objective of this classifier is to create a model that predicts the value of a variable by training a data set [16].

3. Results

A Python file was created, in which the algorithm for predicting the probability of death of a person was built. First, the *Pandas* and *Numpy* libraries were imported, and in line 3 the Excel file with the original data was read.

```
1.      import pandas as pd

2.      import numpy as np

3.      data = pd.read_excel("Causas Mortalidad.xlsx")
```

Through line 5, with the *isnull().sum()* function, you can know the amount of missing data in the data.

4. **data.isnull().sum()**

This function shows that there are three (3) null values in Percentage and Sex:

```
Location      0

Cause         0

Age           0

Total         0

Percentage    3
```

Rate	0
Sex	3

To solve the problem of missing data in Percentage and Sex, we replace the missing data with the averages. This is done for the Percentage column in lines 7, 8, 9 and 10 and for the Sex column in lines 12, 13 and 14.

Line 10 transforms the Sex column from categorical to Numeric. Since it is an object type, mathematical operations cannot be performed. For this it is used the *Numpy* library, which will make a conditional for: 0 female and 1 male.

```
5.  p = data["Porcentaje"].mean()
6.  promedioPorcentaje = int(p)
7.  data['Porcentaje'] = data["Porcentaje"].replace(np.nan, promedioPorcentaje)
8.  data['Sexo']=np.where(data['Sexo']=='Mujeres',0,1)
9.
10. s = data["Sexo"].mean()
11. promedioGenero = int(s)
12. data['Sexo'] = data["Sexo"].replace(np.nan, promedioGenero)
```

Now that all the columns have no missing data, we can proceed to the categorization of the Causes and Age columns. To transform the Causes column from categorical to numerical, *Numpy* is used, where each category is assigned with a value. There are 78 disease categories as lines 13 to 89 show.

```
13. data['Causa']=np.where(data['Causa']=='Malformaciones congÈnitas, deformidades y anomalÌas congÈnitas',1,
14. np.where(data['Causa']=='Feto o reciÈn nacido afectados por complicaciones obstÈtricas y traumatismo del nacimiento',2,
15. np.where(data['Causa']=='Malformaciones congÈnitas del sistema circulatorio',3,
16. np.where(data['Causa']=='Trastornos respiratorios especÌficos del periodo perinatal',4,
17. np.where(data['Causa']=='Enfermedades infecciosas intestinales',5,
18. np.where(data['Causa']=='Infecciones especÌficas del perÌodo perinatal',6,
19. np.where(data['Causa']=='Resto de causas',7,
20. np.where(data['Causa']=='NeumonÌa',8,
21. np.where(data['Causa']=='Feto o reciÈn nacido afectados por ciertas afecciones maternas',9,
22. np.where(data['Causa']=='Tumor maligno del encÈfalo, del ojo y de otras partes del sistema nervioso central',10,
23. np.where(data['Causa']=='Leucemia',11,
24. np.where(data['Causa']=='CaÌdas',12,
25. np.where(data['Causa']=='Enfermedades del hÌgado',13,
26. np.where(data['Causa']=='Lesiones autoinflingidas intencionalmente (suicidios) y secuelas',14,
27. np.where(data['Causa']=='Melanoma y otros tumores malignos de la piel',15,
28. np.where(data['Causa']=='Epilepsia y otros trastornos episÓdicos y paroxÌsticos',16,
29. np.where(data['Causa']=='Agresiones (homicidios) y secuelas',17,
30. np.where(data['Causa']=='Accidentes de transporte de motor',18,
31. np.where(data['Causa']=='Enfermedades isquÈmicas del corazÓn',19,
32. np.where(data['Causa']=='Enfermedad por VIH (SIDA)',20,
33. np.where(data['Causa']=='Enfermedades cerebrovasculares',21,
34. np.where(data['Causa']=='Enfermedades del hÌgado',22,
35. np.where(data['Causa']=='Insuficiencia renal',23,
36. np.where(data['Causa']=='Tumor maligno del colon, de la uniÓn  y ano',24,
```

37. np.where(data['Causa']=='Tumor maligno de la mama de la mujer',25,

38. np.where(data['Causa']=='Tumor maligno del estÓmago',26,

39. np.where(data['Causa']=='Tumor maligno del cuello del &#729;tero',27,

40. np.where(data['Causa']=='Tumor maligno del ovario',28,

41. np.where(data['Causa']=='Enfermedades del sistema osteomuscular y del tejido conjuntivo',29,

42. np.where(data['Causa']=='Tumor maligno de la tr·quea, los bronquios y el pulmÓn',30,

43. np.where(data['Causa']=='Enfermedades hipertensivas',31,

44. np.where(data['Causa']=='Tumor maligno del p·ncreas',32,

45. np.where(data['Causa']=='Tumor maligno de la prÓstata',33,

46. np.where(data['Causa']=='Diabetes mellitus',34,

47. np.where(data['Causa']=='Enfermedades crÓnicas de las vÌas respiratorias',35,

48. np.where(data['Causa']=='Trastornos hemorr·gicos y hematolÓgicos del feto y del reciÈn nacido',36,

49. np.where(data['Causa']=='Accidentes que obstruyen la respiraciÓn',37,

50. np.where(data['Causa']=='Ahogamiento y sumersiÓn accidentales',38,

51. np.where(data['Causa']=='Contratiempos durante la atenciÓn mÈdica y secuelas',39,

52. np.where(data['Causa']=='Tumor maligno del esÓfago',40,

53. np.where(data['Causa']=='Enteritis, colitis no infecciosa y otras enfermedades de los intestinos',41,

54. np.where(data['Causa']=='Tumor maligno de la tiroides y de otras gl·ndulas endocrinas',42,

55. np.where(data['Causa']=='Aneurisma aÓrtico',43,

56. np.where(data['Causa']=='Tumor maligno de los huesos y de los cartÌlagos articulares',44,

57. np.where(data['Causa']=='Tuberculosis y secuelas',45,

58. np.where(data['Causa']=='Deficiencias nutricionales',46,

59. np.where(data['Causa']=='Embarazo, parto y puerperio',47,

60. np.where(data['Causa']=='Tumor maligno del hÌgado',48,

61. np.where(data['Causa']=='Enfermedades del pulmÓn debida a agentes externos',49,

62. np.where(data['Causa']=='Defectos de coagulaciÓn, p&#729;rpura, p&#729;rpura y otras afecciones hemorr·gicas y de la sangre y los trastornos que afectan la inmunidad',50,

63. np.where(data['Causa']=='Meningitis y otras enfermedades inflamatorias del sistema nervioso central',51,

64. np.where(data['Causa']=='Enfermedades del esÓfago y otras enfermedades del estÓmago y del duodeno',52,

65. np.where(data['Causa']=='Enfermedades del apÈndice, hernia y obstrucciÓn intestinal',53,

66. np.where(data['Causa']=='Ulcera',54,

67. np.where(data['Causa']=='Trastornos de la vesÌcula biliar, de las vÌas biliares y del p·ncreas',55,

68. np.where(data['Causa']=='Retardo del crecimiento fetal, desnutriciÓn fetal, gestaciÓn corta y bajo peso al nacer',56,

69. np.where(data['Causa']=='SÌfilis y otras enfermedades venÈreas',57,

70. np.where(data['Causa']=='Enfermedades de la piel y del tejido celular subcut·neo',58,

71. np.where(data['Causa']=='Tumores malignos del labio, de la cavidad bucal y de la faringe',59,

72. np.where(data['Causa']=='Hepatitis viral y secuelas',60,

73. np.where(data['Causa']=='Enfermedades del peritoneo y todas las dem·s enfermedades del sistema digestivo',61,

74. np.where(data['Causa']=='Septicemia, excepto neonatal',62,

75. np.where(data['Causa']=='Enfermedades glomerulares y tubulointersticiales',63,

76. np.where(data['Causa']=='Anemias: nutricionales, hemolÌticas, apl·sicas y otras',64,

77. np.where(data['Causa']=='Tumor maligno del colon, de la uniÓn rectosigmoidea,',65,

78. np.where(data['Causa']=='Trastornos mentales y del comportamiento',66,

79. np.where(data['Causa']=='Insuficiencia cardiaca',67,

80. np.where(data['Causa']=='Tumor maligno de la vejiga urinaria',68,

81. np.where(data['Causa']=='ExposiciÓn al humo, fuego y llamas',69,

82. np.where(data['Causa']=='Tumor maligno de la vesÌcula biliar y de las vÌas biliares',70,

83. np.where(data['Causa']=='ExposiciÓn a la corriente elÈctrica, radiaciÓn y temperatura y presiÓn del aire ambientales extremas',71,

84. np.where(data['Causa']=='Envenenamiento accidental por y exposiciÓn a sustancias nocivas',72,

85. np.where(data['Causa']=='Tumores malignos de sitios mal definidos y secundarios',73,

86. np.where(data['Causa']=='Enfermedad cardiopulmonar y enfermedades de la circulaciÓn pulmonar',74,

87. np.where(data['Causa']=='Epilepsia y otros trastornos episÓdicos y paroxÌstica',75,

88. np.where(data['Causa']=='Tumor maligno del colon, de la uniÓn rectosigmoidea, recto y ano',76,

89. np.where(data['Causa']=='Hemorragia gastrointestinal',77,78
)))
))))))))))))))))))))

Finally, the Age column is categorized with *Numpy* and a conditional is created if it is "less than 1 year old" it is replaced by a 1 in the column or if it is "1 to 4 years old" it is replaced by two in the column and so on (Lines 90 to 97).

90. data['Edad']=np.where(data['Edad']==' Menor a 1 año',1,

91. np.where(data['Edad']=='1 a 4 años',2,

92. np.where(data['Edad']=='5 a 14 años',3,

93. np.where(data['Edad']=='15 a 44 años',4,

94. **np.where(data['Edad']=='45 a 59 años',5,6**)))))

Predictive algorithm application

To predict which gender is more prone to a disease, the following columns have been selected for X: CAUSES and AGE. For Y: GENDER.

The data to be analyzed are determined (X, Y).

95. X = data.iloc[:, [1,2]].values

96. y = data.iloc[:, 6].values

97. **y = y.astype(int)**

X and Y are divided into a test and training set. The test set has 30% of the data. This leaves 603 test data and 1405 training data.

98. from sklearn.model_selection import train_test_split

99. X_train, X_test, y_train, y_test= train_test_split(X, y, test_size=0.3)

In the case of a simple SVM, the parameter is set simply to "linear", since simple SVMs can only classify linearly separable data.

fit calls the SVC class method to train the algorithm on the training data.

100. from sklearn import svm

101. csvm = svm.SVC(kernel="linear")

102. csvm.fit(X_train,y_train)

To make the prediction, the predict method of the SVC class is used and the test x is passed as a parameter.

103. y_pred =csvm.predict(X_test)

The model is evaluated and the test Y and the predicted Y are passed as a parameter

104. from sklearn import metrics

105. print(metrics.recall_score(y_test, y_pred))

The above evaluation shows an accuracy of 94.66%.

The confusion matrix for the test set and prediction is printed.

106. from sklearn.metrics import confusion_matrix

107. cm = confusion_matrix(y_test, y_pred)

108. print(cm)

It is observed that, out of 603 test set data, there were 264 false negatives and 16 false positives.

```
[[ 39 264]
 [ 16 284]]
```

5. Conclusions

According to the development of the predictive algorithm proposed and looking for that this type of developments allow the health sector to detect diseases in advance, in users, making health systems more effective, since the identification of the main causes of mortality can serve as a basis in the design of plans to reduce these rates. The algorithm was designed in the Python programming language, using the Support Vector Machines algorithm and the Salud Data and Salud Bogotá database, which records deaths in the city of Bogotá and their cause.

We worked with a total of 2005 data, of which 1402 (70%) data were taken for training and 603 (20%) data for testing and validation of the algorithm's effectiveness. The various tests performed showed algorithm accuracy percentages ranging from a minimum of 88% to a maximum of 95%. Although it is true that these percentages are not high enough to be able to predict with a certain degree of reliability, it is evident that the main objective was achieved, which is to demonstrate that medicine or the health system can base its decision making on this type of development.

It is also evident that it is possible to continue using this type of predictive tools in future events that affect society in general and thus open a culture of data conservation at all levels, which can provide information and thus contribute to the identification of patterns. Making possible the prevention through the prediction of such unfortunate events or events that affect the community.

Reference

[1] World Health Organization, Global Report on Urban equitable, healthier cities for sustainable development, Geneva: World Health Organization, 2016, pp. 1-241.

[2] F. A. Simanca H. , D. Burgos R. , González Crespo y L. Rodriguez Baena, «Automatización del proceso de tutorización en entornos online a través de la análitica del aprendizaje,» de 13th Iberian Conference on Information Systems and Technologies (CISTI), Caceres, 2018.

[3] Harshith, «towardsdatascience.com,» towardsdatascience.com, 20 11 2019. [En línea]. Available: https://towardsdatascience.com/text-preprocessing-in-natural-language-processing-using-python-6113ff5decd8.

[4] S. F. Weng, L. Vaz, N. Qureshi y J. Kai, «Prediction of premature all-cause mortality: A prospective general population cohort study comparing machine-learning and standard epidemiological approaches,» Journal/ Plos One, vol. 14, nº 3, pp. 1-22, 27 03 2019.

[5] F. A. Simanca H., F. Blanco Garrido, J. Lemus, W. Torres Cerón, J. Palacios Rozo y L. Barbosa Guerrero, «Air Quality Index Prediction Model for the City of Bogotá,

DC,» Advances in Mechanics, vol. 9, nº 3, pp. 542-553, 2021.

[6] A. Gómez Rivadeneira, «Clasificación Internacional de Enfermedades (CIE): Descifrando la CIE-10 y esperando la CIE-11,» Monitor Estratégico, Bogotá, 2015.

[7] OCDE, «COVID-19 en América Latina y el Caribe: consecuencias socioeconómicas y prioridades de política,» Centro de Desarrollo OCDE, París, 2020.

[8] DANE, «Aspectos relacionados con la frecuencia de uso de los servicios de salud, mortalidad y discapacidad en Colombia, 2011.,» OBSERVATORIO NACIONAL DE SALUD, Bogotá, D.C., 2013.

[9] Salud Data, «Saludata.saludcapital.gov.co,» Salud Data, 30 12 2017. [En línea]. Available: https://saludata.saludcapital.gov.co/osb/index.php/datos-de-salud/demografia/causasmortalidad/.

[10] Salud Data, «Saludata.saludcapital.gov.co,» Salud Data, 9 02 2021. [En línea]. Available: https://saludata.saludcapital.gov.co/osb/index.php/datos-de-salud/demografia/tm-bruta/.

[11] J. D. De Estrada Riverón y A. R. Calzadilla, «Factores de riesgo en la predicción de las principales enfermedades bucales en los niños,» Revista Cubana de Estomatologia, vol. 38, nº 2, pp. 111-119, 2001.

[12] Minsalud, «Declaración de Emergencia Sanitaria por causa del Coronavirus COVID19,» Ministerio de Salud y Protección Social, Bogotá D.C., 2020.

[13] H. Robayo Barreto y J. Arias Ayala, «Desarrollo de un modelo de predicción de casos de deserción universitaria en el área de ciencias de la salud en Colombia,» Avenir, vol. 5, nº 1, pp. 24-30, 2021.

[14] F. A. Simanca H, M. Hernández Bejarano, A. Alfaro Tejeiro y J. Palacios Rozo, «Application of the Polynomial Regression Algorithm to Predict Covid-19 Cases Per Day in Colombia,» Advances in Mechanics, vol. 9, nº 3, pp. 49-61, 06 15 2021.

[15] J. Obando Bastidas, A. Peña Pita, L. Obando Vargas y A. Franco Montenegro, «Importance of nonlinear regression models in the interpretation of data from COVID-19 in Colombia,» Revista Habanera de Ciencias Medicas, vol. 19, nº 1, pp. 1-14, 2020.

[16] J. Gironés Roig, J. Casas Roma, J. Minguillón Alfonso y R. Caihuelas Quiles, Minería de datos: Modelos y algoritmos, Barcelona: Oberta UOC Publishing, 2017.

[17] A. Wasilewska y E. Menasalvas, «A Classification Model: Syntax and Semantics for Classification,» de Conference: Proceedings of the 10th international conference on Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing - Volume Part II, Berlin, 2005.

[18] Hablemos de SIG, «BIG DATA aplicada a los datos geográficos,» 14 01 2017. [En línea]. Available: https://shorturl.at/wyJL2.