Diabetes Early Detection and Risk Factors

Ramón Romero

Intelligent Systems

Introduction

Since 2000, diabetes mellitus is the first cause of death in women and the second in men in México.[1]

The diabetes is a deceseas where the glucose in blood is discomposed this is because of the lack or unoptimized use of the insulin, an hormone which regulates the metabolism of carbohydrates, fats and protein by promoting the absorption of nutrients to the cells[2][3]

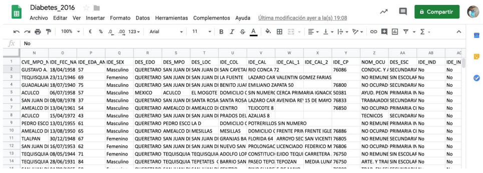
Purpose

Develop a diagnostic evaluation tool which can lead to develop consciousness in the population of Mexico about its actual situation and take actions in order to solve it.

The dataset

Based on data retrieved from Health Services of the State of Querétaro (SESEQ) from its Department of Epidemiology, the model will be generated by trying different supervised learning algorithms to find the best approach to the factors and its weight related to the detection of diabetes.

For this project, advice from doctors with field experience in the area of early detection of chronic diseases and epidemiological factors was implemented.



fx No

	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC
1	CVE_MPO_N	IDE_FEC_NA	IDE_EDA_AN	IDE_SEX	DES_EDO	DES_MPO	DES_LOC	IDE_COL	IDE_CAL	IDE_CAL_1	IDE_CAL_2	IDE_CP	NOM_OCU	DES_ESC	IDE_IND	IDE_IN
2	GUSTAVO A.	18/04/1958	57	Masculino	QUERETARO	SAN JUAN DI	SAN JUAN DI	SAN CAYETAI	RIO CONCA 72			76086	CONDUC. Y A	SECUNDARIA	No	No
3	TEQUISQUIA	23/11/1946	69	Femenino	QUERETARO	SAN JUAN DI	SAN JUAN DI	LA FUENTE	LAZARO CAR	VALENTIN GOMEZ FARIAS			NO REMUNE	SIN ESCOLAF	No	No
4	GUADALAJAI	18/07/1940	75	Masculino	QUERETARO	SAN JUAN DI	SAN JUAN DI	BENITO JUAF	EMILIANO ZAPATA 50			76800	NO OCUPAD	SECUNDARIA	No	No
5	ACULCO	06/07/1958	57	Masculino	MEXICO	ACULCO	EL MOGOTE	DOMICILIO C	SIN NUMERC	CERCA PRIMARIA IGNACIC		50381	AYUD. PEON	PRIMARIA IN	No	No
6	SAN JUAN DI	08/08/1978	37	Masculino	QUERETARO	SAN JUAN DI	SANTA ROSA	SANTA ROSA	LAZARO CAR	AVENIDA RE'	15 DE MAYO	76833	TRABAJADOI	SECUNDARIA	No	No
7	AMEALCO DI	13/04/1961	54	Masculino	QUERETARO	AMEALCO DI	AMEALCO DI	CENTRO	TEJOCOTE 8			76850	NO OCUPAD	PRIMARIA CO	No	No
8	ACULCO	15/04/1972	43	Masculino	QUERETARO	SAN JUAN DI	SAN JUAN DI	PRADOS DEL	AZALIAS 8				TECNICOS	SECUNDARIA	No	No
9	PEDRO ESCO	13/01/1955	61	Masculino	QUERETARO	PEDRO ESCO	LA D	DOMICILIO C	POTRERILLOS SIN NUMERO				NO REMUNE	PRIMARIA IN	No	No
10	AMEALCO DI	13/08/1950	65	Masculino	QUERETARO	AMEALCO DI	MESILLAS	MESILLAS	DOMICILIO C	FRENTE PRIN	FRENTE IGLE	76886	NO OCUPAD	PRIMARIA CO	No	No
11	TLALPAN	30/12/1948	67	Femenino	QUERETARO	SAN JUAN DI	SAN JUAN DI	GRANJAS BA	FLORIDA 64	ARROYO SEC	SAN VICENTI	76805	NO REMUNE	SECUNDARIA	No	No
12	SAN JUAN DI	16/07/1953	62	Femenino	QUERETARO	SAN JUAN DI	SAN JUAN DI	NUEVO SAN	PROLONGAC	LICENCIADO	FEDERICO M	76806	NO OCUPAD	PRIMARIA IN	No	No
13	TEQUISQUIA	08/05/1944	71	Femenino	QUERETARO	TEQUISQUIA	TEQUISQUIA	ADOLFO LOP	CONSTITUCI	EJIDO TEQUI	CARRETERA .	76750	NO REMUNE	SIN ESCOLAF	No	No
14	TEQUISQUIA	28/06/1931	84	Masculino	QUERETARO	TEQUISQUIA	TEPETATES (	BARRIO SAN	PASEO TEPO	TEPOZAN	MEDIA LUNA	76750	ARTE. Y TRAI	SIN ESCOLAF	No	No

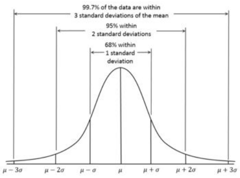
Research

After reading the similar cases where Machine Learning has been useful in order to predict the possibility, a model that has been useful with a decent level of prediction has been the Probabilistic Neural Network, proposed by Specht in 1990 [4].

All probabilistic models have a similar structure to the one presented by Specht:

- **Input layer**
  - Where the data is normalized.
- **Pattern layer**
  - Based on the input received from the input layer, the data is evaluated in order to find likelihood inside each of the elements of the dataset.
  - The most common of the likelihood metrics is the **Euclidean distance** (RBF Kernel) based on a **Gauss Standard Distribution**

$$\sqrt{\sum_{i=1}^{n}(q_i - p_i)^2}.$$



- **Summation Layer**
  - From the weight generated from the pattern layer, it sums the value of each of the features and its corresponding weights

- **Output layer**
    - Based on the summation layer, it is usually a max function where the output is the class or label with the most likelihood.

Development

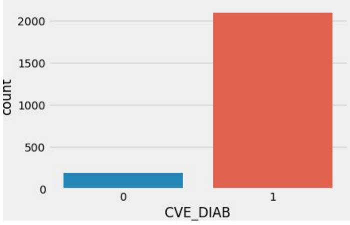
Data Analysis and Encoding

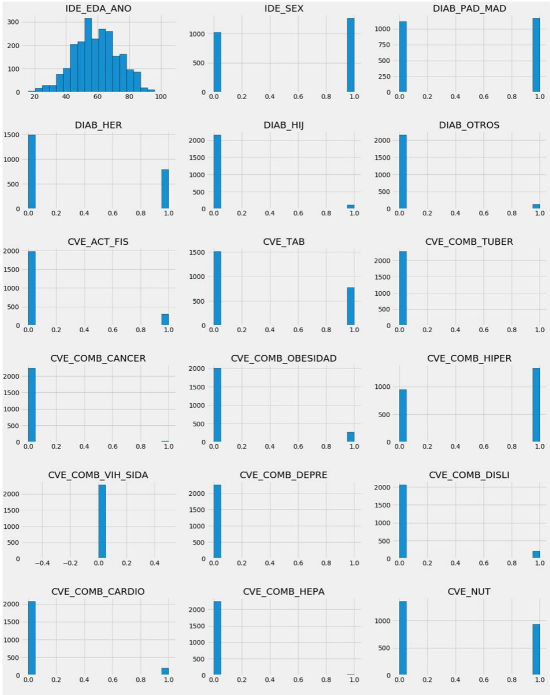
The purpose of the info is to have a medical record of the people who attends to the public hospitals in the city, because of that, features as the name or social number were removed.

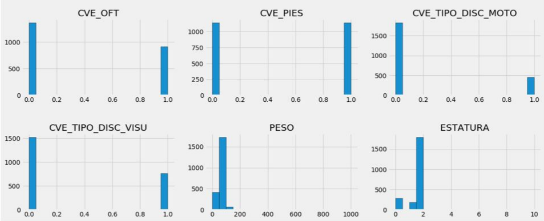
The remaining features were the following.

- Age
- Sex
- Weight
- Height
- Diabetic Familiars
- Addictions and habits
- Chronical Diseases
- Disabilities

Label Distributions







Models

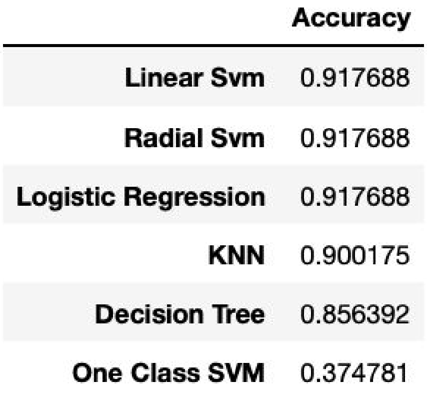
Looking for the best model to predict, the following models were tested.

- Non-probabilistic binary classifiers
  - K Nearest Neighbor (Euclidean)
  - DecisionTreeClassifier
  - Supporting Vector Machines
    - OneClass (Outlier Detection)
    - Linear (Line)
    - Radial (Euclidean)
- Probabilistic binary classifiers
  - Logistic Regression

Implementation

The data set was splitted in 2 subsets, the training set and test set (80 % and 20%)

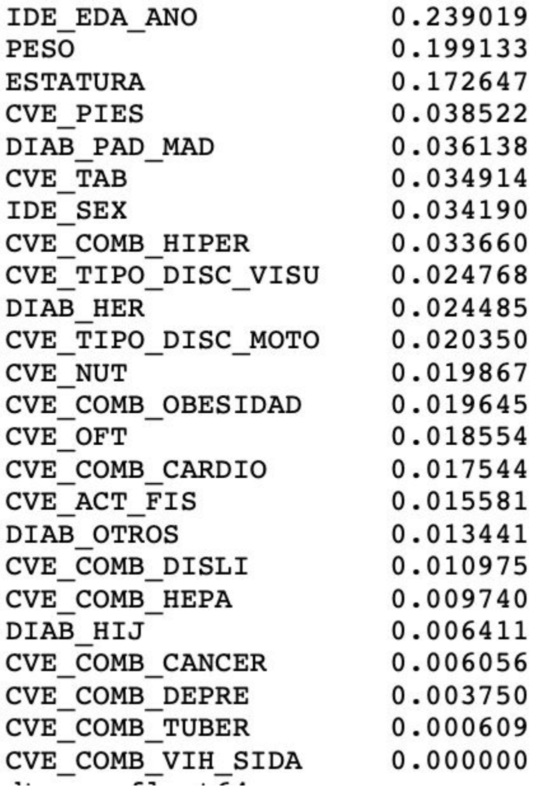
The learning implementation was done by using python and the scikit-learn toolbox, with the following initial results



	Accuracy
Linear Svm	0.917688
Radial Svm	0.917688
Logistic Regression	0.917688
KNN	0.900175
Decision Tree	0.856392
One Class SVM	0.374781

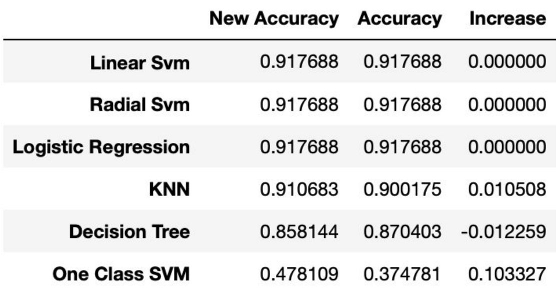
In order to improve the models, another algorithm was added to the sequence: Random Forest Classifier, which is in some point of view similar to a Decision Tree with the main difference that it takes random attributes from the data set in order to find the "nodes" with more impact on the final results.

This is the impact that each attribute has on the final output



```
IDE_EDA_ANO            0.239019
PESO                   0.199133
ESTATURA               0.172647
CVE_PIES               0.038522
DIAB_PAD_MAD           0.036138
CVE_TAB                0.034914
IDE_SEX                0.034190
CVE_COMB_HIPER         0.033660
CVE_TIPO_DISC_VISU     0.024768
DIAB_HER               0.024485
CVE_TIPO_DISC_MOTO     0.020350
CVE_NUT                0.019867
CVE_COMB_OBESIDAD      0.019645
CVE_OFT                0.018554
CVE_COMB_CARDIO        0.017544
CVE_ACT_FIS            0.015581
DIAB_OTROS             0.013441
CVE_COMB_DISLI         0.010975
CVE_COMB_HEPA          0.009740
DIAB_HIJ               0.006411
CVE_COMB_CANCER        0.006056
CVE_COMB_DEPRE         0.003750
CVE_COMB_TUBER         0.000609
CVE_COMB_VIH_SIDA      0.000000
```

For the next implementation was only considered the first three attributes as it they have the most relevant impact on the output.(height, weight and age)



	New Accuracy	Accuracy	Increase
Linear Svm	0.917688	0.917688	0.000000
Radial Svm	0.917688	0.917688	0.000000
Logistic Regression	0.917688	0.917688	0.000000
KNN	0.910683	0.900175	0.010508
Decision Tree	0.858144	0.870403	-0.012259
One Class SVM	0.478109	0.374781	0.103327

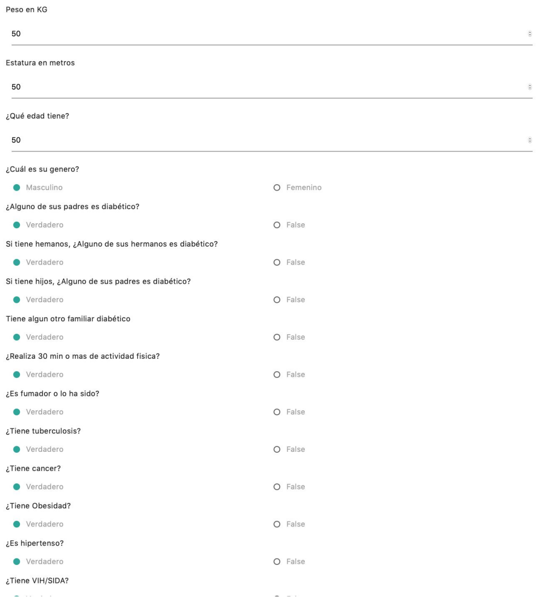
Deployment

After training the models, the weight and normalizers are stored. (learn.py -> outputs)

This models are useful for us as can be easily re-implemented in another form, in this case the models will be used as back-end for an app developed using Flask.

The app consist of 2 elements-views:

- A test:



Peso en KG

50

Estatura en metros

50

¿Qué edad tiene?

50

¿Cuál es su genero?

● Masculino ○ Femenino

¿Alguno de sus padres es diabético?

● Verdadero ○ False

Si tiene hemanos, ¿Alguno de sus hermanos es diabético?

● Verdadero ○ False

Si tiene hijos, ¿Alguno de sus padres es diabético?

● Verdadero ○ False

Tiene algun otro familiar diabético

● Verdadero ○ False

¿Realiza 30 min o mas de actividad fisica?

● Verdadero ○ False

¿Es fumador o lo ha sido?

● Verdadero ○ False

¿Tiene tuberculosis?

● Verdadero ○ False

¿Tiene cancer?

● Verdadero ○ False

¿Tiene Obesidad?

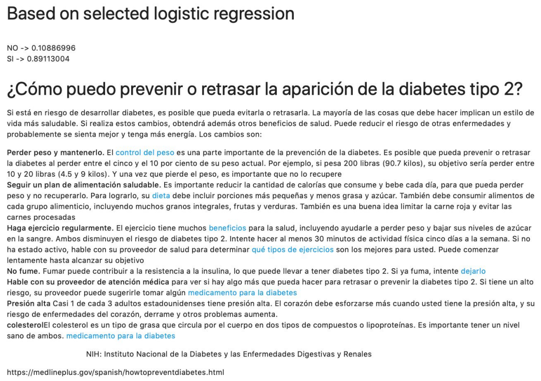
● Verdadero ○ False

¿Es hipertenso?

● Verdadero ○ False

¿Tiene VIH/SIDA?

- The results and recommendations



Based on selected logistic regression

NO -> 0.10886996
SI -> 0.89113004

¿Cómo puedo prevenir o retrasar la aparición de la diabetes tipo 2?

Si está en riesgo de desarrollar diabetes, es posible que pueda evitarla o retrasarla. La mayoría de las cosas que debe hacer implican un estilo de vida más saludable. Si realiza estos cambios, obtendrá además otros beneficios de salud. Puede reducir el riesgo de otras enfermedades y probablemente se sienta mejor y tenga más energía. Los cambios son:

Perder peso y mantenerlo. El control del peso es una parte importante de la prevención de la diabetes. Es posible que pueda prevenir o retrasar la diabetes al perder entre el cinco y el 10 por ciento de su peso actual. Por ejemplo, si pesa 200 libras (90.7 kilos), su objetivo sería perder entre 10 y 20 libras (4.5 y 9 kilos). Y una vez que pierde el peso, es importante que no lo recupere

Seguir un plan de alimentación saludable. Es importante reducir la cantidad de calorías que consume y bebe cada día, para que pueda perder peso y no recuperarlo. Para lograrlo, su dieta debe incluir porciones más pequeñas y menos grasa y azúcar. También debe consumir alimentos de cada grupo alimenticio, incluyendo muchos granos integrales, frutas y verduras. También es una buena idea limitar la carne roja y evitar las carnes procesadas

Haga ejercicio regularmente. El ejercicio tiene muchos beneficios para la salud, incluyendo ayudarle a perder peso y bajar sus niveles de azúcar en la sangre. Ambos disminuyen el riesgo de diabetes tipo 2. Intente hacer al menos 30 minutos de actividad física cinco días a la semana. Si no ha estado activo, hable con su proveedor de salud para determinar qué tipos de ejercicios son los mejores para usted. Puede comenzar lentamente hasta alcanzar su objetivo

No fume. Fumar puede contribuir a la resistencia a la insulina, lo que puede llevar a tener diabetes tipo 2. Si ya fuma, intente dejarlo

Hable con su proveedor de atención médica para ver si hay algo más que pueda hacer para retrasar o prevenir la diabetes tipo 2. Si tiene un alto riesgo, su proveedor puede sugerirle tomar algún medicamento para la diabetes

Presión alta Casi 1 de cada 3 adultos estadounidenses tiene presión alta. El corazón debe esforzarse más cuando usted tiene la presión alta, y su riesgo de enfermedades del corazón, derrame y otros problemas aumenta.

colesterolEl colesterol es un tipo de grasa que circula por el cuerpo en dos tipos de compuestos o lipoproteínas. Es importante tener un nivel sano de ambos. medicamento para la diabetes

NIH: Instituto Nacional de la Diabetes y las Enfermedades Digestivas y Renales

https://medlineplus.gov/spanish/howtopreventdiabetes.html

Observations and improvements

The logistic regression result is shown as an output to the test because different to the other algorithms it output us **probability distribution, not only the binary classification**.(Result for other models are shown on the bottom part of the results page)

By the show in the tests, and also live testing of the app, it seems that the mexican have an added bias to the probability of being diabetic, it is because of the living style (habits and food)

Having a bigger data set may improve the knowledge of the less important features and its impact.

Based on the medical practice, the Random forest algorithm for feature selection correctly predicted the most important aspect to evaluate a possible diabetic patient as were the **height, weigh and age**. Aspects used to calculate body mass index and it classification and predisposition to also other chronic diseases.

From presentation:

The Outlier was the worst of the models,it seems that **the data is too simila**r to find a shape that can separate the ones who are diabetic to the ones which are not

Ridiculous examples that would suggest a clear predisposition to diabetes such as extreme high weight, age and short height, was classified as outlier as they seem not similar to the common population.

References

[1]Rojas Martínez, María Rosalba, et al, “Epidemiología de la diabetes mellitus en México”, en Aguilar Salinas, Carlos A. et al, (eds), Acciones para enfrentar a la diabetes. Documento de postura. Academia Nacional de Medicina de México, México, 2015.

[2] "Diabetes tipo 2: MedlinePlus enciclopedia médica", Medlineplus.gov, 2019. [Online]. Available: https://medlineplus.gov/spanish/ency/article/000313.htm. [Accessed: 11- Apr- 2019].

[3]Stryer L (1995). Biochemistry (Fourth ed.). New York: W.H. Freeman and Company. pp. 773–74. ISBN 0 7167 2009 4.

[4] Specht, D. F. (1990). "Probabilistic neural networks". Neural Networks. 3: 109–118. doi:10.1016/0893-6080(90)90049-Q.

[5]"scikit-learn: machine learning in Python — scikit-learn 0.21.0 documentation", Scikit-learn.org, 2019. [Online]. Available: https://scikit-learn.org/stable/. [Accessed: 11- May- 2019].

[6]"Pima Indians Diabetes Database", Kaggle.com, 2019. [Online]. Available: https://www.kaggle.com/uciml/pima-indians-diabetes-database. [Accessed: 11- May- 2019].