Diabetes Health Indicators: Don't Underestimate Your Health

Francesco Angiulli, Alice Parodi, Matteo Pasotti

17 February 2023

Abstract

How can diabetes be recognized in a person without appropriate exams? Diabetes is a sly disease, as some of its symptoms can appear after it has been present for years [1]. Many people discover they have diabetes by chance through a simple blood test, but what are the indicators that identify a possible presence of diabetes? This study is developed on this question. Through a set of decision tree learning algorithms we will try to evaluate which possible factors trigger diabetes and how much they affect this result.

Contents

1 Introduction

Diabetes is a chronic disease that affects the way the body processes blood sugar. It is caused by either a deficiency of insulin production or a failure of cells to properly respond to insulin. If left uncontrolled, high blood sugar levels can lead to long-term damage of various body systems, including the nerves and blood vessels. This is known as hyperglycemia. It is important for individuals with diabetes to carefully manage their condition to avoid serious health complications.

There are certain groups of people who are particularly at risk of developing diabetes and may already have more or less severe complications. Through a set of decision tree learning algorithms we will investigate which complications increase the probability of developing diabetes. Diabetes is divided into two levels of relevance, but in this study we only consider the presence or absence of the disease in the patient.

2 Presentation of the dataset

Data is produced by CDC's BRFSS, Behavioral Risk Factor Surveillance System [2], a health telephone survey collected each year; in particular, this is related to the year 2015.

Moreover, it has been synthesized using CTGAN. This data set contains the answers of 40.109 Americans to 17 variables plus the target variable *Diabetes*, where 0=no diabetes and 1=diabetes:

- **Age**, *the age of the patient*
  1=18-24, 2=25-29, 3=30-34, 4=35-39, 5=40-44, 6=45-49, 7=50-54, 8=55-59, 9=60-64, 10=65-69, 11=70-74, 12=75-79, 13=80+ (ordinal)
- **Sex**, *the sex of the patient*
  0=female, 1=male (binary)
- **HighChol**, *presence or not of high cholesterol*
  0=no high cholesterol, 1=high cholesterol (binary)
- **CholCheck**, *cholesterol check in the last five years*
  0=no, 1=yes (binary)
- **BMI**, *Body Mass Index of the patient*
  (numeric)
- **Smoker**, *answer to the question 'Have you smoked at least 100 cigarettes in your entire life?'*
  0=no, 1=yes (binary)
- **HeartDiseaseorAttack**, *presence or not of coronary heart disease or myocardial infarction*
  0=no, 1=yes (binary)
- **PhysActivity**, *physical activity in the past 30 days*
  0=no, 1=yes (binary)
- **Fruits**, *consume of fruit one or more times per day*
  0=no, 1=yes (binary)
- **Veggies**, *consume of vegetables one or more times per day*
  0=no, 1=yes (binary)
- **HvyAlcoholConsump**, *for males more than 14 drinks per week and for females more than 7 drinks per week*
  0=no, 1=yes (binary)
- **GenHlth**, *Would you say that in general your health is*
  1=excellent, 2=very good, 3=good, 4=fair, 5=poor (ordinal)
- **MenHlth**, *Days of poor mental health*
  scale from 1 to 30 (numeric)
- **PhsyHlth**, *physical illness or injury days in past 30 days*
  scale from 1 to 30 (numeric)
- **DiffWalk**, *Do you have serious difficulty walking or climbing stairs?*
  0=no, 1=yes (binary)
- **Hypertension**, *presence or not of hypertension*
  0=no hypertension, 1=hypertension (binary)
- **Stroke**, *Ever told you had a stroke*
  0=no, 1=yes (binary)

The dataset has a similar percentage for both presence and absence of the diabetes: 48,91% for no diabetes and 51,09% for diabetes. Moreover there are no missing values, so there are no problems of missing replacement. Despite this, it can be observed that few observations are exactly the same of others, in particular 2456 duplicates, but we decided not to delete them because they can simply be people that share the same values for all the variables. In figure 1, the contingency table between diabetes and age is shown.

Age	No_Diabetes	Yes_Diabetes
1	451	64
2	717	120
3	1002	200
4	1211	409
5	1343	582
6	1719	1076
7	2089	1764
8	2313	2308
9	2374	3170
10	2483	3957
11	1727	3157
12	1026	1889
13	1164	1793

Figure 1: Number of people with and without Diabetes among different Age categories

3 Pre-processing

Pre-processing is a broad area; it is essential in order to make data more suitable for data analysis. The first step is the normalization of the numerical variables: transform them into the interval [0,1], otherwise the machine learning algorithm will be dominated by the variables that use a larger scale. The majority of the variables in the dataset are binary, with values 0 or 1. As a consequence, the normalization in [0,1] is applied to *BMI*, *MenHlth* and *PhsyHlth*.

Secondly, feature selection should be performed. In order to complete this goal, different approaches are followed. The correlation matrix between all the variables shows us that the maximum value of correlation is between *GenHlth* and *PhsyHlth* and it is equal to 0,527, quite low. So, no variable seems to be redundant. Moreover, the correlation between diabetes and the other variables is at most 0,403 showing that there are no irrelevant variables. An important annotation is that in order to compute the correlation matrix, the hot one encoding is performed because the correlation can only be present between numeric variables. Another thing is to filter out variables with low variance, below a user defined threshold. Columns with low variance are likely to distract certain learning algorithms and are therefore better removed. At the beginning, we set 0.16 as threshold for variance upper bound. This because the goal was to remove all features that were either one or zero in more than 80% of the samples, so $0.8(1 - 0.8)$ as threshold. The problem was that it deleted a lot of variables that ended up to be significantly important for the diabetes prediction.

The Variance Inflation Factor (VIF) measures the severity of multicollinearity in regression analysis. It is a statistical concept that indicates the increase in the variance of a regression coefficient as a result of collinearity. In our dataset the highest value of VIF turned out to be 1,79. As a consequence, no one was removed because the automatic threshold was set to 5. It can be affirmed than that there is no multicollinearity in the data.

During the pre-processing, feature construction can be a further step. The idea is combining two or more attributes to build more efficient features for the machine data mining task. For example, our first deduction was to create a variable *Diet* as a combination of *Fruits* and *Veggies*. It is useful if it improves the classification performance, but clearly here it doesn't, probably there is a too high loss of information. Furthermore, as stated above, the correlation between the variables is quite small so there was no need of combining them.

Last but not least, *Age* and *GenHlth* were not transformed to binary attributes because the XGBoost, the model chosen (see section 4.5), categorizes them automatically.

4 Models

In this analysis different machine learning techniques were applied to find medical answers for diabetes prediction, with the aim of discovering the best solution based on the dataset available. The decision on the best model is obtained on the value of Log loss, indicative of how close the prediction probability is to the corresponding actual class. Log loss is similar to the accuracy, but it will favor models that distinguish more strongly the classes, not only in output but in probabilistic outcome. All the value of log loss listed below are obtained after the performance of a parameter optimization, in order to have a faithful comparison to XGboost.

Classification techniques can be divided into:

- **Heuristic models** as Decision trees and Random forests
- **Regression based models** as Logistic regression
- **Separation models** as Support vector machines and Artificial neural networks
- **Probabilistic models** as Naive bayes and Tree-augmented naive bayes

4.1 Heuristic models

Even though they don't allow us to obtain robust and clear results, they are very commonly used. These models are composed on the idea of the tree with roots and leaves, so nodes and arches. In the decision tree, for instance, each record is classified from the top to the bottom thanks to different measures depending on the type of attribute selected. In our analysis, the decision tree has a value of log loss equal to 0,240. It is very important for our further analysis, because our final model is based on the concept of decision tree (see section 4.5).

While decision tree combines some decisions, a random forest combines several decision trees. Its log loss is 0,244 so quite close to the one of the decision tree.

4.2 Regression based models

This type of models uses parametric conditional probability to find the effect of input variables on different output. So the answer is given on the values of the input that influence mostly the class attribute. The log loss associated to the logistic regression is 0,225, lower than both the heuristic models. Unfortunately, the logistic regression allows as explanatory attributes only numeric ones, but in this dataset the majority of the attributes are binary. As a consequence, we think there would be a misalignment with theory behind.

4.3 Separation models

The separation models partition the attributes' space. Artificial neural networks is a large set of models such as multi layer perceptron based on the idea of connected neuron in a linear way. Its log loss has a very good value (0,223), the closest to the one of XGboost. The probabilistic artificial network was also executed and has a log loss of 0,255.

4.4 Probabilistic models

They exploit Bayes formula and compute posterior probability of the class attributes given the exploratory attributes in order to classify records. A widely used example is the naive bayes that is based on the idea of independence between attributes. The naive bayes has a log loss of 0,447, higher than all the other ones analysed before. Differently, AODE was developed to address the attribute-independence problem of the popular naive Bayes classifier and so to give more accurate classifier. Unfortunately in this study the log loss with AODE classifier is higher than the naive bayes one, 0,254.

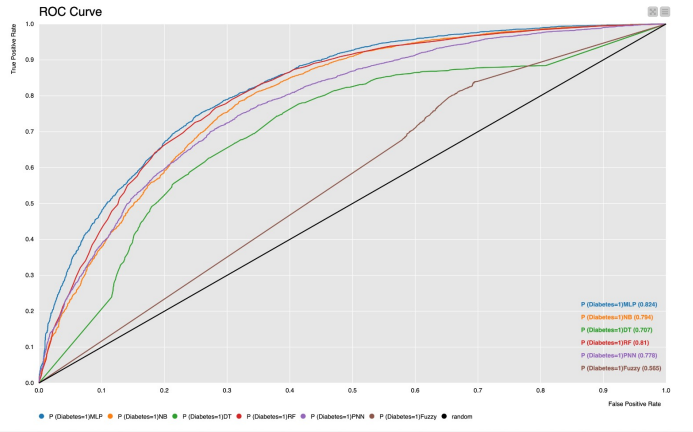

Figure 2: ROC curve of the main models

4.5 XGboost

The final result is that XGBoost is the best model for our dataset and the analysis' goal. Based on the fact that the decision tree has the lowest value of log loss, we searched for a model which could enhance the one stated above. XGboost is a popular machine learning library that is based on the idea of boosting. It is a way for transforming the weak observations into strong ones after some steps. As such, the predictions on the last model will be the result of the previous tree models. The gradient boosting is a way to automatically develop the machine learning: the following model will reduce errors to a minimum. It helps the predictive models and solve problems of multicollinearity (see section 3). To be more

specific, the node "XGboost Tree Ensemble Learner" is used because, in addition to the model, the feature importance is obtained. A paragraph on it can be found here (5).

At the end of parameter optimization (see 4.5.1), the parameters that allow the lowest value of logloss are used by the learner to learn the classification model. The inducer is than queried on the testing data to predict the values of the class attribute. The value of logloss of the test set after having applied the XGboost with the best parameters is 0,2209. The ROC curve related to this model is shown in the Appendix C.

4.5.1 Hyperparameter optimization

After the pre-processing step, the dataset is divided into training and test set with a stratified partitioning on the class attribute diabetes. This is done to be sure that both the presence and the absence of diabetes are in the partitions. The training set is then partitioned once more through the node "X partitioner", which is the starting point of a cross validation loop. It is internal to another loop related to parameters optimization. The last is done in order to find the best parameters possible to the model, starting from the predefined ones using a bayesian optimization strategy. For each set of hyper parameters, the cross validation allows us to find K values of accuracy, with K equal to the number of disjoint subsets of the partition, and compute the average log loss with that parameters. More than 10.000 iterations were computed.

At the end of the external loop, the set of parameters which allows the lower log loss is then used to test the model. Inside the loop, the node "table row to variable" uses the first row of the data table with log loss to define new flow variables with the parameters used in that specific iteration. In such a way, the node that ends the parametrization loop decides which partition results in the lower level of log loss and knows exactly the values of that parameters.

The result of the parameters optimization can be summarized as follows:

- **Lambda**=1,5
  It regulates leaf weights. Increasing it, the model will be more conservative.
- **Alpha**=1,5
  Same idea of Lambda, it regulates leaf weights.
- **Gamma**=0,5
  It is the minimum loss reduction required to make a further partition on a leaf node of the tree. So larger gamma, the more conservative the model will be.
- **Eta**=0,18
  It prevents overfitting. A smaller value results in a more conservative boosting process
- **Subsamples**=0,5
  Setting it to 0.5 means that XGBoost would randomly sample half of the training data prior to growing trees
- **Maximum depth**=3
  Maximum depth of a tree. Increasing it, the model will be more complex and more likely to overfit.

Table 1: Log loss comparison for different classification models

Model	LogLoss
XGboost	0,221
Multilayer perceptron	0,223
Logistic regression	0,225
Decision tree	0,240
Random forest	0,244
AODE	0,254
Probabilistic artificial networks	0,255
Native bayes	0,447

5 Feature importance and Shap

As stated in the section 4.5, XGboost Tree Ensemble Learner was used for its low value of log loss and for feature importance released as output. The table in A gives us

the measures for all the attributes. Each row is an attribute of the training set and the columns are [3]:

- **Weight**, the number of times a feature is used to split the data across all trees
- **Gain**, the average gain across all splits the feature is used in. A higher value here implies that attribute us more important for generating a prediction on diabetes
- **Cover**, the average coverage across all splits the feature is used in
- **Total gain**, the total gain sums up the gain across all splits the feature is used in
- **Total cover**, the total cover sums up the total coverage across all splits the feature is used in

BMI seems to be the attribute that has the major role in determining if a patient has or not diabetes with the highest value of weight and total cover (more details in A). Unfortunately, with feature selection it can be only said how each variable is important to define if a person is diabetic or not.

For this reason, the Shap was executed to determine if each attribute influences the prediction positively or negatively and how much. In Appendices B, the bar chart of the results from the shap loop are represented. The plot is only related to the presence of diabetes; for the absence it would be the same but symmetric. The length of the bars reproduces the degree of influence that each feature has. For instance, as shown in the table A and stated above, *BMI* has the major role with *hypertension* and *GenHlth*. While *hypertension* has a positive impact, *BMI* and *GenHlth* a negative one. If hypertension goes from 0 to 1, the probability of having diabetes increases. Patients with hypertension often exhibit insulin resistance and are at greater risk of diabetes developing than are normotensive individuals. The major cause of morbidity and mortality in diabetes is cardiovascular disease, which is exacerbated by hypertension [4]. Also *HighChol* has a positive influence: if a patient has an high level of cholesterol risks more to have diabetes.

6 Conclusion and Future Developments

To summarize, after a lot of work and analysis, the model chosen to best predict the presence of diabetes is the XGboost Tree Ensemble Learner with specific parameters set obtained with the parameter optimization.

For anyone wishing to resume and continue our analysis, we report some possible changes that could be added in the future. The value of logloss is good, but not perfect, we have still a lot of observations that are wrongly classified. To improve it, a possibile solution can be speak with a domain expert, for instance a doctor, to find new attributes which can add information to the actual dataset. Another option can be to find and merge different dataset thanks to data managment tools. We think that a great step would be studying the different incidence of diabetes concerning different locations. For instance, try to compare the same problem with the same attributes in USA and in Italy. These two nations have opposite cultures and food habits. It can have a huge impact and through that, it could be possible to identify important attributes.

References

[1] CDC, `https://www.cdc.gov/diabetes/basics/diabetes.html`

[2] BRFSS, `https://www.cdc.gov/brfss/index.html`

[3] KNIME, `https://www.knime.com/`

[4] National library of medicine, `https://pubmed.ncbi.nlm.nih.gov/29459239/`

Appendices

A

Feature importance

S Featur...	D Weight	D Gain	D Cover	D Total Gain	D Total C...
Hypertension	52	127.022	2,980.842	6,605.15	155,003.779
GenHlth=2	41	40.485	2,655.396	1,659.867	108,871.233
GenHlth=1	43	37.735	3,805.343	1,622.607	163,629.764
DiffWalk	48	34.956	2,266.058	1,677.864	108,770.78
HighChol	63	26.21	2,209.589	1,651.208	139,204.12
HeartDiseas...	29	12.633	2,273.854	366.355	65,941.76
GenHlth=4	32	11.133	3,008.136	356.257	96,260.366
BMI	243	10.884	2,160.333	2,644.828	524,961.037
CholCheck	33	10.199	3,910.418	336.563	129,043.795
GenHlth=3	10	9.506	2,483.697	95.06	24,836.967
Age=3	27	8.547	4,378.518	230.774	118,219.98
Age=4	23	7.277	4,046.19	167.367	93,062.374
Age=11	31	7.262	3,537.868	225.136	109,673.899
Age=5	22	7.084	4,395.986	155.849	96,711.7
Age=2	22	7.037	4,960.796	154.818	109,137.507
HvyAlcoholC...	36	6.933	3,333.287	249.575	119,998.315
Age=10	30	6.197	3,213.817	185.92	96,414.5
GenHlth=5	38	6.018	2,643.934	228.684	100,469.503
Age=12	31	5.093	3,440.346	157.869	106,650.719
PhysActivity	44	4.982	1,395.63	219.186	61,407.702
Age=1	18	4.836	4,763.161	87.041	85,736.901
Stroke	20	4.775	2,705.644	95.506	54,112.873
Sex	50	4.443	1,740.184	222.168	87,009.192
Age=6	15	4.397	4,715.413	65.959	70,731.196
Age=13	23	4.096	2,955.481	94.21	67,976.073
Veggies	23	2.983	2,682.501	68.607	61,697.522
PhysHlth	121	2.882	1,235.149	348.741	149,452.997
Age=9	17	2.656	3,027.818	45.145	51,472.913
Age=8	11	2.512	3,574.802	27.637	39,322.824
Fruits	27	2.396	2,021.318	64.691	54,575.577
MentHlth	80	2.361	2,369.273	188.879	189,541.801
Age=7	15	2.033	2,745.064	30.502	41,175.955
Smoker	33	1.882	1,317.181	62.116	43,466.988

Figure 3: Feature importance output of XGboost model

B

SHAP

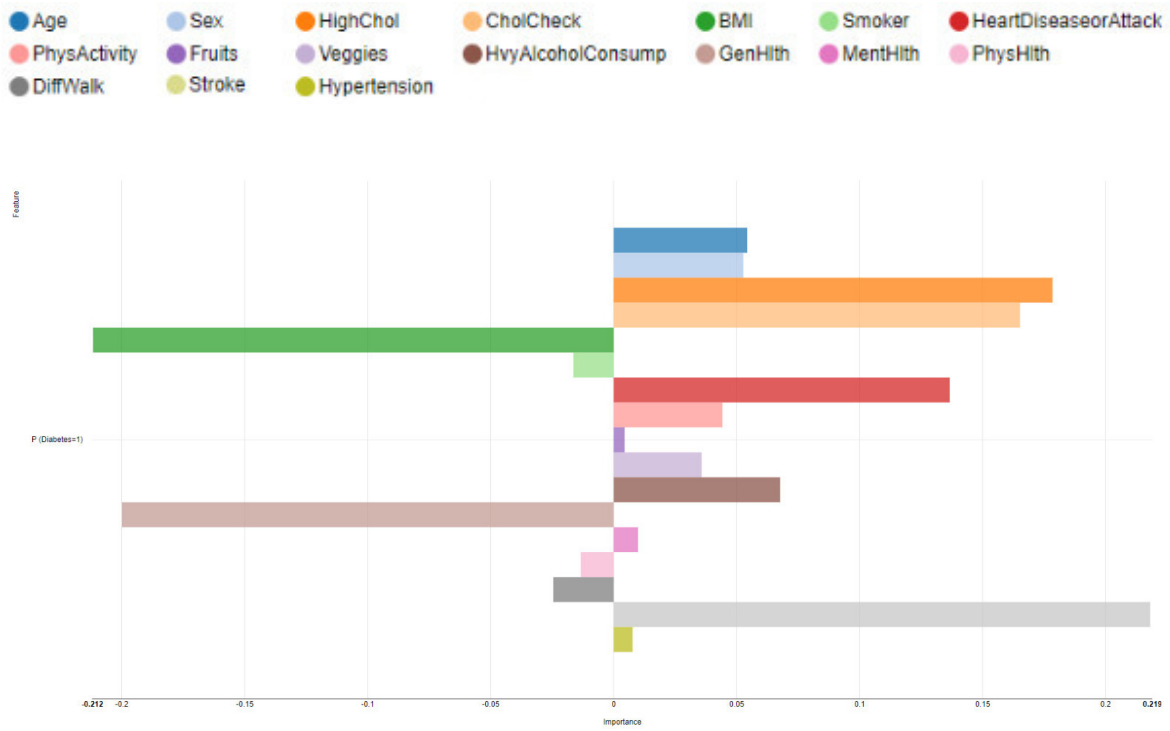

Figure 4: SHAP

C

ROC curve XGboost

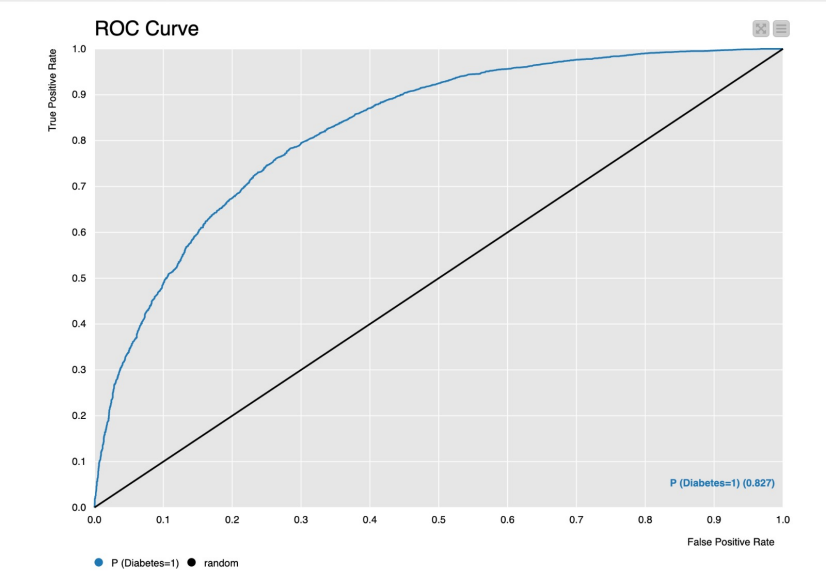

Figure 5: ROC curve XGboost