Recognition of factors indicating cases of diabetes
MC886 - Machine Learning and Pattern Recognition

Group 8
Fábio Santos Villar 234135
Guilherme Tezoli Bakaukas 217332
Juliana Bronqueti da Silva 238389
Lucca Gazotto Vettori 240231
Vitor Rodrigues Pietrobom 245584
Unicamp Computing Institute
Campinas, São Paulo

***Abstract*—Diabetes has become a serious public health problem, being considered by the WHO as an epidemic due to its alarming progression.The group's idea is to develop a model responsible for predicting, based on parameters of a patient's health, if the individual is not diabetic, is pre-diabetic or is diabetic, which could help facilitate early diagnosis and intervention and also reduce medical costs.**

1. Introduction

Diabetes has become a serious public health problem, being considered by the WHO as an epidemic due to its alarming progression.[1] In 2000, the global estimate of adults living with diabetes was 151 million. In 2009 this number had already grown to 285 million, an increase of 88%, and in 2020 the estimate is 463 million people.[2] But this does not only happen outside the country: Brazil is the 5th country in incidence of diabetes in the world, with 16.8 million adult patients (20 to 79 years old), with the estimated incidence of the disease in 2030 reaching 21.5 million.[3] The country is second only to China , India, United States and Pakistan.

Types of diabetes can be classified into:
- Type 1: caused by the destruction of insulin-producing cells, due to a defect in the immune system in which antibodies attack the cells that produce insulin. It occurs in about 5 to 10% of diabetics.
- Type 2: results from insulin resistance and deficiency in its secretion. It occurs in about 90% of diabetics.
- Gestational Diabetes: is the decrease in glucose tolerance, diagnosed for the first time during pregnancy, and may or may not persist after delivery. Its exact cause is not yet known.
- Other types: they result from genetic defects associated with other diseases or with the use of medications.

The incidence of diabetes varies by age, education, income, location, race, and other social determinants of health. Much of the burden of disease falls on those of lower socioeconomic status.

Correct treatment of diabetes means maintaining a healthy life. When not treated properly, blood sugar levels can stay elevated for a long time and cause damage to various organs, causing a chain reaction of health damage. As diabetes is associated with higher rates of hospitalizations, and the value of these hospitalizations cost more for this than for other diseases[4], it becomes even more necessary to invest in its prevention, as early diagnosis can lead to changes lifestyle and more effective treatment.

A lot of research using machine learning has been done to try to understand how different types of data can indicate diabetes. These surveys generally use clinical data, laboratory measurements and biomarkers, which makes the discovery, even if it is via machines, still expensive, since tests need to be carried out. Most studies have used logistic regression and Cox models[5][6] and the prediction model reported by Talmud[6] found, for example, a sensitivity/detection rate of 30% to 40% using research data and biomarkers.

The Behavioral Risk Factor Surveillance System (BRFSS) is a telephone health-related survey that is collected annually by the Centers for Disease Control and Prevention (CDC). Each year, the survey collects responses from more than 400,000 Americans about health-related risk behaviors, chronic health conditions, and use of preventive services. For this project, we used a csv from the dataset available in Kaggle for the year 2015.[7]

This dataset contains responses from 441,455 individuals and has 330 types of information. This information is either questions asked directly to the participants or variables calculated based on the individual responses of the participants. The listing with all the columns, what each one means and what are the possible associated values is in a codebook made available by the CDC.[8]

In order to understand if it is possible to help in the discovery of a possible case of pre-diabetes or diabetes without tests, only by analyzing some data on preventive health practices and risk behaviors related to chronic diseases, we propose this work.

2. Materials and methods

Our main goal is to accurately predict whether someone may have diabetes or is at high risk for diabetes. From this analysis we can also verify which risk factors are more predictive of diabetes risk.

For this we will use the BRFSS as a base.

Our main column of interest is “DIABETE3”, which has the answer to whether or not the individual has the disease, as well as what type it is. In this dataset it is not divided between types I and II, gestational and others, but whether you have diabetes or not, whether gestational or not, or have pre-diabetes. Therefore, we will not be able to define precisely which type of diabetes is between types I and II, only in the form of the data presented.

2.1. Methods

Through the aforementioned data source, we seek to filter your information and perform the proper data cleaning. Initially, we will group diabetes labels into just 3 different groups:

- 0 - subject without diabetes;
- 1 - individual with prediabetes;
- 2 - individual with diabetes.

Thus, it is expected that the model will be able to predict which of these conditions a given individual fits into. For this to be possible, it is essential that there is a good selection of the parameters that we consider relevant for this analysis, as they will be the basis for the construction of the model. Thus, it was decided to select the information from the original dataset present in Figure 1.

```
Index(['DIABETE3', 'BPHIGH4', '_RFCHOL', '_BMI5', '_SMOKER3', 'CVDSTRK3',
       '_MICHD', 'EXERANY2', '_FRTLT1', '_VEGLT1', 'HLTHPLN1', 'MEDCOST',
       'GENHLTH', 'MENTHLTH', 'PHYSHLTH', 'DIFFWALK', 'SEX', '_AGEG5YR',
       'EDUCA', 'INCOME2', 'FRUITJU1'],
      dtype='object')
```

Figure 1. BRFSS filtered columns.

It’s possible to access the information of each column in the codebook[8]. However, to make each parameter easier to recognize, they have been renamed as shown in figure 2.

```
Index(['DIABETE3', 'HIGH_BP', 'HIGH_CHOL', 'BMI', 'SMOKER', 'STROKE',
       'HEART_ATTCK_OR_DISEASE', 'PHYS_ACTIVITY', 'FRUITS', 'VEGG',
       'ANY_HEALTHCARE', 'MEDCOST', 'GEN_HEALTH', 'MENTAL_HEALTH',
       'PHYSICAL_HEALTH', 'DIFFWALK', 'SEX', 'AGE', 'EDUCATION', 'INCOME',
       'FRUIT_JUICE'],
      dtype='object')
```

Figure 2. Dataset columns renamed.

The columns present in the 2013 data matched with those in 2015, except for the information of ’_MICHD’ (Heart Attack or Coronary Disease). However, it’s possible to recreate it by combining the information ’CVDINFR4’ (Heart Attack) and ’CVDCRHD4’ (Coronary Disease) contained in 2013. As for the datatype, we know that, since they are all categorical, whether binary or with more classes, in which they are grouped into levels, thus, continuous data such as those in the “INCOME” column are discretized into predetermined value intervals. Therefore, with these parameters already defined, we can analyze the composition of the dataset. Let y be the column intended for prediction (DIABETE3) and x the remaining columns, it is important to highlight the distribution of the previously mentioned categories in y. For this goal, we created a plot with the amount of data from each label shown in Figure 2.

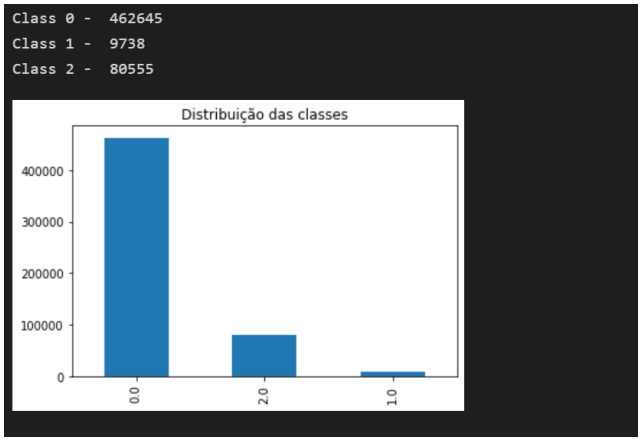

Figure 3. Simulation results for the network.

2.1.1. Balancing Techniques - Undersampling. As we can see, the dataset is very unbalanced, with a much higher amount of data 0 - ’individual without diabetes’. Thus, this could generate a tendency to overestimate the majority class, ignoring, in part, the information of the other classes. However, in order to prevent this problem, we can apply some techniques such as raising the error penalty in minority classes, if the model is suitable for this technique. We can also perform balancing techniques such as oversampling and undersampling. Although, these processes can generate a loss of relevant data (undersampling) or a greater possibility of overfitting, since replicated data (oversampling) is created, we believe that it is the best option for our context.

Therefore, to minimize these risks, we will use the SMOTE (Synthetic Minority Oversampling Technique) algorithm for oversampling, and NearMiss for undesampling. Thus, for the first one, we will have synthetic data closer to the real data, by determining positions between the nearest neighbors, avoiding the replicate data issue. Furthermore, for undersampling, we will use NearMiss version 3, which determines the selection of samples closest to the minority class samples. Then, it will be possible to keep the data of the different classes close together, seeking greater model accuracy.

With the techniques defined, let’s analyze a balance related to minority class 1 - ’Individuals with prediabetes’, by applying undersampling in both remaining classes. As

a result, we got a data distribution of 9738 samples for each class. As the amount of data has significantly reduced, we will also adopt a more comprehensive approach. By balancing the dataset based on class 2. This way we will keep a largest amount of data to insert into the machine learning model.

2.1.2. Models. Now, to select the best model to use, let's analyze the data correlation on our dataset. As we have parameters with discrete values, we will use the spearman method correlation to cover non-linear relationships between these categorical variables. Thus, we can analyze if there is any relevant correlation between the chosen columns and the diabetes information. In this context, we obtained the following result observed in figure 4.

	DIABETE3
DIABETE3	1.000000
HIGH_BP	0.221596
HIGH_CHOL	-0.165481
BMI	-0.194013
SMOKER	0.036232
STROKE	0.091495
HEART_ATTCK_OR_DISEASE	0.152479
PHYS_ACTIVITY	-0.102046
FRUITS	-0.027941
VEGG	-0.043212
ANY_HEALTHCARE	0.014923
MEDCOST	0.026117
GEN_HEALTH	-0.248187
MENTAL_HEALTH	-0.035777
PHYSICAL_HEALTH	-0.137281
DIFFWALK	0.187361
SEX	0.011433
AGE	-0.143890
EDUCATION	0.098669
INCOME	0.136155
FRUIT_JUICE	-0.047434

Figure 4. Diabetes spearman correlation on unbalanced dataset.

As observed, it was not possible to notice a very evident correlation between the chosen parameters. So, to build our model, we chose to discard a linear approach and use a neural network to find the best function that relates diabetes with the chosen variables.

In this matter, we decided to build a model with an input layer of 21 inputs, 2 hidden layers, with 60 neurons each and 3 elements on the output layer. In order to achieve classification between the 3 classes specified, this last layer must have softmax as activation function, which will return the probability, according to the model, that the set of inputs belongs to each class, then, the highest value will be considered as the forecast. On the other hand, for the hidden layers, we chose to use the reLU (Rectified Linear Unit) activation function, which is fast and efficient even for non-linear functions. Because the combination of activation functions of each neuron, obtained by its weights and bias, can generate results close to nonlinear functions.

However, to approximate the results predicted by the model in its training stage, it is essential to define a way to calculate the error. For this, we chose to use the Sparse Categorical Cross-Entropy that works well for classifications, as it returns an error with a logarithmic proportion of the differences between target and forecast. In addition, the optimizer used to define the new parameters of each neuron was Adam, which demonstrates a lot of efficiency to reach the ideal values for the model.

So, with the model built, we will normalize the data on the same scale from 0 to 1, using the MinMaxScalar function, using the minimum and maximum values of each parameter as 0-1 range. After that, we must divide the dataset into training and testing portions, in the proportion of 0.8/0.2 randomly. Thus, the data from 2013 and 2015 will be mixed and we should avoid biased results. Then, the next step is to fit the model and validate its precision.

2.1.3. Balancing Techniques - Oversampling. Using the neural network model, we did not obtain a good accuracy value, 59%. With that, we started to question whether the balancing method would be adequate, since the amount of data used for training was less than 9000 for each of the classes. In order to see how the original data were distributed among the classes, we used dimensionality reduction (Figure 5).

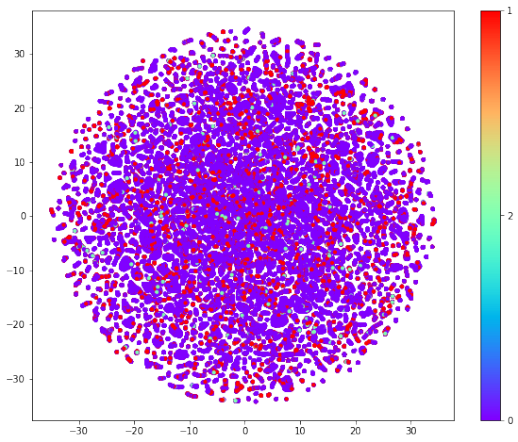

Figure 5. Original data dimensionality reduction

Knowing how the data is distributed we can create an oversampling/undersampling method that has a similar distribution. To do the opposite of what did not perform so well (undersampling), we decided to increase the amount of data from patients who had diabetes (type 2) and decrease the amount of data from patients who did not have diabetes (type 0), leaving the values in the same amount of individuals with pre-diabetes (type 1), totaling about 80 thousand data each. For undersampling we use NearMiss and for oversampling we use Smote.

Rearranging the data and reducing the dimensionality, we obtained Figure 6, which is similar to Figure 5 and does not have blocks with data of the same type, which indicates that this fit in the data was not biased.

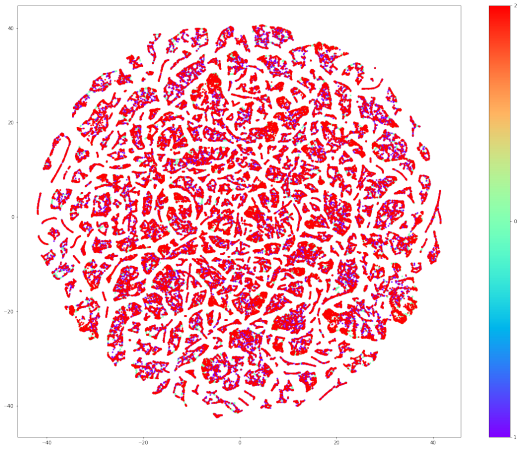

Figure 6. New data dimensionality reduction

With this work on the data we achieved a better accuracy than the method using only undersampling, so we decided to continue with this one to perform our analyzes that will be discussed in the results section.

3. Results and Discussion

As an initial result for the 3-class model, in which we balanced the data for the smallest class 1 size, we got the following table as classification report of Figure 7.

	precision	recall	f1-score	support
0.0	0.60	0.49	0.54	1991
1.0	0.78	0.58	0.66	1964
2.0	0.48	0.71	0.57	1888
accuracy			0.59	5843
macro avg	0.62	0.59	0.59	5843
weighted avg	0.62	0.59	0.59	5843

Figure 7. Classification report 3 classes under sampled

With the application of this balancing method, we achieved an accuracy of 59% in the test process. We imagine that this result was a consequence of using little data as input, since with the under sample of both classes 0 and 2, the amount of data dropped drastically, making it difficult to create a robust model with so much lost data as we can see when comparing Figure 3 with Figure 8.

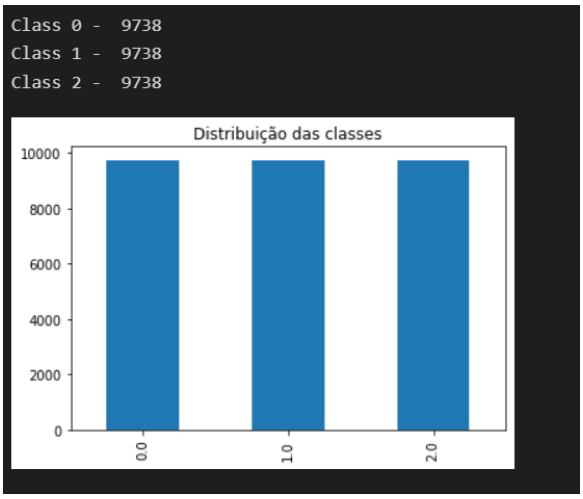

Figure 8. Data distribution after under sampling

To avoid this problem, we decided to balance the dataset by using the size of class 2 samples. Thus, we would maintain a higher amount of data, assuming the risks of a significant over sample on class 1. By applying this method, we were able to see a better accuracy of the model, on Figure 9.

	precision	recall	f1-score	support
0.0	0.68	0.75	0.71	16094
1.0	0.99	0.78	0.87	16204
2.0	0.64	0.71	0.68	16035
accuracy			0.75	48333
macro avg	0.77	0.75	0.75	48333
weighted avg	0.77	0.75	0.75	48333

Figure 9. Classification report 3 classes under and over sampled

Using a two-layer neural network, as well as undersampling data from subjects without diabetes and oversampling data from subjects with diabetes, we obtained 75% accuracy, which is pretty good. Just as with precision we ensure that our result, that is, the prediction, is more accurate, and with the recall we ensure that more classifications are correct, like ours that all cases of diabetes were identified by the model, the most important metric for the The analysis is precision, as the problem is sensitive to classifying a sample as Positive in general, that is, including Negative samples that were falsely classified as Positive.

With that we realized that our model is excellent at detecting prediabetes, with 99% precision, but not so good for without diabetes, 68% precision, and not so good for diabetes, 64% precision.

In addition to performing this analysis using the three classes, as our precision was not as accurate for individuals without diabetes and with diabetes, we replicated the model only now using only two classes: without diabetes (0), pre-diabetes and diabetes (1) . Our results for this new classification are shown in Figure 10.

	precision	recall	f1-score	support
0.0	0.72	0.76	0.74	18050
1.0	0.75	0.70	0.73	18068
accuracy			0.73	36118
macro avg	0.73	0.73	0.73	36118
weighted avg	0.73	0.73	0.73	36118

Figure 10. Classification report 3 classes under and over sampled

The first point noticed was that our accuracy dropped a little, from 75% to 73%. However, the accuracy for no diabetes increased slightly, from 68% to 72%. As the additions and decreases were few, we can conclude that this new categorization is not better for making the prediction, since it does not have much better values of accuracy and precision, nor greater granularity in the diagnosis, which is important to determine the referral. that the individual will have.

4. Conclusion

Since using the 3-class model gave us a low accuracy in the test process, it was necessary to find a way to avoid this problem, since a 59% accuracy was still low. The motives for that were the use of little data as input, mainly.

A better result was achieved by using the size of class 2 samples, which resulted in a higher accuracy of the model, of 75%, because of the increase in the amount of data.

The methods utilized to achieve better precision numbers were fundamental to our predictions, not only the mentioned above, but also undersampling data from subjects without diabetes and oversampling data from subjects with diabetes.

At the end, we realized that the model was excellent to predict prediabetes, but the accuracy to detect no diabetes and to detect diabetes could've been better.

References

[1] Conass. *OMS classifica o Diabetes como epidemia mundial*. URL: https://www.conass.org.br/oms-classifica-o-diabetes-como-epidemia-mundial. (accessed: 20.05.2022).

[2] BVS. *Dia Nacional do Diabetes*. URL: https://bvsms.saude.gov.br/26-6-dia-nacional-do-diabetes-4/#:~:text=O%20Brasil%20%C3%A9%20o%205%C2%BA,chega%20a%2021%2C5%20milh%C3%B5es.. (accessed: 20.05.2022).

[3] IDF. *IDF Diabetes Atlas*. URL: https://diabetesatlas.org/resources/. (accessed: 20.05.2022).

[4] FAPEG. *Internações por diabetes e complicações custam mais que outras doenças*. URL: http://www.fapeg.go.gov.br/internacoes-por-diabetes-e-complicacoes-custam-mais-que-outras-doencas/. (accessed: 20.05.2022).

[5] et al. Abbasi A Peelen LM. *Prediction models for risk of developing type 2 diabetes: systematic literature search and independent external validation study.* URL: https://www.bmj.com/content/345/bmj.e5900. (accessed: 10.07.2022).

[6] et al. ; UCLEB Consortium. Talmud PJ Cooper JA. *Sixty-five common genetic variants and prediction of type 2 diabetes.* URL: https://diabetesjournals.org/diabetes/article/64/5/1830/40514/Sixty-Five-Common-Genetic-Variants-and-Prediction. (accessed: 10.07.2022).

[7] CENTERS FOR DISEASE CONTROL and PREVENTION. *Behavioral Risk Factor Surveillance System*. URL: https://www.kaggle.com/datasets/cdc/behavioral-risk-factor-surveillance-system. (accessed: 20.05.2022).

[8] CDC. *Behavioral Risk Factor Surveillance System*. URL: https://www.cdc.gov/brfss/annual_data/2015/pdf/codebook15_llcp.pdf. (accessed: 20.05.2022).