Predicting Diabetes Disease in the Female Adult Population, Using Data Mining

Carolina Marques[1], Vasco Ramos[1], Hugo Peixoto[2(✉)], and José Machado[2]

[1] University of Minho, Campus Gualtar, Braga, Portugal
[2] Centro Algoritmi, University of Minho, Campus Gualtar, Braga, Portugal
hpeixoto@di.uminho.pt

Abstract. The aim of this study is to predict, through data mining, the incidence of diabetes disease in the Pima Female Adult Population. Diabetes is a chronic disease that occurs either when the pancreas does not produce enough insulin or when the body cannot effectively use the insulin it produces and is a major cause of blindness, kidney failure, heart attacks, stroke and lower limb amputation. The information collected from this population combined with the data mining techniques, may help to detect earlier the presence of this decease. To achieve the best possible ML model, this work uses the CRISP-DM methodology and compares the results of five ML models (Logistic Regression, Naive Bayes, Random Forest, Gradient Boosted Trees and k-NN) obtained from two different datasets (originated from two different data preparation strategies). The study shows that the most promising model as k-NN, which produced results of 90% of accuracy and also 90% of F1 Score, in the most realistic evaluation scenario.

Keywords: Data mining · Diabetes · CRISP-DM · Classification · ML models

1 Introduction

This paper has the purpose of diagnostically predict whether or not a patient has diabetes, considering certain variables, using Data Mining techniques with the help of RapidMiner[1].

Regarding its content, this article includes five sections. After the Introduction, the second section - Background and Related Work - presents a brief description of diabetes disease followed by previous studies and papers on this subject. The third section, Methodology, describes the applied CRISP-DM processes, which includes: business and data understanding, data preparation, modeling and evaluation. Section four presents the analysis and discussion of both

[1] https://rapidminer.com.

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2022
Published by Springer Nature Switzerland AG 2022. All Rights Reserved
S. Spinsante et al. (Eds.): HealthyIoT 2021, LNICST 432, pp. 63–73, 2022.
https://doi.org/10.1007/978-3-030-99197-5_6

the achieved results and the underlying work that was required to achieve those results. Finally, the last section addresses the conclusions that were possible to be drawn and and future work and improvements that could improve the present work.

2 Background and Related Work

2.1 Diabetes

Diabetes is a chronic disease that occurs either when the pancreas does not produce enough insulin or when the body cannot effectively use the insulin[2] it produces. Hyperglycaemia, or raised blood sugar, is a common effect of uncontrolled diabetes and over time leads to serious damage to many of the body's systems, especially the nerves and blood vessels.

About 422 million people worldwide have diabetes, the majority living in low-and middle-income countries, and 1.6 million deaths are directly attributed to diabetes each year. Both the number of cases and the prevalence of diabetes have been steadily increasing over the past few decades. Diabetes of all types can lead to complications in many parts of the body and can increase the overall risk of dying prematurely. Possible complications include kidney failure, leg amputation, vision loss and nerve damage. Adults with diabetes also have two- to three-fold increased risk of heart attacks and strokes. In pregnancy, poorly controlled diabetes increases the risk of fetal death and other complications, [1].

Early diagnosis can be accomplished through relatively inexpensive blood testing, [1].

2.2 Related Work

Data Mining is defined as the process of discovering patterns in data. The patterns discovered must be meaningful in order to lead to some advantage. Useful patterns allow us to make nontrivial predictions on new data, [2,3]. With computerized technology, content and structure started to change very fast in the health sector. Provided health services have to be fast, accurate, qualified and also have to meet the required needs. In order to achieve these goals, healthcare professionals need to have the most accurate and updated information and use this information as a relevant factor in their decision support systems, [4].

Effective use of healthcare data is made possible by data mining, which allows to extract relevant and valuable knowledge from large volumes of data. In healthcare, data mining is used to predict various diseases and to help doctors diagnose, [5–7].

Furthermore, the application of Data-Mining techniques which help to create a streamlined pipeline that includes all relevant phases of this type of work (data analysis, preparation, model development and application, results evaluation,

[2] Insulin is a hormone that regulates blood sugar.

and deployment) and allows for a simpler process of review, improvement and comparison, [8,9].

In [10], the study proposed to identify and classify the presence of diabetes diseases by applying data mining techniques. The dataset contained 520 instances, each having 17 attributes. Seven different classification algorithm including Bayes Network, Naïve Bayes, J48, Random Tree, Random Forest, k-NN and SVM were studied on the dataset. Obtained results indicated that k-NN performed the highest accuracy with 98.07%, being the best method to identify and classify diabetes diseases on the studied dataset. This work also presented interesting topics on how to clean and augment data, both in quantity and quality.

In [11], the study proposed a hybrid prediction model comprised of two different algorithms: the improved K-means algorithm and the logistic regression algorithm, which was based on data mining techniques for predicting type 2 *diabetes mellitus* (T2DM). The main problems trying to be solved were: improve the accuracy of the prediction model, and make the model adaptive to more than one dataset. Some previous studies have developed models with the same premise as the one for this paper, so the goal was to later compare the paper's results to those from this other papers. The dataset used on this paper was the same used on the present study (The Pima Indians Diabetes Dataset). The obtained results were compared to the results of the already mentioned previous studies and showed that the model attained a 3.04% higher accuracy of prediction (the accuracy was around 94%) than the ones used for comparison. Moreover, the model ensured that the dataset quality is sufficient. As a result, the model was shown to be useful in the realistic health management of diabetes.

3 Methodology

The data used in this work was originally provided by the **National Institute of Diabetes and Digestive and Kidney Diseases**, and has the purpose of diagnostically predict whether or not a patient has diabetes, based on some diagnostic measurements and medical indicators. The dataset is available at Pima Indians Diabetes Database - Kaggle.

Regarding the Data Mining process, this work will apply the Cross Industry Standard Process for Data Mining (CRISP-DM) Methodology, which is an hierarchical and iterative process model with six phases that naturally describes the data science life cycle. Those six phases are: **Business Understanding**, **Data Understanding**, **Data Preparation**, **Modeling**, **Evaluation** and **Deployment**, [12–14]. This process methodology was chosen due to its large set of advantages such as standardization of applied processes which makes the whole approach easily replicable, clear evaluation metrics and methods, clear structure of what to study and analyze which increases the changes of success and, finally, the possibility of applying Data Mining models in real scenarios, [15].

3.1 Business Understanding

The purpose of this study, as stated earlier, is to diagnostically predict whether or not a patient has diabetes, considering characteristics such as insulin level, plasma glucose concentration, blood pressure, skin thickness, among others. It is also relevant to point out that this study is focused on a very specific population: all patients are females, at least 21 years old, of Pima Indian heritage.

3.2 Data Understanding

The dataset used for this study, as stated earlier, consists of information regarding women of, at least, 21 years old from the Pima Indian community. It has 768 instances and each instance has 8 attributes and one more column with the respective class.

- **Pregnancies**: Number of times pregnant;
- **Glucose**: Plasma glucose concentration a 2 h in an oral glucose tolerance test;
- **Blood Pressure**: Diastolic blood pressure (mm/Hg);
- **Skin Thickness**: Triceps skin fold thickness (mm);
- **Insulin**: 2-Hour serum insulin (mu U/ml);
- **BMI**: Body mass index ($weight\_in\_kg/(height\_in\_m)^2$);
- **DPB (Diabetes Pedigree Function)**: Diabetes pedigree function;
- **Age**: Age (in years);
- **Outcome**: Class variable that specifies if tested positive for diabetes (0 or 1): 0 if YES, 1 if NO.

For a better understanding of each attribute, the Table 1 was created. It shows the number of missing values, the minimum and maximum value, the average and standard deviation of each attribute.

Table 1. Attribute description.

Attribute	Missing	Min	Max	Avg	Std. dev.
Pregnancies	0	0	17	3.8450	3.369
Glucose	5	0	199	120.894	31.973
Blood pressure	35	0	122	69.105	19.356
Skin thickness	227	0	99	20.536	15.952
Insulin	374	0	846	79.799	115.244
BMI	11	0	67.1	31.992	7.884
DPB	0	0.078	3.420	0.472	0.3315
Age	0	21	81	33.241	11.760

The class variable (outcome) has two different values: YES or NO. Figure 1 shows the data distribution regarding the outcome class that has 268 instances

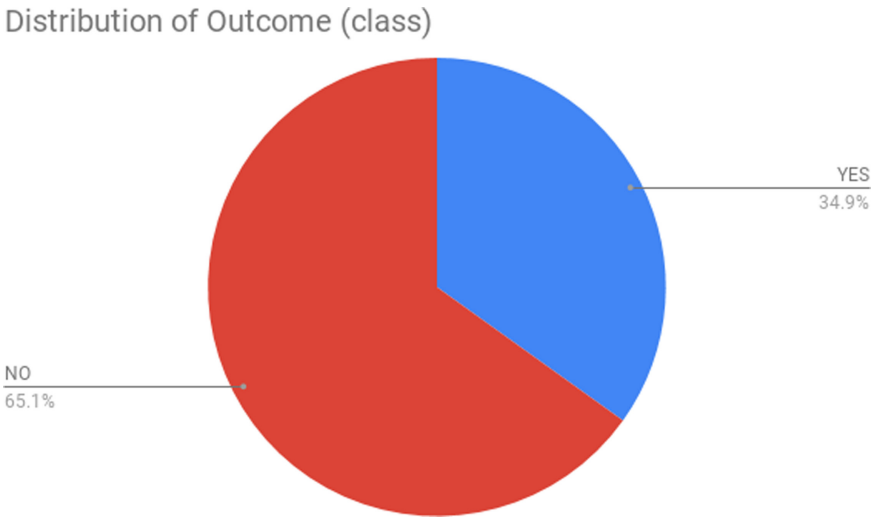

Fig. 1. Distribution of outcome (class)

of YES (34.9%) and 500 instances of NO (65.1%), which means that almost 35% of the cases indeed tested positive for diabetes and the remaining cases didn't.

The visualization of Fig. 1 clearly shows that the distribution of examples across the class Outcome is biased, having a big majority of the instances classified as NO. Having imbalanced or skewed data jeopardizes the ML models that will have poor predictive performance, specifically for the minority class.

3.3 Data Preparation

First Approach

Using the insight gained in the previous phase, the first step to clean the dataset was to map the missing values (that were described as zeros) to NaN values and remove all the instances that had missing values. After that, the dataset was normalized to values between 0 and 1 so that it was possible to find and remove the outliers more accurately. The next step is related to the fact that the dataset is skewed, so the dataset was oversampled using replication techniques over the existing data in order to balance the proportion of positive and negative instances for the diabetes test, increasing the number of instances associated with the presence of diabetes to approximate the number of negative examples. The last step was to shuffle the obtained dataset to ensure randomization in the process.

Second Approach

Being that the first approach was more focused on achieving high predicting results, the obtained dataset was not truly representative of the problem, therefore not reliable. Thus, it was decided to create an alternative approach more

focused on producing a credible and representative dataset, taking into consideration the possible deterioration of the results.

The first approach followed the strategy of removing all the instances that had missing values, whether this new methodology followed the strategy of removing the attribute with more missing values. In the dataset there were two main attributes that had missing values: *insulin*, with 374 instances (49% of all instances) and *skin thickness*, with 227 (29% of all instances). Removing both columns would reduce significantly the number of attributes (would only have 6) and the dataset itself, which would originate a poorer and more error-prone dataset. Hence, in order to create a compromise and keep the maximum number of instances while discarding a big majority of missing values, it was decided to remove only one of those two attributes: *insulin*, which, as already said, had the highest percentage of missing values.

From this point forward, the same steps of the first approach were followed, i.e., normalization, outlier identification and removal, oversampling to fix data imbalance and data shuffling.

As a final note, in both approaches the final dataset had roughly 1250 instances, with a similar distribution of the *outcome* class (the obtained datasets are balanced).

3.4 Modeling

This phase consisted in exploring and choosing the ML models to use, always bearing in mind that this is a problem of classification. Therefore, taking into account [16], it was decided to apply five different ML techniques: **Logistic Regression** (LR), **Naive Bayes** (NB), **Random Forest** (RF), **Gradient Boosted Trees** (GBT), with a learning rate of 0.05, and **k-NN**, with $K = 10$.

In addition to choosing the models to be applied, the strategy for training and evaluating each of the selected ML models was also specified.

First, the dataset is split in two sub-datasets: 70% of dataset is used to train the ML model and the remaining 30% is used to test the obtained model. To train the ML model, it's used the Cross Validation technique with 50 folds and, then, after the training is completed, the ML model is tested with the testing dataset, obtaining the evaluation metrics. The number of folds to use in Cross Validation was chosen after some experimentation, and the selected value was the one that gave better overall results, without any evidence of over-fitting or under-fitting.

Figure 2 shows the implementation of the specified strategy, on RapidMiner.

3.5 Evaluation

Being the problem in hands a classification one, it was decided not to only use accuracy as performance metric, because it becomes misleading. Instead, it was used the combination of **F1 Score**, **Accuracy**, **Precision** and **Recall** to compare models, [17].

To clarify these concepts:

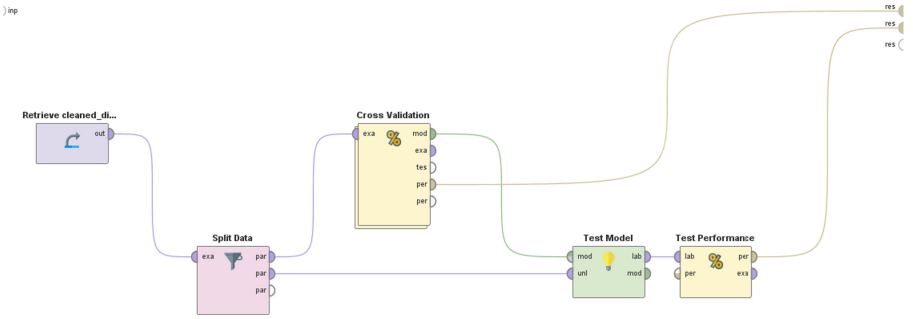

Fig. 2. Base modeling approach

- **Accuracy** - the ratio of correctly predicted examples to the total examples.
- **Precision** - the fraction of correctly classified positive examples from all classified as positive.
- **Recall** - actual positive rate of all positive examples, that is, the fraction of correctly classified examples.
- **F1 Score** - weighted average of Precision and Recall.

These concepts have mathematical representations, as follows:

- $Precision = \frac{TP}{TP+FP}$
- $Recall = \frac{TP}{TP+FN}$
- $F1Score = 2 * \frac{Recall*Precision}{Recall+Precision}$

Furthermore, the results presented next were calculated as the average of three different executions for each ML model.

(Data Preparation) First Approach

To begin with, the models were trained and tested with the dataset that resulted from the first data preparation approach. The Table 2 shows an overview of the results of each models for that dataset.

Table 2. ML model - testing results (1st approach)

ML model	Accuracy (%)	Precision (%)	Recall (%)	F1 score (%)
LR	78.55	79.35	77.63	78.38
NB	76.15	76.66	78.38	75.21
RF	93.51	94.09	93.62	93.76
GBT	96.60	96.62	96.61	96.55
k-NN	97.53	97.61	97.58	97.55

As it can be seen in Table 2, both the Logistic Regression and Naive Bayes model originated very low and unsatisfying results in every metric of evaluation, which is a clear indication that these models are not the most suitable to the prediction and classification at hand. The remaining three models: RF, GBT and k-NN produced far better results, being the k-NN model the one with highest overall performance with an F1 Score of almost 98%.

(Data Preparation) Second Approach
The next step was to train and test de ML models with the dataset from the second data preparation approach, given that the latter describes the problem more accurately. The Table 3 summarizes the obtained results for each ML model.

Table 3. ML model - testing results (2nd approach)

ML model	Accuracy (%)	Precision (%)	Recall (%)	F1 score (%)
LR	76.68	76.74	75.86	77.31
NB	73.32	73.32	73.33	73.51
RF	86.62	86.97	87.68	86.22
GBT	87.54	87.99	87.62	87.93
k-NN	89.94	90.35	90.08	90.17

As it was the case in the first dataset, the overall performance of the LR and NB was worst than the RF, GBT and k-NN models. Similarly, the k-NN model showed itself as the most suitable for the problem at hand, with an accuracy of 89%, precision and recall of 90% and, more importantly, with a F1 score of 90%.

4 Discussion

This section will address and discuss the overall strategies applied during the execution of the present work and, also, the achieved results.

In an early phase of data understanding it was clear that the used dataset was skewed, meaning that there was a bigger weight of instances on a class (in this case, the instances classified as NO), jeopardizing the applicability of ML models, which delivered a poorer predictive performance. Prior to the data preparation phase, there were some early experimentation with the already specified models which proved the previous thesis that the skewed data would produce weaker models and that it was an issue needed to be addressed.

Two datasets were produced with different approaches. The first approach was prepared in a way that the best results would be obtained when applying different models, but turned out to not be representative of the problem since the data was multiplied over and over again to balance the dataset and give the

best possible performance. This resulted in ML models theoretically good (with good metrics in both training and testing, which was due to dataset having too much repeated examples) that had no real applicability in the problem. Because it was considered inadequate, although with good results, a second strategy was developed where the results did not perform as well as in the first approach but are more fit to the problem and, therefore, was considered as the final solution.

When modeling, five different models were chosen to be applied: LR, NB, RF, GBT and k-NN and the dataset was split into 70% for training and the remaining 30% for testing. To train the model a Cross Validation technique was used and after that, the model was tested. This methodology allows the processes to be more consistent and coherent between the multiple tests and execution, which provides more certainty regarding the ability to compare results and reproduce the described conditions in future iterations.

When evaluating the first approach, it was noticed that the overall performance metrics were acceptable when using RF, GBT and k-NN, because they had an F1 Score over 90%. LR and NB were not considered suitable since the achieved values were around 77%. Since k-NN obtained F1 Score values of 97%, it was considered the most valuable model for this approach.

In the final approach, NB and LR presented the worst performance with 77% and 73% of F1 Score, respectively. Improvement was seen on the other three models, with performances over 86%. RF and GBT had similar results but GBT had an overall improvement, having a F1 Score of 86% and 87%, respectively. With a F1 score of 90%, the k-NN model proved itself as the most suitable to the given problem and dataset.

Finally, although the results were not as good as in [11], the work in that paper was found difficult to replicate, even with the same dataset, primarily due to the low detail on the applied data preparation techniques, which leads to an even bigger difficulty of comparing results.

5 Conclusions and Future Work

This work had the purpose to build a model capable of predicting whether a person, more specifically, a women from the Pima Indian community, has diabetes or not. Given the dataset, acceptable results were achieved with the k-NN model with the second approach of data preparation with an overall performance of 90% of Accuracy, Precision, Recall and F1 Score.

The biggest obstacles to attain a successful model were mainly related to the dataset: it had much more instances of non-diabetes cases than diabetes ones and also had a large amount of missing values.

The work that is more prone to future improvement is: try to use a better oversampling technique to generate synthetic data instead of replicate the existent data, try a different overall approach to deal with the missing values and, finally, try other ML models of unsupervised learning.

Acknowledgements. This work is funded by "FCT-Fundação para a Cincia e Tecnologia" within the R&D Units Project Scope: UIDB/00319/2020.

References

1. Organization, W.H: Diabetes - Fact Sheet. https://www.who.int/en/news-room/fact-sheets/detail/diabetes. Accessed 05 June 2021
2. Aljumah, A.A., Ahamad, M.G., Siddiqui, M.K.: Application of data mining: diabetes health care in young and old patients. J. King Saud Univ. Comput. Inf. Sci. **25**(2), 127–136 (2013). https://doi.org/10.1016/j.jksuci.2012.10.003, https://www.sciencedirect.com/science/article/pii/S1319157812000390
3. Witten, I.H., Frank, E., Hall, M.A.: Chapter 1 - what's it all about? In: Witten, I.H., Frank, E., Hall, M.A. (eds.) Data Mining: Practical Machine Learning Tools and Techniques, pp. 3–38. 3rd edn. The Morgan Kaufmann Series in Data Management Systems, Morgan Kaufmann, Boston (2011). https://doi.org/10.1016/B978-0-12-374856-0.00001-8, https://www.sciencedirect.com/science/article/pii/B9780123748560000018
4. Raghupathi, W., Raghupathi, V.: Big data analytics in healthcare: promise and potential. Health Inf. Sci. Syst. **2014** (2014)
5. Cruz, M., Esteves, M., Peixoto, H., Abelha, A., Machado, J.: Application of data mining for the prediction of prophylactic measures in patients at risk of deep vein thrombosis. In: Rocha, Á., Adeli, H., Reis, L.P., Costanzo, S. (eds.) New Knowledge in Information Systems and Technologies, pp. 557–567. Springer International Publishing, Cham (2019). https://doi.org/10.1007/978-3-030-16187-3_54
6. Konda, S., Rani, B., Govardhan, D.: Applications of data mining techniques in healthcare and prediction of heart attacks. Int. J. Comput. Sci. Eng. **2**, 250–255 (2010)
7. Peixoto, H., et al.: Predicting postoperative complications for gastric cancer patients using data mining. In: Cortez, P., Magalhães, L., Branco, P., Portela, C.F., Adão, T. (eds.) Intelligent Technologies for Interactive Entertainment, pp. 37–46. Springer International Publishing, Cham (2019). https://doi.org/10.1007/978-3-030-16447-8_4
8. Loreto, P., Peixoto, H., Abelha, A., Machado, J.: Predicting low birth weight babies through data mining. In: Rocha, Á., Adeli, H., Reis, L.P., Costanzo, S. (eds.) New Knowledge in Information Systems and Technologies, pp. 568–577. Springer International Publishing, Cham (2019). https://doi.org/10.1007/978-3-030-16187-3_55
9. Silva, C., Oliveira, D., Peixoto, H., Machado, J., Abelha, A.: Data mining for prediction of length of stay of cardiovascular accident inpatients. In: Alexandrov, D.A., Boukhanovsky, A.V., Chugunov, A.V., Kabanov, Y., Koltsova, O. (eds.) Digital Transformation and Global Society, pp. 516–527. Springer International Publishing, Cham (2018). https://doi.org/10.1007/978-3-030-02843-5_43
10. Alpan, K., İlgi, G.S.: Classification of diabetes dataset with data mining techniques by using weka approach. In: 2020 4th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), pp. 1–7 (2020). https://doi.org/10.1109/ISMSIT50672.2020.9254720
11. Wu, H., Yang, S., Huang, Z., He, J., Wang, X.: Type 2 diabetes mellitus prediction model based on data mining. Inf. Med. Unlock. **10**, 100–107 (2018). https://doi.org/10.1016/j.imu.2017.12.006, https://www.sciencedirect.com/science/article/pii/S2352914817301405

12. Portela, F., Santos, M.F., Machado, J., Abelha, A., Rua, F., Silva, Á.: Real-time decision support using data mining to predict blood pressure critical events in intensive medicine patients. In: Bravo, J., Hervás, R., Villarreal, V. (eds.) Ambient Intelligence for Health, pp. 77–90. Springer International Publishing, Cham (2015). https://doi.org/10.1007/978-3-319-26508-7_8
13. Guide, I.S.M.C.D: ftp://ftp.software.ibm.com/software/analytics/spss/documentation/modeler/14.2/en.CRISP_DM.pdf (2011)
14. Witten, I.H., Frank, E.: Data mining: practical machine learning tools and techniques with java implementations. ACM SIGMOD Rec. **31**(1), 76–77 (2002)
15. Wirth, R., Hipp, J.: Crisp-dm: towards a standard process model for data mining. In: Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining (2000)
16. Wu, X., Kumar, V., Quinlan, J.R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G.J., Ng, A., Liu, B., Philip, S.Y., et al.: Top 10 algorithms in data mining. Knowl. Inf. Syst. **14**(1), 1–37 (2008)
17. Hossin, M., Sulaiman, M.N.: A review on evaluation metrics for data classification evaluations. Int. J. Data Mining Knowl. Manage. Process **5**(2), 01–11 (2015). https://doi.org/10.5121/ijdkp.2015.5201