EAI Endorsed Transactions
on Smart Cities

Research Article EAI.EU

Predicting Diabetes Disease for healthy smart cities

Hugo Peixoto*, Vasco Ramos, Carolina Marques and José Machado

Algoritmi Center, University of Minho, Campus Gualtar, Braga 4710, Portugal

Abstract

INTRODUCTION: Diabetes is a chronic condition that affects a large portion of the population and is the leading cause of numerous health problems. Its automatic detection could improve the communities' overall well-being.
OBJECTIVES: The primary goal was to introduce advancements to the subject of healthy smart cities by studying an approach for predicting the occurrence of diabetes in the Pima Female Adult Population using data mining.
METHODS: This study uses CRISP-DM to analyze the results of six different models acquired from three different iterations of the same dataset.
DISCUSSION: This study found that the most promising model is k-NN, which obtained results of almost 92% of F1 Score with the third data preparation strategy.
CONCLUSION: Acceptable results were achieved with the k-NN model and the third data preparation strategy, but more research into improving the data preparation processes and their influence on the outputs of each model is needed.

Received on 15 March 2022; accepted on 21 April 2022; published on 22 April 2022

Keywords: Data Mining, Diabetes, CRISP-DM, Classification, ML Models, Smart Cities, Smart Health

Copyright © 2022 Hugo Peixoto *et al.*, licensed to EAI. This is an open access article distributed under the terms of the Creative Commons Attribution license, which permits unlimited use, distribution and reproduction in any medium so long as the original work is properly cited.

doi:10.4108/eetsc.v6i18.589

1. Introduction

Smart cities are designed with the goal of enhancing and providing a good quality of life for its residents by expanding urban infrastructure, encouraging innovation, and improving the healthcare system available to them, all while adhering to the values of equality, efficiency, and foresight [1, 2].

Smart health is a prominent area of the smart city concept. With the COVID-19 pandemics, the research and application of eHealth related topics have only increased in our society. A recent in-depth report on smart cities published by Deloitte [3], unveiled that key applications of smart health in smart cities vary from simplified processes for diagnosing diseases to streamlining their treatment, not to mention supporting well-being through early intervention and prevention using digital technologies. One of the key examples of this argument is a diabetes prevention program developed by the New York State Department of Health[1], which allows participants to use virtual monitoring and interact with life coaches and other participants.

As a result, it is evident that illness prevention and early and automatic detection, particularly of chronic diseases, is a critical application of academic research in smart cities with a massive impact on people's lives.

In this context, this paper is the evolution of a previously published work, in [4], to deepen the study previously conducted and explore improvement opportunities identified in that work. As such, building on what was already achieved, the main goal is to diversify the application of Data Mining techniques previously applied to identify the occurrence of diabetes illness on patients using a dataset from the Pima Female Adult Population, taking into consideration several factors.

This article is divided into five sections in terms of content. The second section, Background and Related Work, follows the Introduction with a brief summary

*Corresponding author. Tel: +351 253 604 430 ; Fax: +351 253 604 471 ; Email: hpeixoto@di.uminho.pt

[1] https://www.health.ny.gov/health_care/medicaid/redesign/ndpp/index.htm

of what is smart health and its role in the smart city concept, followed by an explanation on diabetic illness, and, lastly, earlier research and related work on the issue at hand. The third part, Methodology, discusses the CRISP-DM procedures in use, which include business and data comprehension, data preparation, modeling, and assessment. The analysis and discussion of both the attained outcomes and the underlying work that was necessary to reach those results are presented in section four. Finally, the last part discusses the conclusions as well as future work and improvements on the current work.

2. Background and Related Work

2.1. Smart Cities and Smart Health

Smart cities are a collection of technology techniques and mechanisms woven together by digital systems capable of collecting and analyzing massive amounts of data generated by a variety of sources, including low-cost sensors, mobile devices, and networks. All of this information is gathered and evaluated in order to develop intelligent and automated mechanisms that improve the overall quality of life for residents [5].

As such, the endeavor to transform current cities into smart ones can and will undoubtedly impact and disrupt different sectors, such as urban mobility, transportation, resources (water, power and heating energy, etc) and waste management, environment, politics and governance, economy and, of course, health and healthcare [6, 7].

Smart health, as its parent concept, can also be seen as a complexion of different concepts, such as mobile health and electronic health. These concepts include every system and application that focuses on improving and streamlining the complexity of clinical processes, and illness treatment. The area more specific to smart health is the preparation and analysis of all the generated health data and use that data to extract relevant knowledge that is crucial to develop intelligent and, most importantly, automated processes such as early illness diagnosis and identification.

This last effort of being able to achieve early, intelligent and automated illness diagnosis is being undertaken by a multitude of different research teams, on different illnesses, by the usage of machine learning and data mining techniques [8, 9].

2.2. Diabetes

Diabetes is a chronic disease that occurs when the pancreas fails to generate enough insulin[2] or when the body's own insulin is used ineffectively. Uncontrolled diabetes leads to hyperglycemia, or high blood sugar, which over time causes catastrophic damage to many body systems, including neurons and blood vessels.

It can cause problems in many areas of the body and increase the risk of early death in any form. Kidney failure, amputation of a limb, vision loss, and nerve damage are all possible outcomes. Adults with diabetes have a two to threefold increased risk of heart attacks and strokes. Furthermore, untreated diabetes during pregnancy raises the risk of fetal mortality and other complications. Finally, early detection is possible with relatively simple blood tests.

This disease affects approximately 422 million people worldwide, the majority of whom live in low- and middle-income countries, and causes 1.6 million deaths each year. Both the number of cases and the prevalence of diabetes have risen dramatically in recent decades [10].

2.3. Related Work

The practice of detecting patterns in data is known as data mining. In order to yield a benefit, the patterns detected must be relevant. Useful patterns allow us to make nontrivial predictions on new data [11, 12].

In the health industry, content and structure began to change at a rapid pace as a result of computerized technology. The health services that are provided must be quick, accurate, and qualified, as well as satisfy the requirements. To attain these objectives, healthcare professionals must have the most up-to-date and correct information, which they must employ as a meaningful aspect in their decision-making processes [13].

Data mining, which allows for the extraction of relevant and valuable knowledge from vast amounts of data, provides for effective use of healthcare data. Data mining is utilized in healthcare to assist doctors detect and forecast various ailments [14–16].

Furthermore, the use of Data-Mining techniques aids in the creation of a streamlined pipeline that includes all relevant phases of this type of work (data analysis, preparation, model development and application, results evaluation, and deployment) and simplifies the review, improvement, and comparison process [17, 18].

In [19], the study suggested to identify and categorize the existence of diabetes illnesses by using data mining approaches. There were 520 instances in the dataset, each with 17 characteristics. The dataset was used to test seven different classification algorithms, including Bayes Network, Nave Bayes, J48, Random Tree, Random Forest, k-NN, and SVM. According to the obtained results, k-NN had the maximum accuracy of 98.07%, making it the best approach for identifying and classifying diabetic illnesses on the examined

[2]Insulin is a hormone that helps to keep blood sugar levels in check.

dataset. This work also discussed methods to clean and supplement data, both in terms of amount and quality.

For predicting type 2 *diabetes mellitus* (T2DM), the study in [20] developed a hybrid prediction model consisting of two separate algorithms: the modified K-means algorithm and the logistic regression algorithm, both of which were based on data mining techniques. The key issues that needed to be addressed were improving the prediction model's accuracy and making the model adaptable to other datasets. Previous research have produced models based on the same assumption as the one used in this work, therefore the purpose was to compare the results of this study to those of the other studies. The dataset utilized in this investigation was the same as that used in the previous one (The Pima Indians Diabetes Dataset). When the acquired findings were compared to the results of the previously stated research, it was discovered that the model had a 3.04% greater prediction accuracy (approximately 94%) than the ones used for comparison. Furthermore, the model guaranteed that the dataset is of appropriate quality. As a result, the model has been demonstrated to be beneficial in the actual management of diabetic health.

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 Kaggle[3].

In respect to data mining, this project will use the cross industry standard process for Data Mining (CRISP-DM) Methodology, which is a six-phase hierarchical and iterative process model that accurately covers the data science life cycle. **Business Understanding**, **Data Understanding**, **Data Preparation**, **Modeling**, **Evaluation**, and **Deployment** are the six phases [21, 22].

This methodology was chosen because of its numerous benefits, including standardization of applied processes, which makes the entire approach easily replicable, clear evaluation metrics and methods, a clear structure of what to study and analyze, which increases the chances of success, and finally, the ability to apply Data Mining models in real-world scenarios [23].

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

As previously indicated, the dataset used in this study contains data on Pima Indian women aged 21 and above. It has 768 instances, each with eight characteristics and one extra column containing the respective class.

- **Pregnancies**: Number of times pregnant;
- **Glucose**: Plasma glucose concentration a 2 hours 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 $(weight\_in\_kg/(height\_in\_m)^2)$;
- **DPF**: 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.

To better understand each attribute, it was created the table 1. It displays the amount of missing values, as well as the lowest and maximum values, and the average and standard deviation for each property.

The class variable outcome has two possible values: YES or NO. Figure 1 depicts the data distribution for the outcome class, which contains 268 occurrences of YES (34.9%) and 500 instances of NO (65.1%), indicating that almost 35% of the patients tested positive for diabetes while the other cases did not.

Figure 1 clearly illustrates that the distribution of cases throughout the class Outcome is skewed, with a large majority of the occurrences classed as NO. The ML models will have poor prediction accuracy if the data is uneven or biased, particularly for the minority class.

Furthermore, a feature analysis was also conducted in order to understand which features have the greatest and least weight in determining whether or not the individual has diabetes, through a correlation matrix between the different attributes.

Figure 2 presents the obtained correlation matrix. As it can be seen, the least relevant features are **insulin** and **skin thickness**, as they present very low

[3] https://www.kaggle.com/uciml/pima-indians-diabetes-database

Table 1. Attribute description

Attribute	Missing Values	Min	Max	Avg	Std. Dev.
Pregnancies	0	0	17	3.9	3.4
Glucose	5	0	199	121	32.0
Blood Pressure	35	0	122	69.1	19.4
Skin Thickness	227	0	99	20.5	16.0
Insulin	374	0	846	79.8	115.2
BMI	11	0	67.1	32.0	7.9
DPF	0	0.1	3.4	0.5	0.3
Age	0	21	81	33.2	11.8

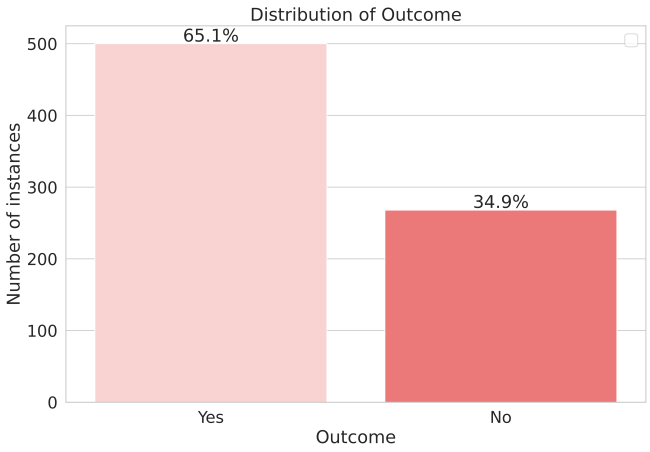

Figure 1. Distribution of Outcome (class)

correlation values for the outcome class compared to the other attributes and, in absolute values, almost null since their correlation factors are below 0.1 (0.066 and 0.09, respectively). On the other hand, it is possible to perceive that **BMI** and **glucose** are the two most important features for predicting the outcome class, with glucose being the most relevant since its correlation factor is the highest (0.48) and with a significant lead over BMI (0.31).

3.3. Data Preparation

First Strategy. In order to have a starting point with few modifications to understand the potential of the dataset in a nearly unmodified state, while also serving as a basis for comparing the results of subsequent strategies and corresponding improvements to the results, the first strategy of data preparation was defined as a two-step process.

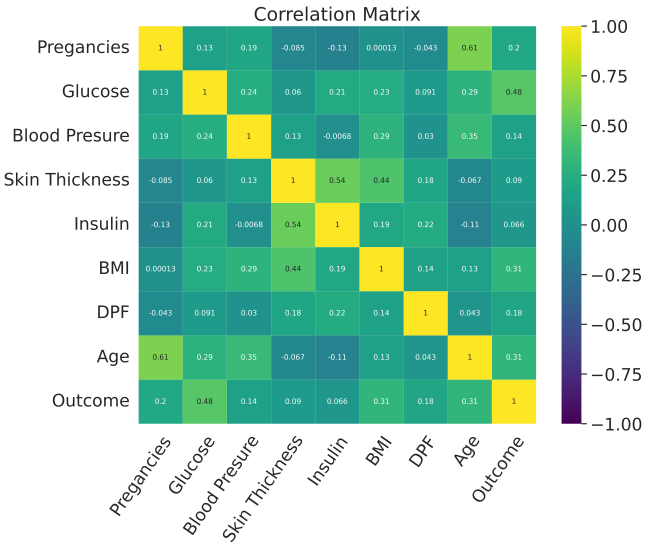

Figure 2. Correlation Matrix

First, the mapping of all missing values, which were described as zeros, to NaN values so that the models would not fail because of non expected data. Second, normalizing the data to values between 0 and 1. Other than that, no further transformation operations were performed, nor were any originally available attributes removed.

Second Strategy. Since the first strategy was simply to establish a baseline of comparison and evaluation, the second data preparation strategy involved adding further transformations to the dataset to improve its potential for success once it is used to train the ML models.

Thus, the previous steps of mapping the missing values and normalizing the data were retained and supplemented with the removal of all instances with missing values and the removal of identified outliers once the normalization process was complete. It was also taken into account the skewness of the dataset, which was addressed by applying replication techniques to the existing data to increase the number of instances associated with the presence of diabetes and to balance the proportion of positive and negative instances for the diabetes test.

Finally, the last step was to shuffle the obtained dataset to introduce randomization in the process.

Third Strategy. Since the second strategy was more focused on achieving high prediction results, the obtained dataset was not truly representative of the problem and therefore not reliable. Thus, it was decided to develop a third data preparation strategy more focused on producing a credible and representative dataset, taking into account the possible degradation of the results.

The second strategy took the approach of removing all instances with missing values, alternatively this strategy took the approach of removing only the attribute with the most missing values. In the dataset, there were two main attributes with missing values: *insulin* with 374 instances (49% of all instances) and *skin thickness* with 227 instances (29% of all instances). These were also the attributes identified as the least relevant features for predicting the occurrence of diabetes, which reinforced the option of removing at least one of them. Removing both columns would significantly reduce the number of attributes (there would be only six) and the dataset itself, resulting in a poorer and more error-prone dataset. Therefore, in order to compromise and maintain the maximum number of instances while discarding much of the missing values, it was decided to remove only one of these two attributes: *insulin*, which, as mentioned earlier, had the highest percentage of missing values and lowest correlation factor.

From this point on, the same steps as in the second strategy were followed, i.e., normalization, identification and removal of outliers, oversampling to correct the data imbalance, and data shuffling.

Finally, it should be noted that in both the second and third strategies, the final dataset comprised about 1250 instances, with a similar distribution of the *outcome* class (the obtained datasets are balanced).

3.4. Modeling

In this phase, the goal was to find and select the best machine learning models, keeping in mind that this is a classification problem. As a result, considering [24], six different machine learning techniques were selected: **Logistic Regression** (LR), **Naive Bayes** (NB), **Support Vector Machine** (SVM), **Random Forest** (RF), **Gradient Boosted Trees** (GBT), and **k-NN**, with $K = 10$.

In addition to the models to be used, the technique for training and evaluating each of the selected ML models was also selected.

To begin, the dataset was divided into two subsets: 70% of the dataset was used to train the machine learning model, while the remaining 30% was used to test the model. The Cross-Validation technique was used to train the machine learning model with 50 folds. Then, the model was evaluated with the testing dataset, yielding the evaluation metrics. After several tests, the number of folds for cross-validation was determined by choosing the value that provided the best overall results while meeting the requirement of not promoting over-fitting or under-fitting phenomena.

The implementation of the stated strategy, on RapidMiner, is shown in Figure 3. Figure 3 shows the implementation of the specified strategy, on RapidMiner.

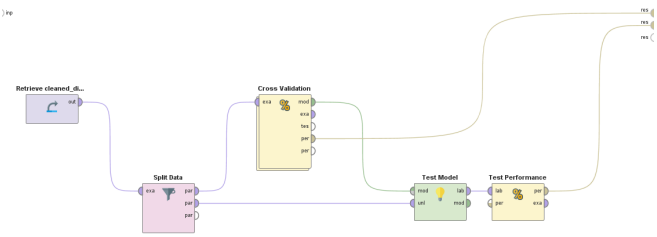

Figure 3. Base Modeling Approach

3.5. Evaluation

Being the problem at hand a classification one, it was decided not to use accuracy alone as the performance measure, since this would be misleading. Instead, the combination of **Accuracy**, **Precision**, **Recall**, and **F1 Score** was used to compare the models [25].

To clarify these concepts:

- **Accuracy** - the ratio of correctly predicted examples to the total examples.
- **Precision** - the ratio of correctly classified positive examples to all examples classified as positive.
- **Recall** - actual positive rate of all positive examples, i.e. the proportion of correctly classified examples.
- **F1 Score** - weighted average of Precision and Recall.

These concepts have mathematical representations, as follows:

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

In addition, the results presented below were calculated as the average of three different executions for each ML model.

First Strategy. The evaluation process began with the dataset that resulted from the first data preparation strategy. The table 2 summarizes the achieved results for each machine learning model.

As it is possible to perceive the dataset in its almost natural state does not produce good results, since every model fell short of reaching the minimum level of 75% in every metric, for the exception of the k-NN model that was able to achieve approximately 75% of accuracy and precision. Furthermore, all models achieved virtually the same results, proving that in most cases the quality of the dataset impacts directly the results, despite the selected model and that a poor dataset, both in quality and size, can undermine

Table 2. Testing results (first strategy)

ML Model	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
LR	72.10	74.02	69.35	71.61
NB	73.75	71.88	71.38	71.63
RF	73.58	71.07	69.34	70.17
SVM	71.89	73.05	69.38	70.18
GBT	69.21	70.55	66.63	68.53
k-NN	75.33	75.15	73.83	74.48

the entire data-mining process, meaning the data preparation phase is one of extremely importance to the whole process.

Second Strategy. The next step was to train and test the models with the dataset that resulted from the second data preparation strategy. The table 3 shows an overview of the results of each model for that dataset.

Table 3. Testing results (second strategy)

ML Model	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
LR	77.98	79.11	75.49	77.26
NB	77.31	77.85	78.57	78.21
RF	93.72	94.14	93.32	93.72
SVM	94.59	98.00	91.97	94.89
GBT	96.45	96.47	96.88	96.67
k-NN	98.06	97.26	98.01	97.71

The first thing clearly noticeable in the results presented in the table 3 is the big improvement of the results when compared to the results obtained with the first dataset, specially since the applied models are the same. With this strategy it was also possible to understand what models were more appropriate to the problem at hand, given some of them performed better than the other, which hadn't happening in the previous case. As it can be seen in table 3, both the Logistic Regression and Naive Bayes models were the ones that originated poorest 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 four models: RF, SVM, 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%.

Third Strategy. Finally, the models were trained and tested with the dataset from the third data preparation strategy, given that the latter describes the problem more accurately. The table 4 summarizes the obtained results for each model.

Table 4. Testing results (third strategy)

ML Model	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
LR	76.55	75.48	75.97	75.73
NB	73.46	73.23	73.61	73.42
RF	85.83	86.79	87.68	87.23
SVM	73.77	97.91	75.40	85.20
GBT	87.39	88.14	87.99	88.06
k-NN	91.14	92.02	91.36	91.69

As was the case with the second dataset, these are also much better results than those obtained with the dataset from the first data preparation strategy, despite being considerably worse than the results from the second strategy. These two observations are not without reason and were quite predictable. Since the second and third strategies built on the data preparation executed on the first strategy and added more transformations to further improve the quality of the data it is understandable the results from these two resulting datasets are better than the results from the dataset from the first strategy. On the other hand, the second data preparation strategy had its core focus on improving as much as possible the dataset to achieve high evaluation results, whereas the third strategy tried to create a compromise between a dataset that would produce good results but that would also faithfully represent the problem and context in question, something the second dataset fails considerably.

Analyzing the table 4, it can be seen that the overall performance of the LR, NB and SVM 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 91%, precision and recall of 91.4% and, more importantly, with a F1 score of 92%.

4. Discussion

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

The first thing worth pointing out is the dataset itself and the goal of this work, which is to contribute to smarter and efficient cities and societies with regard to its health by providing a proposal on a model to

automatically identify the occurrence of diabetes in a person. The dataset used is very focused and particular to females from the Pima population. However, the findings and advances presented in this paper can be extrapolated and replicated with other datasets that contain a broader diversity of population, reinforcing the relevance and applicability of what was developed in the course of this work.

Early on in the data analysis process, it became clear that the dataset was skewed, implying that there was a higher weight of instances on a class (in this case, the instances classified as NO), jeopardizing the applicability of machine learning models, which would result in poorer predictive performance. The initial data preparation method did not address this issue on purpose to establish the earlier assumption that skewed data would result in worse models and that it was a problem that needed to be addressed.

Three datasets were produced with different data preparation strategies and core goals. The first strategy aimed to transform the original dataset as little as possible to create a baseline of results to compare with and to highlight the problems of the dataset and their effect on the results achieved by the machine learning models. The second strategy was created so that the best results could be obtained when different models were used, however it turned out to be unrepresentative of the situation since the data was multiplied several times to balance the dataset and provide the greatest possible performance. As a result, models that were theoretically good (with strong metrics in both training and testing, due to the dataset containing too many repeated examples) but had no real application in the situation were created. Because it was deemed insufficient, while producing good results, a third strategy was devised with the goal of striking a balance between good outcomes and accurate portrayal of the problem. The models in this third strategy did not perform as well as those in the second strategy, but they were more suited to the problem, and, therefore, the third data preparation strategy it was chosen as the final approach.

When modeling, six distinct models were used: LR, NB, RF, SVM, GBT, and k-NN, with 70 percent of the dataset being used for training and the remaining 30 percent for testing. A cross-validation approach was employed to train the model, and then the model was tested. This technique ensures that procedures are consistent and coherent throughout various tests and executions, allowing for more confidence in the capacity to compare findings and recreate the specified circumstances in subsequent iterations.

When evaluating the first strategy, it was noticeably that all models performed poorly no matter how good or sophisticated the model was, meaning the dataset was the problem and that it needed further improvement and transformation in order to achieve good results.

In regard to the second strategy's evaluation, it was evident that the overall performance metrics were acceptable when utilizing RF, SVM, GBT, and k-NN, as they had an F1 Score superior to 90%. LR and NB were not considered suitable given the achieved values were around 77%. Since k-NN achieved F1 Score values of around 98%, it was thought to be the best model for this particular strategy.

In the final strategy, NB and LR presented the worst performance with 73% and 76% of F1 Score, respectively, followed by the SVM with an F1 Score of roughly 85%. RF and GBT had similar results but GBT had an overall improvement, having an F1 Score of 87% and 88%, respectively. With a F1 score of around 92%, the k-NN model proved itself as the most suitable to the given problem and dataset.

Finally, it is worth evaluating the whole research process through a SWOT analysis, i.e., the main strengths, weaknesses, opportunities and threats in this study.

Strengths. One of the major strengths of this work is the used methodology and approach that, as previously stated, provides consistency and coherence, meaning it is reliable and easy to replicate, as well as provides confidence in the achieved results.

Weaknesses. The dataset's low representation of the global population and the ability to predict false positives can be seen as drawbacks to the effective evaluation of this work.

Opportunities. This study has the potential to solve a global problem by predicting the occurrence of diabetes disease in each and every individual, which, if detected early, could save many lives. Although only the female Pima population was used in this dataset, it can be scaled to other women around the world. With the increase of developments in this field, this study can contribute to healthy smarter cities.

Threats. The most serious threat discovered was the unbalanced and poorly-representative data. This can lead to difficulties in fully describing the problem in a more global approach.

5. Conclusions and Future Work

The goal of this research was to contribute to the scientific progress of smart cities, particularly smart health, by developing a model capable of predicting whether or not a person, specifically a Pima Indian woman, has diabetes. Given the dataset, the k-NN model using the third strategy of data preparation produced satisfactory results while preserving an accurate depiction of the problem, with an overall

performance of 92% for Accuracy, Precision, Recall, and F1 Score.

The dataset presented the most significant challenges in developing a successful model: it had many more non-diabetic cases than diabetes patients, as well as a substantial number of missing variables. Furthermore, the results although based on a highly specific dataset are a substantial indicator of the reasonableness of a faster, smarter diabetes diagnostic with the help of data mining processes and techniques, which will positively contribute to earlier diagnosis, resulting in an improvement of the lives of the population affected by this disease.

Finally, the work more prone to future improvement is the research and experimentation of more complex and sophisticated oversampling techniques to produce synthetic data instead of replicating the existing data and the a thorough study on how the obtained k-NN model behaves with other datasets that represent the same problem, i.e., instead of focusing on a small group of people, diabetes occurrence is identified and classified across the entire population, or at least a more general population.

Acknowledgement. This work is funded by "FCT—Fundação para a Ciência e Tecnologia" within the R&D Units Project Scope: UIDB/00319/2020.
The grant of Vasco Ramos is supported by the project "Integrated and Innovative Solutions for the well-being of people in complex urban centers" within the Project Scope NORTE-01-0145-FEDER- 000086.

References

[1] Zota, R.D. and Clim, A. (2019) Smart healthcare for smart cities doi:10.13140/RG.2.2.26449.89446.

[2] Ghosh, R. and Kumar, S. (2020) Mobile health applications during epidemic management in India: a review. *EAI Endorsed Transactions on Smart Cities* **5**. doi:10.4108/eai.5-10-2020.166546.

[3] Antunes, M.E., Barroca, J.G. and de Oliveira, D.G. (2021) *Urban Future with a Purpose: 12 Trends Shaping the Future of Cities*. Tech. rep., Deloitte.

[4] Carolina Marques, Vasco Ramos, H.P. and Machado, J. (2021) Predicting diabetes disease in the female adult population. In *Proceedings of the 8th EAI International Conference on IoT Technologies for HealthCare* (Springer International Publishing). In Press.

[5] Rocha, N.P., Dias, A., Santinha, G., Rodrigues, M., Queirós, A. and Rodrigues, C. (2019) Smart cities and public health: A systematic review. *Procedia Computer Science* **164**: 516–523. doi:10.1016/j.procs.2019.12.214.

[6] (2017), Smart cities. URL https://ec.europa.eu/info/eu-regional-and-urban-development/topics/cities-and-urban-development/city-initiatives/smart-cities_en. Last accessed: 2022-03-07.

[7] What Is a Smart City? URL https://www.cisco.com/c/en/us/solutions/industries/smart-connected-communities/what-is-a-smart-city.html. Last accessed: 2022-03-07.

[8] Jain, R., Chotani, A. and Anuradha, G. (2021) Disease diagnosis using machine learning: A comparative study. In Lee, K.C., Roy, S.S., Samui, P. and Kumar, V. [eds.] *Data Analytics in Biomedical Engineering and Healthcare* (Academic Press), 145–161. doi:10.1016/B978-0-12-819314-3.00010-0.

[9] Singh, P., Singh, N., Singh, K.K. and Singh, A. (2021) Diagnosing of disease using machine learning. In Singh, K.K., Elhoseny, M., Singh, A. and Elngar, A.A. [eds.] *Machine Learning and the Internet of Medical Things in Healthcare* (Academic Press), 89–111. doi:10.1016/B978-0-12-821229-5.00003-3.

[10] Organization, W.H., Diabetes - Fact Sheet, https://www.who.int/en/news-room/fact-sheets/detail/diabetes. Last accessed: 2021-06-05.

[11] Aljumah, A.A., Ahamad, M.G. and Siddiqui, M.K. (2013) Application of data mining: Diabetes health care in young and old patients. *Journal of King Saud University - Computer and Information Sciences* **25**(2): 127–136. doi:10.1016/j.jksuci.2012.10.003, URL https://www.sciencedirect.com/science/article/pii/S1319157812000390.

[12] Witten, I.H., Frank, E. and Hall, M.A. (2011) Chapter 1 - what's it all about? In Witten, I.H., Frank, E. and Hall, M.A. [eds.] *Data Mining: Practical Machine Learning Tools and Techniques (Third Edition)*, The Morgan Kaufmann Series in Data Management Systems (Boston: Morgan Kaufmann): 3–38. doi:10.1016/B978-0-12-374856-0.00001-8, URL https://www.sciencedirect.com/science/article/pii/B9780123748560000018.

[13] W, R. and V, R. (2014) Big data analytics in healthcare: promise and potential. *Health Inf Sci Syst. 2014* .

[14] Cruz, M., Esteves, M., Peixoto, H., Abelha, A. and Machado, J. (2019) 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. and Costanzo, S. [eds.] *New Knowledge in Information Systems and Technologies* (Cham: Springer International Publishing): 557–567.

[15] Konda, S., Rani, B. and Govardhan, D. (2010) Applications of data mining techniques in healthcare and prediction of heart attacks. *International Journal on Computer Science and Engineering* **2**: 250–255.

[16] Peixoto, H., Francisco, A., Duarte, A., Esteves, M., Oliveira, S., Lopes, V., Abelha, A. *et al.* (2019) Predicting postoperative complications for gastric cancer patients using data mining. In Cortez, P., Magalhães, L., Branco, P., Portela, C.F. and Adão, T. [eds.] *Intelligent Technologies for Interactive Entertainment* (Cham: Springer International Publishing): 37–46.

[17] Loreto, P., Peixoto, H., Abelha, A. and Machado, J. (2019) Predicting low birth weight babies through data mining. In Rocha, Á., Adeli, H., Reis, L.P. and Costanzo, S. [eds.] *New Knowledge in Information Systems and Technologies* (Cham: Springer International Publishing): 568–577.

[18] Silva, C., Oliveira, D., Peixoto, H., Machado, J. and Abelha, A. (2018) Data mining for prediction of length of stay of cardiovascular accident inpatients.

In Alexandrov, D.A., Boukhanovsky, A.V., Chugunov, A.V., Kabanov, Y. and Koltsova, O. [eds.] *Digital Transformation and Global Society* (Cham: Springer International Publishing): 516–527.

[19] Alpan, K. and İlgi, G.S. (2020) Classification of diabetes dataset with data mining techniques by using weka approach. In *2020 4th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*: 1–7. doi:10.1109/ISMSIT50672.2020.9254720.

[20] Wu, H., Yang, S., Huang, Z., He, J. and Wang, X. (2018) Type 2 diabetes mellitus prediction model based on data mining. *Informatics in Medicine Unlocked* **10**: 100–107. doi:10.1016/j.imu.2017.12.006.

[21] Portela, F., Santos, M.F., Machado, J., Abelha, A., Rua, F. and Silva, Á. (2015) Real-time decision support using data mining to predict blood pressure critical events in intensive medicine patients. In Bravo, J., Hervás, R. and Villarreal, V. [eds.] *Ambient Intelligence for Health* (Cham: Springer International Publishing): 77–90.

[22] (2011) Ibm spss modeler crisp-dm guide URL https://inseaddataanalytics.github.io/INSEADAnalytics/CRISP_DM.pdf. Last accessed: 2022-03-02.

[23] Schröer, C., Kruse, F. and Gómez, J.M. (2021) A Systematic Literature Review on Applying CRISP-DM Process Model. *Procedia Computer Science* **181**: 526–534. doi:10.1016/j.procs.2021.01.199. CENTERIS 2020 - International Conference on ENTERprise Information Systems / ProjMAN 2020 - International Conference on Project MANagement / HCist 2020 - International Conference on Health and Social Care Information Systems and Technologies 2020, CENTERIS/ProjMAN/HCist 2020.

[24] Wu, X., Kumar, V., Quinlan, J.R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G.J. *et al.* (2008) Top 10 algorithms in data mining. *Knowledge and information systems* **14**(1): 1–37.

[25] M, H. and M.N, S. (2015) A review on evaluation metrics for data classification evaluations. *International Journal of Data Mining & Knowledge Management Process* **5**(2): 01–11. doi:10.5121/ijdkp.2015.5201.