IJDACR
ISSN: 2319-4863

IJDACR
International Journal Of Digital Application & Contemporary Research

International Journal of Digital Application & Contemporary Research
Website: www.ijdacr.com (Volume 8, Issue 12, July 2020)

Supervised Learning Methods for Predicting Diabetes: A Systematic Review of the Literature

Manoj Niwariya
Asst. Professor
Department of Computer Applications
Makhanlal Chaturvedi National University of Journalism and Communication (MCNUJC),
Bhopal, M.P. (India)
manoj_bec@yahoo.com

Dr. Anil Rajput
Professor
Department of Computer Science & Maths Science
Govt. C.S.A. PG College,
Sehore, M.P. (India)
dranilrajput@hotmail.com

Dr. Shailesh Jaloree
Professor and Head
Department of Applied Mathematics and Computer Science,
SATI, Vidisha, M.P. (India)

Abstract –**Artificial intelligence (AI) and its benefits in the field of medicine have generated a great revolution. For this reason, we want to identify the supervised learning methods (a sub-area of artificial intelligence) and the factors used for the prediction of diabetes that have been more significant in terms of technique (of which highlight decision tree and its derivatives) and results. For the identification of these methods, a systematic review of the literature was carried out. Machine learning methods were extracted from all the articles found to consider them as antecedents. There are several supervised learning methods that can predict diabetes in which some are hybrids and others pure, one better than others depending on the case study. Finally, after a review of the selected articles, the pre-processing stage in the development of these models is highlighted to achieve a higher precision score.**

Keywords – **Artificial Intelligence, Supervised Learning, Diabetes Prediction.**

I. Introduction

With the appearance of DENTRAL [1], the first expert system, whose objective was the inference of molecular structures, the enormous benefit of researching on how to improve existing methods or develop new methods of artificial intelligence was revealed. For this reason, they have been applied to the medical field with the aim of increasing success in medical cases [2] [3], such as the diagnosis of diseases or the detection of cancer cells. Due to this, many Machine Learning methods (supervised and unsupervised learning) have emerged for different clinical cases. This article will cover diabetes.

The use of artificial intelligence is closer than you might think, particularly in medicine, it is already used as a support for doctors when detecting life-threatening injuries on mammograms. In addition, we can mention the collaboration of AliveCor and Mayo Clinic, to develop a machine learning solution to determine the potassium levels in a person's blood through an electrocardiographic (ECG) signal from a smart watch. And the technology giants, Google AI and its DeepMind division, have carried out a work that allows the accurate evaluation of urgent eye conditions, the prediction of results in the hospital environment and a major prospective study of cancer pathology slides [4].

In 1980 the number of cases with diabetes was 108 million people and this increased to 422 million in 2014. Globally, cases with diabetes have prevailed in adults over 18 years of age, increasing from 4.7% in 1980 to 8, 5% in 2014. In 2015, approximately 1.6 million deaths were directly caused by diabetes. Another 2.2 million deaths were attributable to high blood glucose levels in 2012. All of these data are from the World Health Organization (WHO) [5].

There are three main types of diabetes: type 1 diabetes or insulin dependent, that is, the pancreas does not produce enough insulin to contribute to metabolism; the second, type 2 diabetes or diabetes mellitus or independent insulin, that is, the problem is the ineffective use of insulin; the third, gestational diabetes, that is, hyperglycemia during the pregnancy period and the possibility of having type 2 diabetes on the part of the children and the mother. [5]

On the other hand, the subjective use of one method over another can be counterproductive. In other words, replicating the method without considering the factors with which that precision was obtained with a method, which was indicated as the best for the tested case, can yield results that are relatively false.

The objective of this research is to know the most significant methods and factors for the prediction of diabetes and what tools are used to implement these solutions. This in turn will give a better idea when applying a supervised learning method to predict diabetes.

II. Literature Review

In this section some terms that are used later are conceptualized.

A. Artificial Intelligence

To define this term, one must first conceptualize intelligence, from the Latin "intelligence". It is understood that intelligence is not one but several intelligences, and that each person has to some degree more than one than the other, this theory proposed by Howard Gardner in 1983, defines 7 intelligences but others have been added until having 9 types of intelligences [3]: logical-mathematical intelligence, visual-spatial intelligence, verbal-linguistic intelligence, intrapersonal intelligence, interpersonal intelligence, bodily / kinesthetic intelligence, naturalistic intelligence, musical intelligence and existential or spiritual intelligence.

Due to the amplitude of the intelligences, it is difficult to measure them, even with the IQ test, which is only for logical-mathematical and visual-spatial intelligence. [3]

The best serious definition: intelligence is the ability to adapt, thus allowing them to solve the problems they encounter [3].

Then it will be said that artificial intelligence "is that the machine of the impression of being intelligent when solving a problem, for example imitating human behavior or implementing more flexible strategies than the ones allowed by classical programming" [3]. Here we find a certain notion of adaptability which is consistent with the previous definition of intelligence.

In addition, the observation of the mathematician, cryptographer, computer scientist and known as one of the fathers of computer science, Alan Turing, cannot be missed. For him, discussing the meaning of the words "think" and "intelligence" to conceptualize the capabilities of the computer seemed to him a nuisance and a waste of time. This is why I devised the Turing test, based on the imitation game, to check whether an entity was intelligent or not. The Turing test itself consists of asking the entity questions and, according to the answers it gives it, deciding whether it is intelligent or not; as "simple" as having a conversation to identify if you know the subject or not [6].

B. Models, Tasks and Methods of Artificial Intelligence

The knowledge extracted in the form of relationships, patterns or rules inferred from the data and (previously) unknown, or in the form of a more concise description (that is, a summary of them). These relationships or summaries constitute the model of the analyzed data. Models can be predictive and descriptive [7].

The tasks are the reason why you want to come up with artificial intelligence methods or techniques (decision tree, neural networks, Naive Bayes, fuzzy logic, etc.) [7]. For example: there is a need to predict the state of the climate; this would be the task, the weather forecast, and the method would be how the task will be accomplished.

In addition, it is considered that for "the same technique, different algorithms have been developed that differ in the form and concrete criteria with which the model is constructed." [7]

C. Machine Learning Algorithm

Machine Learning is an area of artificial intelligence as can be seen in Figure 1. When we speak of machine learning, "machine learning" refers to the ability of computers to recognize patterns or learn from them entered data [8]. And when referring to machine learning algorithms, it will be understood as input data, which represent certain experience, which will result in another contribution to the experience [8]. Generally the algorithm is implemented as the computer interpretable solution [3]. A Machine Learning model is implicit in Figure 1.

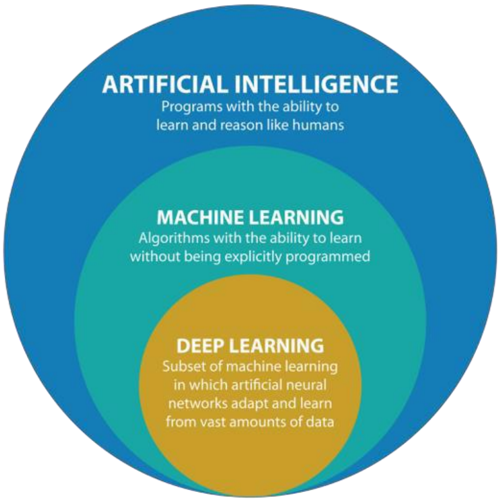
ARTIFICIAL INTELLIGENCE
Programs with the ability to learn and reason like humans
MACHINE LEARNING
Algorithms with the ability to learn without being explicitly programmed
DEEP LEARNING
Subset of machine learning in which artificial neural networks adapt and learn from vast amounts of data

Figure 1: Artificial intelligence, machine learning, and deep learning [3]

D. Supervised Learning

It is understood that process in which there are one or more expected outputs for input values, carried out in the training phase [8] [3]. The training data are object pairs, that is, one is the input data and the other is the expected result.

III. Method of Systematic Literature Review

A. Need for Systematic Review

Machine Learning solutions have been maturing over the years as can be seen in Fig. 2 [9] and Fig. 3 [10] on the trends in artificial intelligence solutions.

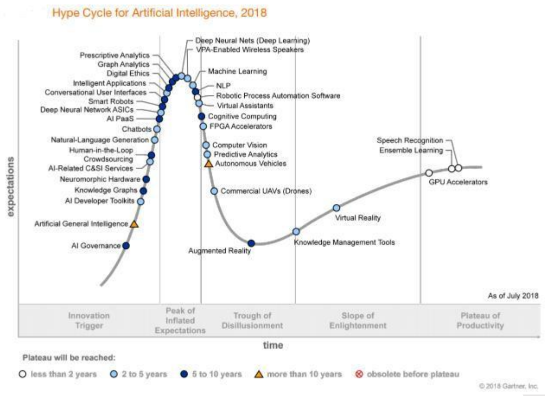

Fig. 2. Gartner's Hype Cycle for Artificial Intelligence, 2018 [9]

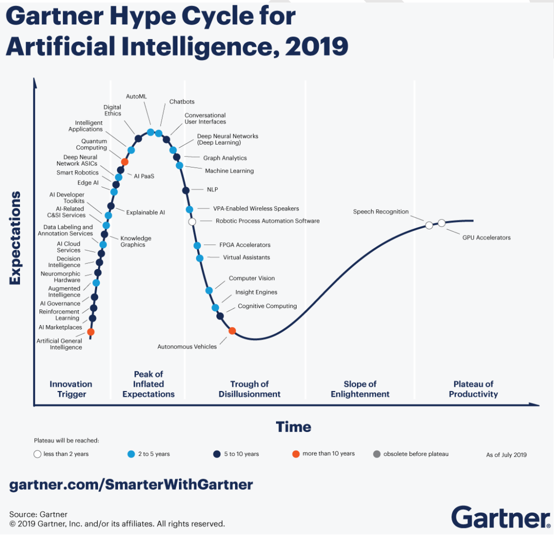

Fig. 3. Gartner's Hype cycle for artificial intelligence, 2019 [9]

Many of these solutions, such as Deep learning applications (see Figures 1 and 2), have supported the medical field.

On the other hand, the subjective use of one method or algorithm over another may be counterproductive. That is, replicating the method or algorithm without considering the factors with which that precision was obtained with a model, which was indicated as the best for the tested case, can yield results that are relatively false.

That is why a literature review is planned to determine which of the Machine Learning supervised learning methods are most significant to prediction, in this case diabetes. And it will be complemented by what is related to Machine Learning algorithms, tools and factors.

Finally, it is required to identify the tools that are used when developing and / or implementing the proposed Machine Learning solution.

B. Questions for Systematic Review

To define and structure the research questions, reference was made to the previous section. The following table (Table 1) presents the proposed questions and the motivation of each one.

Additionally, Table 2 presents the proposed frequently asked questions with the aim of visualizing the evolution and trend of the studies over time.

Table 1. Research and motivation questions

No.	Question	Motivation
1	What are the most significant supervised learning methods, models, and algorithms for predicting diabetes?	Identify the methods, models and algorithms used during the predictive diagnosis of diabetes
2	What are the most significant factors that were taken for the prediction of diabetes in the results?	To identify the factors by which the prediction of diseases used during the predictive diagnosis of diabetes benefited.
3	What are the tools to implement a supervised learning solution for diabetes prediction?	Identify the tools (hardware, software, and methodology) to implement a diabetes prediction solution.

Table 2. Frequently asked questions

No.	Question	Motivation
1	Determine what is the number of publications by type of article?	Specify the number of studies published by type of article to be able to identify their concentration.
2	How has the continuity of publications on the subject evolved over time?	Identify the continuity of the publications in order to establish the relevance of the topic over time.
3	In which publications have studies related to the topic been found?	Identify which application domain concentrates the largest number of publications on this topic

C. Definition of Search Strings

For the elaboration of the search chain, the PICO strategy was used through an iterative process in which adjustments were made for the selection of the outcomes. PICO will be broken down below.

Population:

Entity: supervised learning methods Main term 1: methods

Alternate terms: techniques

Proof: The term selection was due to the object of study of the review to be carried out and the alternative terms that represent close to the main term are determined [7].

Main term 1: Models

Verifier: The term is selected due to the relative popularity with which it refers to Machine Learning solutions.

Main term 2: Algorithms

Proof: The term is selected due to the particularity for the cases that are used. There is a certain difference between algorithm and method.

Main term 3: supervised learning

Alternate term: artificial intelligence

Proof: the term is selected due to the type of analysis to be carried out and these alternative terms are obtained as they are related to the main term that is to be found.

Intervention:

Entity: prediction of diabetes Main term 1: prediction Alternate terms: diagnosis

Proof: the term is selected due to the type of analysis to be performed and these alternative terms are obtained as they are related to the main term.

Entity: diabetes

Main term 1: diabetes

Alternate Terms: type 1 diabetes, type 2 diabetes, diabetes mellitus

Proof: the term is selected due to the type of analysis to be carried out and these alternative terms are obtained as they are the main types of diabetes.

Comparison: does not apply because no contrast is made in the RSL.

Outcomes:

Entity: Proposal and software prediction experience Main term 1: proposal

Alternate terms: experience

Proof: these terms are placed because it is what you want to obtain as an outcome of the search.

Language. English was chosen as the language for the search string due to its continuous usefulness for the elaboration of articles in prestigious databases.

Using the recommendations of the PICO strategy, the search string was obtained as a result from logical operators among the previously defined elements: (Population) AND (Intervention) AND (Comparison) AND (Outcome).

Table 3 shows the string obtained by each of the elements of the PICO strategy, from which the search string is elaborated.

Table 3. Terms and logical connectors to be used in the search

TERMS	CONCEPT
Population	(Method OR technical) AND model AND algorithm AND (supervised learning OR artificial intelligence)
Intervention	prediction AND (diabetes OR diabetes type 1 OR diabetes type 2 OR mellitus diabetes)
Comparison	Does not apply
Outcome	(Proposal OR experience)

D. Inclusion and Exclusion Criteria

Following the guide prepared by Kitchenham [11], after executing the search string in the indexed libraries, the results should be evaluated in order to determine the primary studies that directly answer the research questions asked. The following criteria were taken into account for the evaluation of the studies:

Inclusion Criteria:

1. All articles from digital libraries and indexed sources will be taken into consideration.
2. Articles containing studies of prediction methods or results of comparative analyzes of prediction methods will be accepted.
3. All articles within the defined temporal range are accepted.
4. Articles that come from scientific journals, journals, procedures and conferences are considered.
5. The articles must come from the area of artificial intelligence and related.

Exclusion Criteria:

1. Duplicate items will be excluded.
2. Articles that are not in English will be excluded.
3. Articles with similar content will be excluded, leaving only those with the most complete content.
4. Secondary studies and abstracts are rejected.
5. Articles whose title has no relation to the object of study are rejected.

Temporality: The studies carried out in the last 5 years are considered due to the fact that it is necessary to analyze diabetes prediction techniques and methods that are still in force. Furthermore, it is considered due to the increasing advance in Machine Learning (especially in supervised learning), after the 1970s (1970s). [1]

E. Quality Criteria

Following the guidelines established in the Kitchenham guide, the quality of the selected studies is evaluated [11] [12] [13]. A list of criteria is defined in order to check the compliance of each article. Each criterion is accompanied by a score based on the Rouhani scale, which is explained below: Yes it complies (S) = 1, it partially complies (P) = 0.5 and it does not fulfill (N) = 0.

IV. Results

A. Search Results

The first step in selecting studies consists of executing the search string in the selected digital libraries. Table 4 shows the results and the search strings used. For the IEEE Xplore database, an adjustment was made to improve the search.

Table 4. Search results

Database	Date	Total
Search String		
Science Direct	**May 2019**	**3694**
(method OR technical) AND model AND algorithm AND (supervised learning OR artificial intelligence) AND prediction AND (diabetes OR diabetes type 1 OR type 2 diabetes OR mellitus diabetes) AND (proposal OR experience)		
IEEE Xplore	**May 2019**	**145**
(method OR model OR algorithm) AND (supervised learning OR artificial intelligence) AND prediction AND (diabetes OR diabetes type 1 OR diabetes type 2 OR mellitus diabetes)		
ACM Digital Library	**May 2019**	**1633**
(method OR technical) AND model AND algorithm AND (supervised learning OR artificial intelligence) AND prediction AND (diabetes OR diabetes type 1 OR type 2 diabetes OR mellitus diabetes)		
Google Scholar	**May 2019**	**16900**
(method OR technical) AND model AND algorithm AND (supervised learning OR artificial intelligence) AND prediction AND (diabetes OR diabetes type 1 OR type 2 diabetes OR mellitus diabetes)		

B. Results of Applied Filters

Selection of primary studies.

Next, the details of the steps carried out for the selection of studies are presented:

- Step 1: The search string was executed and the language was limited to those written in English. The indexed bookstore that produced the most results was Google Scholar. Likewise, the inclusion and exclusion criteria presented for this step according to Inclusion and Exclusion Criteria were applied to said list.
- Step 2: On the list of results of Step 1, we proceeded with the exclusion and inclusion of the articles for the object of study according to what is defined in the criteria described in Inclusion and Exclusion Criteria.
- Step 3: The articles from Step 2 were excluded because they did not match the title defined in the described criteria and the inclusion criteria according to Inclusion and Exclusion Criteria were applied.
- Step 4: To proceed with the download of the articles, the content of the remaining articles was reviewed taking into account the summary, introduction and conclusions, and those that were not relevant according to what was defined in the criteria presented were excluded according to Inclusion and Exclusion Criteria.

Table 5 shows the results obtained from the selection of articles.

Table 5. Results of the study selection process

Database	Discovered Items	Step 1	Step 2	Step 3	Step 4
Science Direct	3694	2303	298	8	2
IEEE Xplore	145	145	103	47	21
ACM Digital Library	1633	193	128	4	2
Google scholar	16900	14300	8640	27	3
Total	**22372**	**16941**	**9169**	**86**	**28**

C. Review Summary

Table 6. Summary of the methods found in the articles selected at criteria

Article	Technique	Precision	Variables	Data source	Pre-processing
Art01 [14]	Naïve Bayes, Logistic regression	NB=0.653, LR= 0.661 NB=0.73, LR=0.735	- Waist-hip + TG in men - Rib-hip + TG in women	Korean Health and Genome Epidemiology Study database	Not applied
Art02 [15]	KNN	100%	All Variable Of PIMA Indian diabetes dataset	Pima Indian Diabetes -Data cleaning	Reduction of data
Art03 [16]	Logistic Regression (LR) and Naïve Bayes (NB)	0.741 And 0.739 in the women. 0.687 And 0.686 in the men.	- AGE, HEIGHT, WEIGHT, FC, NC, RFcR, ANcR, and WRcR (women) - AGE, HEIGHT, NC, AC, NFcR, RFcR, NHcR, RAcR, WAcR, WCcR, CHcR, HRcR, and HWcR (men)	Korean Health and Genome Epidemiology Study database	Synthetic Minority Over-sampling Technique (SMOTE)
Art04 [17]	Decision tree and J48	90.04%	All variable of PIMA Indian diabetes dataset	Pima Indian Diabetes Dataset	Replace missing and impossible values with the mean. We use K-means to remove incorrectly classified samples

Art05 [18]	Levenberg–Marquardt algorithm	0.71	All variable of PIMA Indian diabetes dataset	Pima Indian Diabetes Dataset	Not applied
Art06 [19]	Neuro-Fuzzy (ANFIS), Levenberg-Marquardt back-propagation algorithm	90.32% and 71.10%	Age, BMI, Diastolic blood pressure, Diabetes Pedigree, Glucose plasma concentration	Diabetes Database Bhubaneswar, Odisha, India	Not applied
Art07 [20]	- Random forest - Naive Bayes - ID3 algorithm -AdaBoost algorithm	85% 79.89% 78.57% 84.19%	Age, Height, Weight, Waist and Hip. Waist circumference (CC) or waist-hip (WHR) better discriminate diabetes cases among those without, compared to BMI.	University of Virginia dataset	All unrelated functions are removed, including total cholesterol, high-density lipoprotein, stabilized glucose, high-density lipoprotein, cholesterol / HDL ratio and systolic blood pressure first. Missing values have been removed. Rodent control
Art08 [21]	-J48 -Naïve Bayes -SVM with Poly Kernel -SVM with RBF kernel -Multilayer perceptron	AUC score 0.928 0.915) 0.942 0.827 0.911	- Age, sex, polydipsia, polyphagia, polyuria, family history, high blood pressure, diet, - Physical activity, blurred vision, smoking, weight loss, height and weight (BMI), waist circumference	Actual data collected from a renowned hospital in Chhattisgarh state of India	Missing values are replaced with the median
Art09 [22]	-Ensemble Perceptron Algorithm (EPA) -Perceptron (PA) algorithm	0.75 0.72	Age and BMI	National Health and Nutrition Examination Survey (NHANES) of United States	Not applied
Art10 [23]	-AdaBoost algorithm with decision stump -Support vector machine -Naive Bayes -Decision tree	80.72% 79.687% 79.687% 77.6%	All variable of PIMA Indian diabetes dataset. Triceps thickness of the skin fold (locally obtained indirectly). 2 hours insulin serum (obtained indirectly from local data) Body mass index (in the data for	Pima Indian Diabetes Dataset and local data	Missing values are replaced with corresponding mean value attributes in the global dataset

			validation it was obtained with height and weight). Function of pedigree diabetes (in the local data it is obtained indirectly)		
Art11 [24]	Extreme learning machine (ELM) and Back-Propagation	0.5964 and 0.0575	All variable of PIMA Indian diabetes dataset	Pima Indian Diabetes Dataset	Normalized to have a certain range of values
Art12 [25]	Back-Propagation, Decision Tree J48, Naive Bayes And Support Vector Machines	83.11 78.26 78.97 81.69	All variable of PIMA Indian diabetes dataset	Pima Indian Diabetes Dataset	Min-max normalization technique and Selection of characteristics with chi-square
Art13 [26]	Support Vector Machines, Naive Bayes (NB), K-nearest neighbors (KNN) and Decision tree C4.5 (DT)	0.65 0.685 0.708 0.72	Age, Sex, Weight, Diet, Polyuria, Water consumption. Excessive thirst, Blood pressure, Hypertension Tiredness, Vision problem, Kidney problem. Hearing loss, Skin itching, Genetics	Chittagong Medical Center (CMC), Bangladesh	The exact numerical value of the attributes is not significant to predict diabetes. As such, we convert the values of numerical attributes into nominal.
Art14 [27]	Random forest, Support Vector Machines and Logistic regression	0.89 0.825 0.844	SNPs, BMI and Age	Girona Biomedical Research Institute diabetes database	Most relevant SNPs
Art15 [28]	Feed-forward neural network using the window model	Root Mean Square Error=1.26	Monitored glucose levels	AIDA, mathematical simulator of diabetes	Not applied
Art16 [29]	Fuzzy decision tree based on GINI index	75.8%	All variable of PIMA Indian diabetes dataset	Pima Indian Diabetes Dataset	Records with lost data were deleted

The most used models in this review work were the decision tree [17] and its derivatives such as the random forest [27] and others ([17], [21]). This may be due to the estimation of the importance of the variables [30], the great capacity to handle big data and the stability in front of missing data.

From the articles presented in Table 5, considering that they did not pre-process, it is observed that [19] it reached 90.32 with a hybrid model, Neuro-Fuzzy (ANFIS). Furthermore, [22] I also present a score of 0.75 to differences from the other methodologies that did not pre-process. It could not be confirmed that one is better than the other due to the different variables used by the models, but this could give an indication.

In [31] a technique is described that jointly uses three decision tree algorithms (ID3, C4.5 and CART), with the aim of improving individual precision, which are combined by Bagging (selected from other methods of combination) which consists of voting by majority of the models generated for the classification. Stratified sampling was also used to solve the class imbalance. This hybrid technique was tested with two data sets, and gave good results for both data sets. Compared to the other bagging combination methods it gave the best results which were for the precision, sensitivity, specificity and f-measure 91.56%, 95.63%, 68.33% and 79.71%, respectively.

When the quality of the data set and a hybrid model are put together, it is possible to obtain what [32] obtained. A hybrid of SVM and Naive Bayes was used in which 97.6% was obtained, in which both algorithms have to coincide in their predictions, if they are different, greater monitoring will be given (as explained in the article). The dataset is proprietary obtained from Kosovo. After acquiring initial patient data, and after extensive laboratory examinations and continuous monitoring, as specified in the article.

The proposal of the least angle regression (LARS) with PCA described in [33] is interesting, because it implements data normalization and a selection of variables with PCA implicitly, all this gave an 89.53% area under the ROC curve (AUC), which is a ratio between sensitivity and specificity.

One of the factors is the variables used in the elaboration of the supervised learning model, it can be mentioned in [16] and [27], in which it is shown that the use of combinations of variables or measures increases the precision of the model. In this, the variables that contributed the most, through techniques such as those of the random forest [27] and those proposed by the SMOTE technique [16], were considered to the prediction task, from which they were introduced to the elaboration of the model.

The quality of the data set is essential to work with Machine Learning, that is why in [34] 87.5% of precision was obtained with the random forest model in which no data preparation technique was applied to the data set processing possibly due to the absence of any defect in the data set (the article does not mention any defect). This may be the case for [35].

Another factor is the inequality or imbalance of the data that will be used for the elaboration of the model, this means that the records used for the training have a predisposition to a particular class, and in other words, there is considerably more records of a particular class. This can be solved using the synthetic minority over-sampling technique (SMOTE) [13] [16].

Other evidence of SMOTE is in [36] where the Pima India Diabetes data set was used from which two data sets were derived. The first set was applied the elimination of missing values (missing value) and SMOTE for oversampling. A selection of complex characteristics (5 algorithms) was applied to the second data set. Tests were done on both sets of data, and the second set of data scored relatively better, that is, it did not improve at all at all compared to the first set of data. This shows that SMOTE can make a great contribution without much effort.

Furthermore, in [37] one more quality of SMTE is shown where the scores obtained by decision tree, probabilistic neural network (PNN) and Naïve Bayes were improved, showing improvements of 64%, 51% and 5% respectively, when SMOTE was applied. It can be mentioned that for the decision tree that obtained 0.215 (sensitivity), 0.992 (specificity) and 0.336 (f-measure), after applying SMOTE 0.726 (sensitivity), 0.802 (specificity) and 0.436 (f-measure) were obtained. The dataset is from Tehran Lipid and Glucose Study (TLGS).

In [38] it also uses an oversampling technique although it does not specify which one. With which random forest obtained an accuracy of 84% for the Pima Indians Diabetes data set.

Regarding article [15], it is shown that after applying three pre-processing techniques (data cleaning, sample reduction and PCA), PCA being the technique that I reduce from eight variables to two, which allowed the algorithm to k - nearest neighbors will reach a precision of 100% in the cross validation as in the conventional validations. It should be noted that after data cleaning and sample reduction the dataset using, PIMA Indians diabetes (see Table 10), decreased from 768 to 696 samples. To see the performance of the methodology proposed by the article, see table 12, in which D1 means that no pre-processing technique is applied to the dataset, D2 is the dataset after applying data cleaning and D3 means the dataset after applying sample reduction. In addition, AUC validation was used, which gave 1, which indicates 100%.

With respect to the article [17], it is observed that the application of pre-processing techniques such as replacing missing values with the mean values of the attribute and eliminating the badly classified samples by the K-means algorithm influenced the score achieved. However, the criterion of having eliminated by K-means some of the samples is doubtful, this would be contradicting the classification of the same data, and would presume that K-means would be sufficient to decide which sample was erroneously classified with what I would see in vain the elaboration of the J48 model proposed in said article.

Regarding the article [21] where a pre-processing technique is used to replace the missing values with the mean of the attribute to obtain the AUC score of 92.8%, unlike the other models tested by the article, of a set of data from 145 samples. With this it is understood that J48 is a good classifier even with little data.

Recital [28] in which a data set from the AIDA simulator was used; in which up to 40 case studies can be simulated with different age groups, diseases, and food intake; which reached the mean square error of 1.26 ml / d, being the average of the 10 cases tested in the article, it is possible that, because of the simulated data, no pre-processing was necessary.

When comparing the slight pre-processing that was done in [29], with [17], [23] and [25] (for example), it can be seen that these others achieved a higher score being the model and the origin of the same data set.

However, it is worth mentioning [39] where a chi-square test was performed to validate the dependence of the predictor variables used, although the results of this distribution were not applied, and a cleaning of the conjunct was also performed of data they had for their subsequent classification with a decision tree with which a score of 75% was obtained. It is possible that the reason for the low score is the deficiency in the cleanliness of the data or that a single pre-processing technique is not sufficient for the data set they have. With what can be said, that by applying a pre-processing technique does not ensure the optimization of the classification; you have to apply the necessary techniques and do it well.

It is possible that [40] and its 82.35% accuracy is misleading, because although Min Max Scaling has been used as a normalization technique (which transforms the values between 0 to 1, to have a normal distribution), the problem of imbalance presented by the PIMA Indians diabetes dataset I use. Only precision is specified as a metric but not specificity or sensitivity or others.

The use of MATLAB software to evaluate the effectiveness of the proposed model [41] and two of its proposals are programmed in the C ++ language, of the others there is no mention. Reference [42] mentions the large number of methods that the tool has and indicates personalization according to the requirements of the study as one of its greatest advantages.

For the application of web information systems, they used tensorflow.js for the implementation of the Machine Learning model [40].

To apply the study [25] they developed the Rstudio models with the programming language R. On the other hand, the experiments in [30] were carried out with the programming language R 3.2.1 and the packages for random forest were used and SVM.

In [16] the SPSS software is used to analyze the results. On the other hand, in [25] RStudio software with the R language is used to implement it.

A cloud-based tool from Microsoft is the Azure Machine Learning Studio (AMLS), which I use [43] to develop a decision tree model, which has a graphical user interface to build and run machine learning models. Work is made easy with a number of drag and drop functions on the interface. And the deployment can be done with a few clicks.

V. Conclusion

In order to collect the different sources, duly filtered, that are related to this research, the RSL was necessary, in addition to serving the PICO method for posing research questions and literature. It is worth mentioning that using a hybrid technique such as [41] that consists of electromagnetism-like mechanism algorithm (EM) using ROST-type opposite sign test (OST) combined with 1 NN, with a data set without missing values but with imbalance of data; it can cause a misleading value in precision. That is why the precision result is 73.03% while the kappa index is 0.0389 (with the kappa index being from 0 to 1, where 0.00 - 0.20 means very low agreement). Thus showing disagreement.

Another example is [12] that uses a hybrid model (SVM, ANN and Naïve Bayes) that obtained a score of 58.3% (specificity), 86.8% (sensitivity) and 77% (accuracy) with the Pima Indians Diabetes data set in a selection of the predictor variables was made with the MATLAB software (the article does not specify the procedure). This hybrid model could have had a higher score if the data imbalance had been considered, as was done in other articles.

One of the most significant supervised learning methods, according to this research, was the decision tree and its derivatives. To propose a supervised learning model for the prediction of diabetes, either pure or hybrid, it should be considered as factors above all the pre-processing and the predictive variables with which it will work, this involves having a little help from the subject expert to guide the techniques to apply. The precision lies not only in the model used but also in the variables with which it is worked, so it will be necessary to determine the most influential variables that increase the precision of the model.

Among the tools used for supervised learning, the one that stood out the most was MATLAB, due to the set of tools it has to work with machine learning and its easy use. Consider that the more data or cases the learning model has, the more accurate it will be. An opportunity for this research is the implementation of decision tree methods and their derivatives to the variables indicated [14] considering the pre-processing techniques required by the selected data set.

REFERENCES

[1] Lindsay, Robert K., Bruce G. Buchanan, Edward A. Feigenbaum, and Joshua Lederberg. "DENDRAL: a case study of the first expert system for scientific hypothesis formation." *Artificial intelligence* 61, no. 2 (1993): 209-261.

[2] Figge, Helen. "Deploying Artificial Intelligence against Infectious Diseases." *US PHARMACIST* 43, no. 3 (2018): 21-24.

[3] Rich, Charles, and Richard C. Waters, eds. *Readings in artificial intelligence and software engineering*. Morgan Kaufmann, 2014.

[4] Anahad O'Connor, "How Artificial Intelligence Could Transform Medicine – The New York Times," 2019. [Online]. Available: https://www.nytimes.com/2019/03/11/well/live/how-artificial-intelligence-could-transform-medicine.html. [Accessed: 01-March-2020].

[5] Mathers, Colin D., and Dejan Loncar. "Projections of global mortality and burden of disease from 2002 to 2030." *Plos med* 3, no. 11 (2006): e442.

[6] Barr, Avron, and Edward A. Feigenbaum, eds. *The handbook of artificial intelligence*. Vol. 2. Butterworth-Heinemann, 2014.

[7] Tan, Pang-Ning, Michael Steinbach, and Vipin Kumar. *Introduction to data mining*. Pearson Education India, 2016.

[8] Freeman, Ion, Ashley Haigler, Suzanna Schmeelk, Lisa Ellrodt, and Tonya Fields. "What are they Researching? Examining Industry-Based Doctoral Dissertation Research through the Lens of Machine Learning." In *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pp. 1338-1340. IEEE, 2018.

[9] Laurence Goasduff, "Artificial intelligence trends." [Online]. Available: https://www.gartner.com/smarterwithgartner/top-trends-on-the-gartner-hype-cycle-for-artificial-intelligence-2019/ [Accessed: 01-March-2020].

[10] Sicular, Svetlana, and Kenneth Brant. "Hype cycle for artificial intelligence, 2018." *Gartner (July 24, 2018)).< https://www. gartner. com/doc/3883863/hype-cycle-artificial-intelligence* (2018).

[11] Kitchenham, Barbara, O. Pearl Brereton, David Budgen, Mark Turner, John Bailey, and Stephen Linkman. "Systematic literature reviews in software engineering–a systematic literature review." *Information and software technology* 51, no. 1 (2009): 7-15.

[12] Li, Lin. "Diagnosis of diabetes using a weight-adjusted voting approach." In *2014 IEEE International Conference on Bioinformatics and Bioengineering*, pp. 320-324. IEEE, 2014.

[13] Nnamoko, Nonso Alex, Farath N. Arshad, David England, and Jiten Vora. "Meta-classification model for diabetes onset forecast: A proof of concept." In *2014 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 50-56. IEEE, 2014.

[14] Lee, Bum Ju, and Jong Yeol Kim. "Identification of type 2 diabetes risk factors using phenotypes consisting of anthropometry and triglycerides based on machine learning." *IEEE journal of biomedical and health informatics* 20, no. 1 (2015): 39-46.

[15] Panwar, Madhuri, Amit Acharyya, Rishad A. Shafik, and Dwaipayan Biswas. "K-nearest neighbor based methodology for accurate diagnosis of diabetes mellitus." In *2016 Sixth International Symposium on Embedded Computing and System Design (ISED)*, pp. 132-136. IEEE, 2016.

[16] Lee, Bum Ju, Boncho Ku, Jiho Nam, Duong Duc Pham, and Jong Yeol Kim. "Prediction of fasting plasma glucose status using anthropometric measures for diagnosing type 2 diabetes." *IEEE journal of biomedical and health informatics* 18, no. 2 (2013): 555-561.

[17] Chen, Wenqian, Shuyu Chen, Hancui Zhang, and Tianshu Wu. "A hybrid prediction model for type 2 diabetes using K-means and decision tree." In *2017 8th IEEE International Conference on Software Engineering and Service Science (ICSESS)*, pp. 386-390. IEEE, 2017.

[18] Saji, Sumi Alice, and K. Balachandran. "Performance analysis of training algorithms of multilayer perceptrons in diabetes prediction." In *2015 International Conference on Advances in Computer Engineering and Applications*, pp. 201-206. IEEE, 2015.

[19] Swain, Aparimita, Sachi Nandan Mohanty, and Ananta Chandra Das. "Comparative risk analysis on prediction of Diabetes Mellitus using machine learning approach." In *2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)*, pp. 3312-3317. IEEE, 2016.

[20] Xu, Weifeng, Jianxin Zhang, Qiang Zhang, and Xiaopeng Wei. "Risk prediction of type II diabetes based on random forest model." In *2017 Third International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB)*, pp. 382-386. IEEE, 2017.

[21] Sowjanya, K., Ayush Singhal, and Chaitali Choudhary. "MobDBTest: A machine learning based system for predicting diabetes risk using mobile devices." In *2015 IEEE International Advance Computing Conference (IACC)*, pp. 397-402. IEEE, 2015.

[22] Mirshahvalad, Roxana, and Nastaran Asadi Zanjani. "Diabetes prediction using ensemble perceptron algorithm." In *2017 9th International Conference on Computational Intelligence and Communication Networks (CICN)*, pp. 190-194. IEEE, 2017.

[23] Vijayan, V. Veena, and C. Anjali. "Prediction and diagnosis of diabetes mellitus—A machine learning approach." In *2015 IEEE Recent Advances in Intelligent Computational Systems (RAICS)*, pp. 122-127. IEEE, 2015.

[24] Pangaribuan, Jefri Junifer. "Diagnosis of diabetes mellitus using extreme learning machine." In *2014 International Conference on Information Technology Systems and Innovation (ICITSI)*, pp. 33-38. IEEE, 2014.

[25] Woldemichael, Fikirte Girma, and Sumitra Menaria. "Prediction of Diabetes Using Data Mining Techniques." In *2018 2nd International Conference on Trends in Electronics and Informatics (ICOEI)*, pp. 414-418. IEEE, 2018.

[26] Faruque, Md Faisal, and Iqbal H. Sarker. "Performance Analysis of Machine Learning Techniques to Predict Diabetes Mellitus." In *2019 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, pp. 1-4. IEEE, 2019.

[27] López, Beatriz, Ferran Torrent-Fontbona, Ramón Viñas, and José Manuel Fernández-Real. "Single Nucleotide Polymorphism relevance learning with Random Forests for Type 2 diabetes risk prediction." *Artificial intelligence in medicine* 85 (2018): 43-49.

[28] Asad, Muhammad, Usman Qamar, Babar Zeb, Aimal Khan, and Younas Khan. "Blood glucose level prediction with minimal inputs using feedforward neural network for diabetic type 1 patients." In *Proceedings of the 2019 11th International Conference on Machine Learning and Computing*, pp. 182-185. 2019.

[29] Varma, Kamadi VSRP, Allam Appa Rao, T. Sita Maha Lakshmi, and PV Nageswara Rao. "A computational intelligence approach for a better diagnosis of diabetic patients." *Computers & Electrical Engineering* 40, no. 5 (2014): 1758-1765.

[30] Xao, Wenxiang, Fengjing Shao, Jun Ji, Rencheng Sun, and Chunxiao Xing. "Fasting blood glucose change prediction model based on medical examination data and data mining techniques." In *2015 IEEE International Conference on Smart City/SocialCom/SustainCom (SmartCity)*, pp. 742-747. IEEE, 2015.

[31] Bashir, Saba, Usman Qamar, Farhan Hassan Khan, and M. Younus Javed. "An efficient rule-based classification of Diabetes using ID3, C4. 5, & CART ensembles." In *2014 12th International Conference on Frontiers of Information Technology*, pp. 226-231. IEEE, 2014.

[32] Tafa, Zhilbert, Nerxhivane Pervetica, and Bertran Karahoda. "An intelligent system for diabetes prediction." In *2015 4th Mediterranean Conference on Embedded Computing (MECO)*, pp. 378-382. IEEE, 2015.

[33] Qiu, Shaoming, Jiahao Li, Bo Chen, Ping Wang, and Xiue Gao. "An improved prediction method for diabetes based on a feature-based least angle regression algorithm." In *Proceedings of the 3rd International Conference on Machine Learning and Soft Computing*, pp. 232-238. 2019.

[34] Rallapalli, Sreekanth, and T. Suryakanthi. "Predicting the risk of diabetes in big data electronic health Records by using scalable random forest classification algorithm." In *2016 International Conference on Advances in Computing and Communication Engineering (ICACCE)*, pp. 281-284. IEEE, 2016.

[35] Sarwar, Abid, and Vinod Sharma. "Comparative analysis of machine learning techniques in prognosis of type II diabetes." *AI & society* 29, no. 1 (2014): 123-129.

[36] Nnamoko, Nonso, Abir Hussain, and David England. "Predicting Diabetes Onset: An Ensemble Supervised Learning Approach." In *2018 IEEE Congress on Evolutionary Computation (CEC)*, pp. 1-7. IEEE, 2018.

[37] Ramezankhani, Azra, Omid Pournik, Jamal Shahrabi, Fereidoun Azizi, Farzad Hadaegh, and Davood Khalili. "The impact of oversampling with SMOTE on the performance of 3 classifiers in prediction of type 2 diabetes." *Medical decision making* 36, no. 1 (2016): 137-144.

[38] Dutta, Debadri, Debpriyo Paul, and Parthajeet Ghosh. "Analysing Feature Importances for Diabetes Prediction using Machine Learning." In *2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, pp. 924-928. IEEE, 2018.

[39] Anand, Ayush, and Divya Shakti. "Prediction of diabetes based on personal lifestyle indicators." In *2015 1st International Conference on Next Generation Computing Technologies (NGCT)*, pp. 673-676. IEEE, 2015.

[40] Dey, Samrat Kumar, Ashraf Hossain, and Md Mahbubur Rahman. "Implementation of a web application to predict diabetes disease: an approach using machine learning algorithm." In *2018 21st international conference of computer and information technology (ICCIT)*, pp. 1-5. IEEE, 2018.

[41] Wang, Kung-Jeng, Angelia Melani Adrian, Kun-Huang Chen, and Kung-Min Wang. "An improved electromagnetism-like mechanism algorithm and its application to the prediction of diabetes mellitus." *Journal of biomedical informatics* 54 (2015): 220-229.

[42] Sisodia, Deepti, and Dilip Singh Sisodia. "Prediction of diabetes using classification algorithms." *Procedia computer science* 132 (2018): 1578-1585.

[43] Srivastava, Yashi, Pooja Khanna, and Sachin Kumar. "Estimation of Gestational Diabetes Mellitus using Azure AI Services." In *2019 Amity International Conference on Artificial Intelligence (AICAI)*, pp. 321-326. IEEE, 2019.