







## Article

# Identification of People with Diabetes Treatment through Lipids Profile Using Machine Learning Algorithms

Vanessa Alcalá-Rmz <sup>1,†</sup>, Carlos E. Galván-Tejada <sup>1,\*,†</sup>, Alejandra García-Hernández <sup>1</sup>,  
Adan Valladares-Salgado <sup>2</sup>, Miguel Cruz <sup>2</sup>, Jorge I. Galván-Tejada <sup>1</sup>, Jose M. Celaya-Padilla <sup>1</sup>,  
Huizilopoztli Luna-García <sup>1</sup> and Hamurabi Gamboa-Rosales <sup>1</sup>

- <sup>1</sup> Unidad Académica de Ingeniería Eléctrica, Universidad Autónoma de Zacatecas, Jardín Juárez 147, Centro, Zacatecas 98000, Mexico; vdrar.06@uaz.edu.mx (V.A.-R.); alegarcia@uaz.edu.mx (A.G.-H.); gatejo@uaz.edu.mx (J.I.G.-T.); jose.celaya@uaz.edu.mx (J.M.C.-P.); hlugar@uaz.edu.mx (H.L.-G.); hamurabigr@uaz.edu.mx (H.G.-R.)
- <sup>2</sup> Unidad de Investigación Médica en Bioquímica, Hospital de Especialidades, Centro Médico Nacional Siglo XXI, Instituto Mexicano del Seguro Social, Av. Cuauhtémoc 330, Col. Doctores, Del. Cuauhtémoc, Mexico City 06720, Mexico; adan.valladares@imss.gob.mx (A.V.-S.); miguel.cruzlo@imss.gob.mx (M.C.)
- \* Correspondence: ericgalvan@uaz.edu.mx; Tel.: +52-492-544-0968
- † These authors contributed equally to this work.



**Citation:** Alcalá-Rmz, V.; Galván-Tejada, C.E.; García-Hernández, A.; Valladares-Salgado, A.; Cruz, M.; Galván-Tejada, J.I.; Celaya-Padilla, J.M.; Luna-García, H.; Gamboa-Rosales, H. Identification of People with Diabetes Treatment through Lipids Profile Using Machine Learning Algorithms. *Healthcare* **2021**, *9*, 422. <https://doi.org/10.3390/healthcare9040422>

Academic Editor: Saleh A. Naser

Received: 31 December 2020

Accepted: 8 March 2021

Published: 6 April 2021

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

**Abstract:** Diabetes incidence has been a problem, because according with the World Health Organization and the International Diabetes Federation, the number of people with this disease is increasing very fast all over the world. Diabetic treatment is important to prevent the development of several complications, also lipid profile monitoring is important. For that reason the aim of this work is the implementation of machine learning algorithms that are able to classify cases, that corresponds to patients diagnosed with diabetes that have diabetes treatment, and controls that refers to subjects who do not have diabetes treatment but some of them have diabetes, bases on lipids profile levels. Logistic regression, K-nearest neighbor, decision trees and random forest were implemented, all of them were evaluated with accuracy, sensitivity, specificity and AUC-ROC curve metrics. Artificial neural network obtain an accuracy of 0.685 and an AUC value of 0.750, logistic regression achieve an accuracy of 0.729 and an AUC value of 0.795, K-nearest neighbor gets an accuracy of 0.669 and an AUC value of 0.709, on the other hand, decision tree reached an accuracy of 0.691 and a AUC value of 0.683, finally random forest achieve an accuracy of 0.704 and an AUC curve of 0.776. The performance of all models was statistically significant, but the best performance model for this problem corresponds to logistic regression.

**Keywords:** type 2 diabetes; diabetic treatment; logistic regression; random forest; K-nearest neighbor; decision trees; computer-aided diagnosis; statistical analysis

## 1. Introduction

Diabetes is part of a diseases set called Noncommunicable Diseases (NCDs), it is known as a chronic disease, two important characteristics are that it is of a long duration and it is the result of a scheme of several factors as: genetic, environmental and behaviours factors [1]. Diabetes is characterized by hyperglycemia, it refers to a complex disorder that involve profound alterations in the metabolism of fats, proteins and carbohydrates, resulting from defects in insulin secretion, insulin action, and even both of them [2,3]. There are several problems that can be developed like: long-term damage, failure or dysfunction of various organs, vascular complications, all of them shorten the life expectancy of who is diagnosed with this disease [4].

The control of diabetes and the problems that can be developed is the main objective for doctors and diagnosed patients, for that reason, the patient must have a diabetes treatment. Antidiabetic drugs are pharmacological agents that have been approved for

hyperglycemic treatment in diabetes type 2. These drugs are classified as biguanides, sulfonylureas, meglitinides, thiazolidinediones, dipeptidyl peptidase IV inhibitors and  $\alpha$ -glucosidase inhibitors [5].

The lipid control levels are important in diabetic patients [6]. A high cholesterol level can lead to the accumulation of plaques on the walls of blood vessels, and this can block arteries and cause high blood pressure, stroke, heart disease, or heart attack. High triglycerides are associated with the risk of developing metabolic syndrome, which can increase the risk of heart disease and other disorders, including diabetes [7,8].

Recent studies in the health area, have been adopting machine learning and deep learning algorithms, due to the high performance in several healthcare applications, this is part of diseases diagnosis or classification making implementations of algorithms based on computer-aided diagnosis (CAD) where prediction models are used when it is necessary to know in the future the behavior of some data related to any disease, for example diabetes [9]. Machine learning techniques are implemented to discover patterns from medical data sources and provide excellent capabilities to predict diseases or classify diseases [10].

On the other hand, it is important to mention that diabetes incidence has been a problem, because according with the World Health Organization (WHO) and the International Diabetes Federation (IDF), the number of people with this disease is increasing very fast all over the world [1]. Despite the impact that diabetes has had on society and the efforts made to design effective therapeutic protocols and drugs, it has been documented that most current therapies for this disease are developed in the absence of defined molecular targets or a complete delineation of the pathogenesis of the disease. For this reason and due to the continuous increase in knowledge of pathophysiological mechanisms and the side effects of therapeutic protocols, drug design and discovery has become a major challenge in the field of diabetes research. The contribution of data mining and machine learning in this area are focused on helping in different aspects, such as: making recommendations and improvements to the effectiveness of medication, making predictions, as well as suggestions for more personalized medications, also helping to design more effective blood glucose reduction factors, as well as improving the planning and dosage of medications, and applying the administration of drugs in a more specific way [11].

For that reason, the main objective of this study is to analyze the relation that exists between the diabetes treatment and lipids profile, implementing machine learning algorithms.

The contribution of this work is that given the lipids profile, age and gender, the machine learning algorithm can determine if a person is on diabetes treatment, because there were five machine learning algorithms implemented, and based on a comparative, it can determine which of them provided the best model to give a solution to this problem, permitting us to know if there is a relationship between subjects who have or do not have diabetic treatment and their lipids profile, being a first approach to help with the lipids profile control in subjects with type 2 diabetes and how the medication modifies parameters to optimize the treatment developing a computational assisted diagnosis (CAD). Due to it being considered an important first step for future research in this area.

#### *Related Work*

Diabetes represents one of the greatest challenges of this century, because it is a major cause of death and disability worldwide [12]. Mainly type 2 diabetes, due to it being an expanding health problem [13]. This disease is influenced by genetic risk and diverse environmental factors [14].

For that reason, there are some machine learning approaches focused in this disease. A related work uses machine learning paradigm to detect diabetes disease, National Health and Nutrition Examination Survey (NHANES 2009–2012) diabetes dataset was used, they implemented naïve bayes, decision tree, adaboost and random forest to predict diabetes disease. The highest accuracy results were obtained from a combination of logistic regression and random forest, that was 94.15% and an AUC of 0.94 [15].

Another approach consist in the classification of diabetic patients through lipids profile levels, Guerrero-Flores et al. [16] implemented three algorithms, which are logistic regression, decision trees and support vector machine, the AUC values obtained are from 0.613 to 0.727.

Almatrooshi et al. [17] proposed an integration of two systems to create a system that is able to detect diabetes and after that recommend a proper plan or medication to overcome diabetes, they evaluated and tested four different approaches to detection diabetes, the most accurate approach was random forest with an accuracy of 79.2% and F-measure of 0.787.

On the other hand, Koren et al. [18] decided to test the utility of machine learning applied to big data, specifically in the identification of the potential role of concomitant drugs not taken for diabetes which may contribute to lowering blood glucose. They implemented logistic regression to predict the probability of treatment success with the matched drug and this constituted the propensity score. The basis metric was HgA1c, then patients with levels <6.5 were classified as successfully treated and according to the results, 54% of the patients were successfully treated.

In the approach proposed by Alcalá-Rmz et al. [19] 19 para-clinical features were used to determine the health status of the patients. Among the 19 features there were the lipids profile of each subject. They developed a model that was evaluated through a statistical analysis based on the calculation of the loss function, accuracy, area under the curve (AUC) and receiving operating characteristics (ROC) curve. This model was able to obtain an accuracy of 0.94, and AUC values of 0.98.

Hosseini et al. [20] proposed an algorithm that is based on the notion of Markov blankets in Bayesian networks. They applied the algorithm with the aim to optimize medication prescriptions for diabetic patients, taking in count different features, for example if the patient suffers of multiple comorbidities and if the subject is currently taking multiple medications. With this study, they evaluated the features with a bayessian network, and achieve a precision of 88.75% and an AUC of 71.15%

The objective of the Wu et al.'s [21] study, is to assess several machine learning algorithms and screen out a model that can be used to predict patients' non-adherence risks. For this work, 401 patients were selected from 630 candidates, of which 85 were evaluated as poor adherence (21.20%). A total of 16 featured were evaluated in the model, 300 models were built based on 30 machine learning algorithms. The highest results corresponds to an AUC of 0.866.

Kowsher et al.'s [22] research consist of a comparative study of 7 machine learning classifier algorithms and an artificial neural network approach to predict the detection and treatment of diabetes. The training dataset was comprised by the information of 9483 diabetic patients. The performance evaluation metrics were accuracy and precision, looking for the best algorithm, the best accuracy performance was from artificial neural network with 95.14%.

In the Wright et al.'s study, the sequential pattern mining approach is used, the main objective was to identify the temporal relationships between medication prescriptions, and in this way predict the follow-up medication to be prescribed for a patient [23].

Finally, Oh et al. proposed a novel method to follow trajectories, the method was focused on studying the trajectories of type 2 diabetes, using electronics health record systems. They were able to identify the trajectory that most people follow, which consists of a sequence of diseases ranging from hyperglycemia to hypertension, impaired fasting glucose and type 2 diabetes [24].

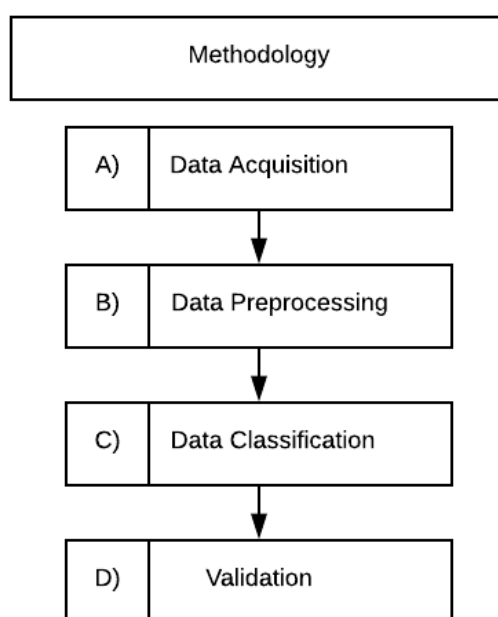
In recent years there are many related works that involves machine learning as a useful tool for many approaches that pretend to give a solution to the diabetes problem, as mentioned before, machine learning is being used with different objectives from the classification of diabetes patients to determining the optimal diabetic treatment for diabetic patients, all of them pretend to give solutions to society as far as this disease is concern. In addition, some related work has found hyperlipidemia to be a relevant factor in the study

of type 2 diabetes. In this paper we propose the implementation of five machine learning classifiers, which are neural networks, logistic regression, k-nearest neighbor, decision trees and random forest that are able to classify, if a subject has a diabetic treatment or if a subject does not have treatment through lipids profile values like: total cholesterol, High Density Lipoprotein (HDL), Low Density Lipoprotein (LDL) and triglycerides.

## 2. Materials and Methods

The methodology followed to analyze the database provided by the Medical Research Unit in Biochemistry, “Centro Médico Siglo XXI”, “Instituto Mexicano del Seguro Social”, is presented in this section, as well as the data description, data preprocessing, data classification and the validation of the results. The study focuses on the classification of subjects who have diabetes treatment and subjects who do not have the treatment.

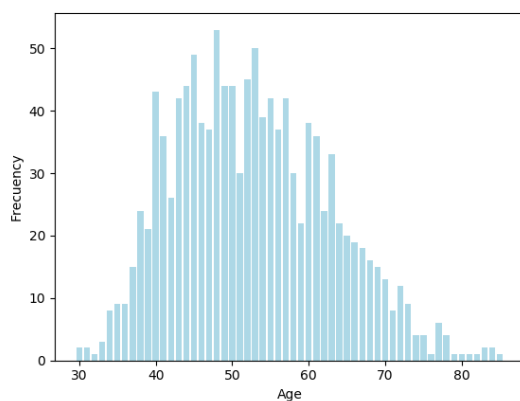
The methodology flowchart is shown in Figure 1. Section (A) corresponds to data acquisition. The next section (B) refers to data preprocessing, in this stage some techniques were implemented to analyze the database, solve outliers problems and separate the data in two subsets. In section (C) the implemented machine learning algorithms is explained: neural networks, logistic regression, K-nearest neighbor, decision tree and random forest. Finally, section (D) represents the validation process for each model, using statistical parameter like sensitivity, specificity, AUC, ROC curve and accuracy.



**Figure 1.** Flowchart of the methodology followed.

### 2.1. Dataset Description

The dataset contains information referring to 1198 Mexican subjects, 537 cases that have diabetes and diabetes treatment, and 661 controls who do not have diabetes treatment but some of them have diabetes and some of them do not have diabetes. Also, in the dataset there are 499 males and 561 females, Figure 2 corresponds to age distribution, the age range of the dataset is between 30 and 85 years old.



**Figure 2.** Age distribution.

Also the dataset is composed by another 4 input features, which are described in Table 1. The output feature indicates “1” whether a subject has diabetes treatment or “0” does not have treatment. The cases treatment includes metformin, glibenclamide, pioglitazone, rosiglitazone, acarbose or insuline.

**Table 1.** Description of features contained in the dataset.

Feature	Description
Age	Subject age
Gender	Subject Gender
CHOL	Total Cholesterol (mg/dL)
HDL	High Density Lipoprotein (mg/dL)
LDL	Low Density Lipoprotein (mg/dL)
TG	Triglycerides (mg/dL)
TX	0—absence of diabetic treatment 1—diabetic treatment

## 2.2. Data Preprocessing

The dataset features were normalized using z-score method, this method consists of transforming the data to a distribution with mean 0 and standard deviation 1. Z-score method was used with two purposes, the first one was to define the same numeric scale for the data, the second one was to identify outliers in the dataset, because while calculating the Z-score the data were re-scaled and centered and there were looking for data points which were too far from zero, then the data points which were way too far from zero were treated as the outliers. A threshold of 3 or  $-3$  was used, because it is the commonly used [25]. Also the boxplots was used to visualize the data distribution before and after removing outliers for each feature.

Boxplot is known as a “box and whisker plot” and it is a method to identify outliers. The diagram is comprised by a box with the interquartile range (IQR), the box has a line horizontally in the middle, this represents the median score [25,26]. On the other hand, there is an upper quartile which represents the 75th percentile, also the bottom of the box shows the lower quartile which represents the 25th percentile. Finally, the long extensions from the box represents the highest and lowest values in the expected normal distribution [27,28].

Figure 3 shows the boxplot for subjects ages, the number 1 represent the boxplot ages before deleting the outliers and number 2 refers to boxplot ages after removing outliers. People over 80 years old were deleted from the dataset.

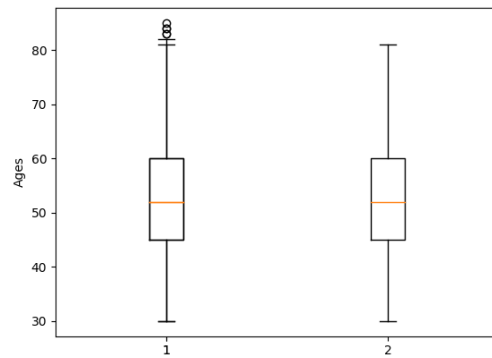


Figure 3. Age Boxplots.

Figure 4a presents the boxplot for HDL and Figure 4b shows the boxplot for LDL features, in both cases 1 corresponds to boxplot before removing the outliers and number 2 represents the boxplot after deleting outliers.

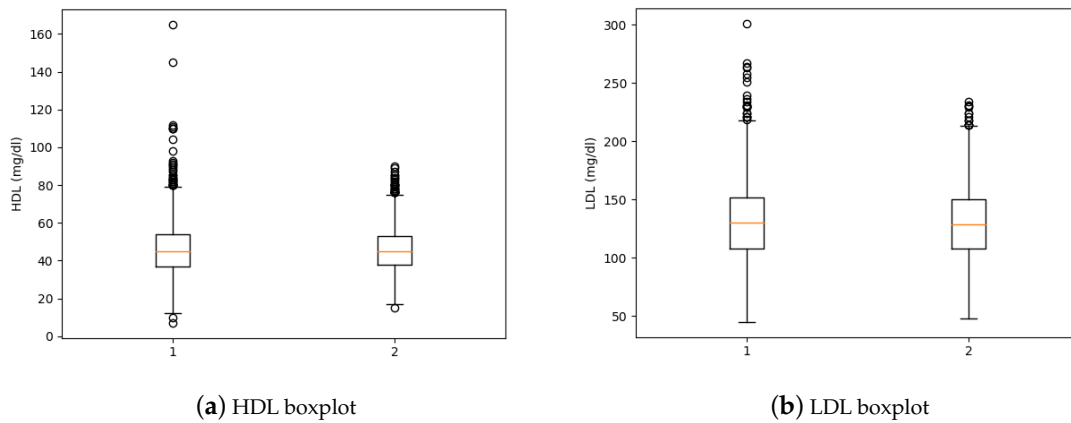


Figure 4. HDL and LDL Boxplots.

Finally, the boxplots for total cholesterol are shown in Figure 5a and the boxplots for triglycerides are presented in Figure 5b. As in the previous diagrams, the number 1 indicates the boxplot before deleting the outliers, and number 2 refers to boxplot after removing outliers, according to the threshold selected.

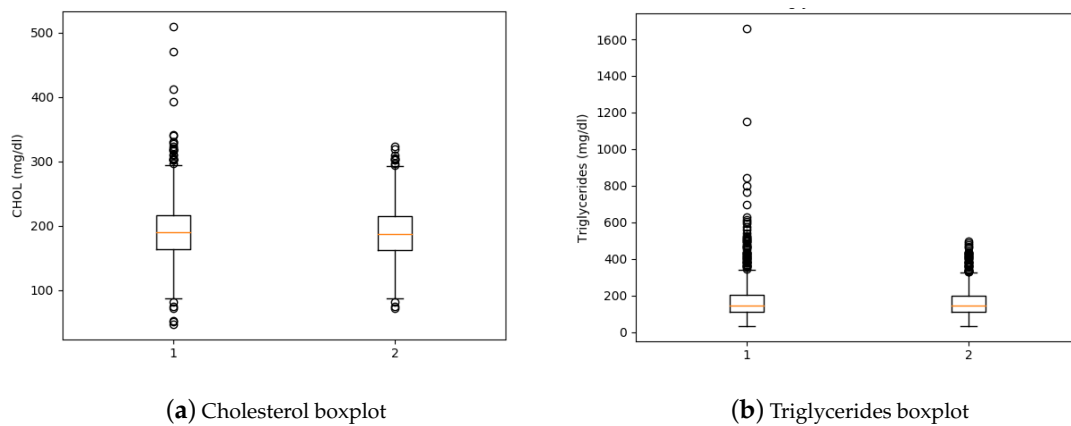


Figure 5. Cholesterol and Triglycerides Boxplots.

Once the outliers corresponding to the chosen threshold have been eliminated, the dataset decreased from 1198 subjects to 1060, where 459 patients have diabetes and diabetes treatment and 601 subjects who do not have diabetic treatment but some of them have diabetes.

Due to the dataset size, it was decided to perform a blind test, which consists of dividing the data set into two subsets, one for training and the other for testing. Then, the main dataset was randomly divided in two sets. The first one refers to the training set, involving 70% of all dataset, and the second was the test set, this subset corresponds to the remaining 30%.

### 2.2.1. Data Classification

In this section it is explained the machine learning algorithms, the tools and packages used to the classification stage.

The subjects who do not have a diabetes treatment are labeled as “0”, which are control subjects, the case subjects were labeled with “1” and corresponds to patients with diabetes treatment. The implemented algorithms were: neural networks, logistic regression, K-nearest neighbor, decision tree and random forest, using the scikit-Learn, keras and tensorflow packages, for Python.

- Scikit-Learn: is a Python library that provides different supervised and unsupervised learning algorithms to solve regression, classification, clustering problems, etc. It is built upon packages like Numpy, Pandas, Scipy and matplotlib [29].
- Keras: is known as a high-level Artificial Neural Network API, which is written in Python, designed specifically to enable fast experimentation with deep neural networks, it focuses on being user-friendly, modular, and extensible [30].
- Tensorflow: is an open-source software library for dataflow programming. It is a symbolic math library. Furthermore it is used for machine learning applications such as neural networks [31].
- Deep Neural Network: Neural Networks looking for a solution using a correlation between features. Neural Networks acquire their knowledge by detecting relationships and patterns between the data, learning through experience. It is composed by hundreds of neurons that can be modifiable, connected with weights, organized in layers, which can also be modified. Learning rule, transfer functions and architecture itself are parameters that determine the deep neural network behavior [32]. Two activation functions were used, the first one was Rectified Linear Unit (ReLU), which is shown in Equation (1), it was implemented in all dense layers, except outter layer, the function assigns “0” to neurons that have a value lower than “0”, and the original value is assigned when the value is above or equal to “0” [33].

$$ReLU(z) = \begin{cases} \text{if } z < 0 & 0 \\ \text{if } z \geq 0 & z \end{cases} \quad (1)$$

The second function was softmax based on general logistic function, represented in Equation (2) where  $\sigma(z)$  is a  $K$ -dimensional vector of  $z$ . It gives a vector of arbitrary values located between 0 and 1 [34].

$$\sigma(z)_j = \frac{e^{z_j}}{\sum_{k=1}^K e^{z_k}}, j = 1, \dots, K \quad (2)$$

In this study, the independent variables are; age, gender, hdl, ldl, chol and tg, the dependent variable is tx, that indicates if the subject has or does not have diabetes treatment. The main parameters used were 5 dense layers, with 100 neurons each one, except the last one that corresponds to the output and has 2 neurons; 3 dropout layers, with a rate of 0.25, 0.50 and 0.25 respectively; the optimizer implemented was “Adam” and 100 epoch.



- Logistic Regression: consists of measuring the relationship between the categorical dependent variable and the independent variables, by estimating probabilities using a logistic function. It is used to predict binary response based on one or more predictor variables. In general terms, logistic regression can be defined as is shown in Equation (3), where  $p$  is the chance of the distinctive of interest [35,36].

$$\text{logit}(p) = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_3 + \dots + b_k X_k \quad (3)$$

The model obtained by the logistic regression permit us to know the relationship between the output, which corresponds to “0” if the subject is not taking antidiabetic medication or “1” does not have treatment; and the entry values, which are age, gender, hdl, ldl, chol an tg; through the analysis of the results, it is possible to know if the relationship exists or not.

- K-Nearest Neighbor (K-NN): is a simple classification algorithm, based on a distance metric, that evaluate the similarity of two features vectors. The main objective is to select the class label for the new input, which appears frequently in its  $k$  near neighbors. In other words, the purpose is computing similarities between unknown sample and training samples, the idea is to find the top  $k$  nearest neighbors of the unknown sample [37]. The K-NN algorithm has four principal steps [38,39]:
  - Select a “K” value (Neighbors)
  - For each example, calculate the distance between the query example and the current example from the data. After that, add the distance and the index of the example to an ordered collection. This distance is called Euclidean distance, which is shown in Equation (4), where  $D(a,b)$  means Euclidean length between  $b$  and  $a$  [40].
  - Sort the calculated distances in ascending order.
  - Obtain the top  $k$  rows from the sorted array.
  - Get the most frequent class.
  - Return the predicted class.

$$D(a,b) = \sqrt{\sum_{i=1}^n (b_i - a_i)^2} \quad (4)$$

In this study the K-NN algorithm is implemented,  $k\_neighbor = 16$ , the same 6 input features are used and in the same way, the output feature is “tx”, which indicates if a subject has diabetes treatment or not.

- Decision Tree: refers to a machine learning model, commonly used in classification problems. The algorithm analyzes the data by making decisions based on asking a series of question. It is a classifier model that correspond to a supervised learning algorithm [41,42]. In other words, this is a predictive model, which is used effectively to classify datasets. This model determines the best decisions in the analysis process, splitting the data into subset. In the learning stage, the model manages to maximize the information gain  $I$  in a given node, and it is represented as in Equation (5).

$$I = H(S) - \sum_{i \in L, R} \frac{|S^i|}{|S|} H(S^i) \quad (5)$$

- Random Forest (RF): it is an algorithm widely used in medical areas, is a supervised method that uses multiples decision trees to create a forest. RF builds multiple decision trees and merge them into a single tree, the purpose is to achieve a high prediction accuracy. In this model, there are settings that constructed many classification and regression trees using randomly selected training datasets and random subsets of prediction variables for modeling outcomes, then the results from the tree are added to give a prediction [43]. It is possible to use entropy to determine how nodes branch



in a decision tree, as is shown in Equation (6), taking into consideration the probability of a certain outcome in order to make a decision on how the node should branch.

$$Entropy = \sum_{i=1}^C -p_i * \log_2(p_i) \quad (6)$$

The algorithms were implemented in Python, which is an interpreter, object-oriented, high-level programming language with dynamic semantics. It can be used for a variety of applications, one of them is the data analysis [44].

### 2.2.2. Evaluation

In the evaluation stage the metrics used to compare the models performance are presented. The outputs are presented in a confusion matrix, where the diagonal represents the observations that are correctly classified and the values outside of diagonal corresponds the observations that were incorrectly classified, also the class error for each model is calculated, based on confusion matrix values. In addition the accuracy and the Receiver Operating Characteristic (ROC) were calculated for each machine learning algorithm implemented.

The accuracy metric calculates the average performance of the algorithms, the purpose of this metric is to calculate the percentage of samples that are correctly classified as is shown in the Equation (7), where *TP* corresponds to true positives, *TN*, true negatives, *FP* false positives and *FN* false negatives.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (7)$$

In addition, the ROC curve is a parameter used to measure the classification precision of the model, through the sensitivity and specificity. Sensitivity refers to the proportion of subjects with a positive condition that were correctly classified and it is calculated with Equation (8), where *TP* are the true positives and *FP* are the false positives [45].

$$Sensitivity = \frac{TP}{TP + FN'} \quad (8)$$

Specificity refers to the proportion of true negatives, which means the subjects with a negative condition that were correctly classified and it is calculated with Equation (9), where *TN* are the true negatives and *FN* are the false negatives [45].

$$Specificity = \frac{TN}{FN + TP'} \quad (9)$$

The ROC curve is used to visualize the performance of the classifiers, this is complemented with the area under the ROC curve (AUC), this complement represents the probability that a random positive sample is correctly identified.

In this study the ROC curves were calculated for each classifier algorithm.

The implementation of this work was performed with a laptop DELL g7, Intel Core i7-8750H 2.20 GHz, 16 GB, 500 GB SSD, Windows 10, 64-bit; and with the version of Python 2.7.

## 3. Results

In this study a dataset is used with a total of 7 features, which 6 of them are the input data for the classifiers, and the remaining feature is the output feature, which indicates with the label "0" the absence of diabetic treatment and with label "1" if the patient has a diabetic treatment. The main objective is to look for the machine learning algorithm with the highest performance in the binary classification explained above.

In the preprocessing step, they were analyzed with boxplot and z-score, the possible outliers for each feature, a threshold was also selected to remove the points data which were too far from zero, for this reason the dataset was reduced from 1198 subjects to 1060,

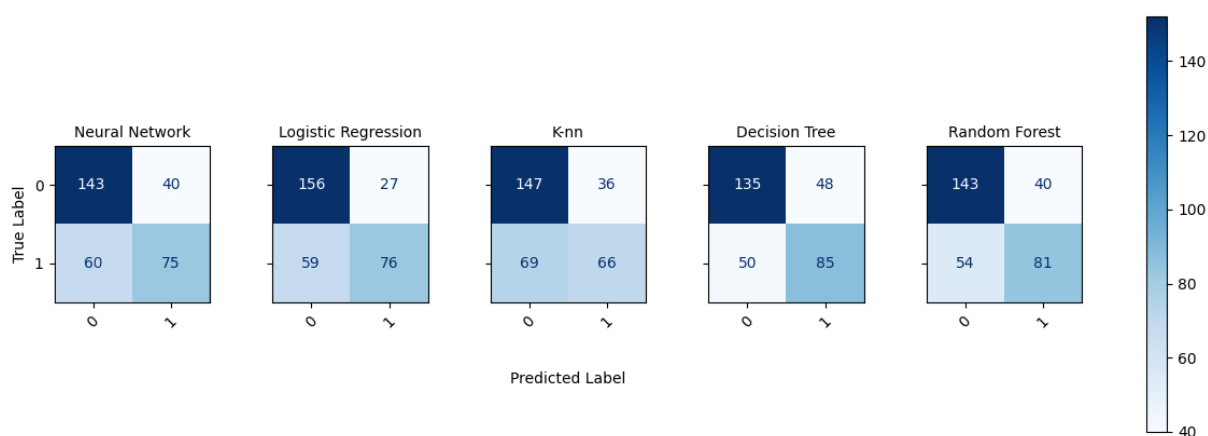
where 459 patients were cases and 601 subjects were control. After this step, the dataset was divided in two subsets, one for training, containing 70% of the data (421 controls/321 cases), and one for testing containing 30% of the data (180 controls/138 cases).

Once finished the preprocessing stage, the implementation of each classifier was carried out, it was calculated the accuracy, sensitivity, specificity and AUC value for each model. In Table 2 the classifier name and the values obtained in each metric are presented. Logistic regression achieved the best performance according to the metrics values obtained, followed by random forest, which got an accuracy value of 0.704 and an AUC value of 0.776, this means that the model is able to classify 77.6% of the data correctly, after that, neural network obtained an accuracy of 0.685 and an AUC of 0.750. The K-NN and Decision Tree obtained the worst performance, but both models got a statistical significance values.

**Table 2.** Performance comparison based on accuracy and AUC.

Classifier	Accuracy	Sensitivity	Specificity	AUC
Neural Network	0.685	0.781	0.555	0.750
Logistic Regression	0.729	0.852	0.562	0.795
K-NN	0.669	0.803	0.488	0.709
Decision Tree	0.691	0.737	0.629	0.683
Random Forest	0.704	0.781	0.600	0.776

In addition, the confusion matrix of each implemented algorithm are presented in Figure 6. The diagonal of each matrix contains the predictions that were correctly classified, and the off-diagonal of each matrix represents the observations that were incorrectly classified.



**Figure 6.** Confusion matrix of each implemented algorithm.

Furthermore, Table 3 shows the error class for each algorithm, the error was calculated from the confusion matrix values, the minimum error value is presented by logistic regression for class 0 and the minimum error value is presented by decision tree for class 1.

**Table 3.** Class error for each algorithm implemented, based on confusion matrix.

	Class Error	
	0	1
Neural Network	0.218	0.444
Logistic Regression	0.147	0.437
K-nn	0.196	0.511
Decision Tree	0.262	0.370
Random Forest	0.218	0.400

Also, each model was validated calculating their ROC curves based on the performance of each model, which are presented in Figure 7. The ROC curve for neural network is presented in orange line, with an AUC value of 0.75. ROC curve for the logistic regression is presented in dark blue, with an AUC value of 0.79. The ROC curve for k-nn model is shown in blue color, with an AUC value of 0.71. The ROC curve calculated for decision tree classifier is presented in pink line, obtaining an AUC value of 0.68, and finally, the ROC curve calculated for random forest model is presented in purple, with an AUC value of 0.78.

The performance of logistic regression and random forest models was similar, on the other hand, the performance of k-nn and decision tree were lower, but with a statistical significance in this area.

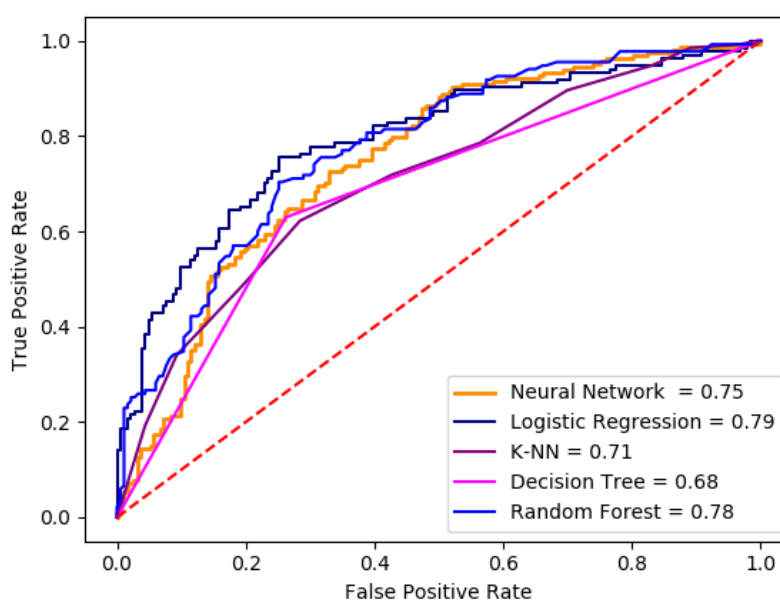


Figure 7. ROC curves obtained with the average performance of each implemented model.

#### 4. Discussion

The related work indicates the importance of analyzing the drugs that are prescribed to diabetic patients, there are different approaches, but all of them approach the same objective, which is to know the impact that drugs have on patients, as well as to find patterns that help the pharmaceutical industry to improve drugs, based on the needs that arise. This work is a first approach to know the impact that different diabetic medications have on patient's health, specifically within the lipid profile, because an uncontrolled lipid level can lead to different complications [7,8].

In this work five machine learning algorithms are implemented: neural networks, logistic regression, K-NN, decision tree and random forest; each model was validated through accuracy, sensitivity, specificity, AUC and ROC curves metrics.

The evaluated dataset is comprised by 1060 Mexican subjects, where 459 are subjects diagnosed with diabetes that follow a diabetic treatment and the remaining 601 are control subjects with or without diabetes but none of them follow a diabetic treatment. The features analyzed are: Age, Gender, HDL, LDL, Cholesterol, Triglycerides and Treatment (TX), it is important to mention that these features were selected with the aim to classify subjects with diabetes treatment from subjects who do not have diabetic treatment, through their lipids profile.

The highest accuracy obtained in the testing stage corresponds to logistic regression model, which achieve an accuracy of 0.729 with a sensitivity of 0.852 and 0.795 in AUC metric, there is no big difference between this model and random forest model, due to its performance it indicates an accuracy of 0.704, a sensitivity value of 0.781 and 0.776 in the AUC metric, another model with high performance is neural network. which performance

corresponds to an accuracy of 0.685, 0.781 of sensitivity and an AUC of 0.750. These three models are able to classify more than 70% of subjects correctly. K-NN and decision tree are not far from the other two, both of them achieve significant values that are over 66% in accuracy, sensitivity and AUC metrics.

In addition, Figure 7 shows that all curves presented statistically significant values  $> 67\%$ . These curves refers to the proportion of true positives and true negatives.

On the other hand, this study demonstrate that it is possible to identify subjects who have diabetes treatment from those who do not have diabetes treatment, showing the relevance of cholesterol, HDL; LDL and triglycerides features. The models obtained can be useful for doctors to know if patients are following the established treatments correctly or simply know if a new patient has been taking an antidiabetic drug, allowing for a better control of the patient's medical history. Furthermore, according to [46] subjects that are taking sulphonylurea therapy, have observed effects on their lipids profile, also a group treated with insulin along with metformin had significant improvement in the lipids levels. Because, once that the Hemoglobin A1c (HbA1c) is improved through diabetic treatment, the lipids profile can significantly improve [47]

It is important to remember that the dataset is composed by mexican subjects information, for this reason, the results can be implemented in tools that allow to improve the Mexican health.

Another advantage of the models implemented is that they do not require high computational cost, because it is not necessary to acquire a special equipment.

## 5. Conclusions

The results obtained in this work, permits us to conclude that the database is adequate for the aim in this study, also, it allows to classify subjects who have a diabetic treatment or that do not have a diabetic treatment, based on the lipids profile, age and gender.

On the other hand, specificity achieves a low value for each model, it is necessary to remember that in the medical area it is more important to find true positives states, because the sensitivity metric should be higher.

Also, it is possible to develop a tool based on lipids profile to detect whether a subject has a diabetes treatment or not, implementing any of the models obtained in this study.

The ROC curves in Figure 7 shows that decision tree is one of the worst AUC with 0.68, which means that only 68% of the subjects were correctly classified, but it is important to mention that the performance values might be improved increasing the observation numbers in the dataset. Besides, logistic regression was able to classify 79% of the total subjects, random forest has the second place, because it achieved a 78% of the subjects correctly classify, neural network was able to classify 75% of the subjects in the correct way and K-NN obtained an AUC value of 0.71, which means that 71% of the subjects in the dataset were classified correctly.

In addition, the results obtained demonstrate that the lipids profile is an important feature in this classification, because it can be modeled by the classifiers implemented, also the results show a relationship between lipids profile and a subject with diabetic treatment.

This work is considered an important basis to search for a specific relationship between the different medications prescribed to a diabetic patient and the impact they have on their lipid profile.

## 6. Future Work

As future work we propose to change the machine learning algorithms parameters, and also find a way to increase the database observation.

On the other hand, it could be interesting to implement other machine learning algorithms or apply a different classification approach, and it also could be important to do an analysis that shows in a clear way the relationship between lipids profile and the diabetic treatment of the subjects.

**Author Contributions:** Conceptualization, C.E.G.-T., A.V.-S. and H.G.-R.; Data curation, V.A.-R., C.E.G.-T., M.C., J.I.G.-T. and J.M.C.-P.; Formal analysis, V.A.-R., C.E.G.-T., A.G.-H., A.V.-S. and J.M.C.-P.; Funding acquisition, J.I.G.-T. and H.G.-R.; Investigation, V.A.-R., C.E.G.-T., A.G.-H., A.V.-S., M.C. and H.L.-G.; Methodology, V.A.-R., C.E.G.-T., A.V.-S., M.C., J.I.G.-T. and H.L.-G.; Project administration, A.G.-H., A.V.-S., M.C., J.I.G.-T. and H.G.-R.; Resources, A.V.-S., M.C., J.I.G.-T., H.L.-G. and H.G.-R.; Software, V.A.-R. and J.M.C.-P.; Supervision, C.E.G.-T. and H.L.-G.; Validation, V.A.-R., C.E.G.-T., A.G.-H. and A.V.-S.; Visualization, J.M.C.-P.; Writing—original draft, V.A.-R., C.E.G.-T., A.G.-H. and A.V.-S.; Writing—review & editing, V.A.-R. and A.G.-H. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** The study was conducted in accordance with the Declaration of Helsinki, and the protocol was approved by the Ethics Committee of Instituto Mexicano del Seguro Social and Comision Nacional de Investigacion Cientifica (R-2011-785-018).

**Informed Consent Statement:** Informed consent was obtained from all subjects involved in the study.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Alwan, A. *Global Status Report on Noncommunicable Diseases 2010*; World Health Organization: Geneva, Switzerland, 2011.
2. Turtle, J.R.; Burgess, J.A. Hypoglycemic action of fenfluramine in diabetes mellitus. *Diabetes* **1973**, *22*, 858–867. [CrossRef]
3. American Diabetes Association. Diagnosis and Classification of Diabetes Mellitus. *Diabetes Care* **2010**, *33* (Suppl. 1), S62–S69. [CrossRef] [PubMed]
4. Socarras, M.B.; Blanco, J.; Vazquez, A.; Gonzáles, D.; Licea, M. Factores de riesgo de aterosclerosis en la diabetes mellitus tipo 2. *Rev. Cubana Med.* **2003**, *42*, 17–25.
5. Moses, R.G. Combination therapy for patients with Type 2 diabetes: Repaglinide in combination with metformin. *Expert Rev. Endocrinol. Metab.* **2010**, *5*, 331–342. [CrossRef] [PubMed]
6. Sugeran, D.T. Blood Lipids. *JAMA* **2013**, *310*, 1751. [CrossRef]
7. MedlinePlus. Triglycerides. Available online: <https://medlineplus.gov/triglycerides.html> (accessed on 5 December 2020).
8. NIH. Blood Cholesterol. Available online: <https://www.nlm.nih.gov/health-topics/blood-cholesterol> (accessed on 5 December 2020).
9. Qayyum, A.; Qadir, J.; Bilal, M.; Al-Fuqaha, A. Secure and robust machine learning for healthcare: A survey. *arXiv* **2020**, arXiv:2001.08103.
10. Shailaja, K.; Seetharamulu, B.; Jabbar, M. Machine Learning in Healthcare: A Review. In Proceedings of the IEEE 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, India, 29–31 March 2018; pp. 910–914.
11. Kavakiotis, I.; Tsave, O.; Salifoglou, A.; Maglaveras, N.; Vlahavas, I.; Chouvarda, I. Machine learning and data mining methods in diabetes research. *Comput. Struct. Biotechnol. J.* **2017**, *15*, 104–116. [CrossRef] [PubMed]
12. Etienne, G.K. Trends in diabetes: Sounding the alarm. *Lancet* **2016**, *387*, 1485–1486.
13. DeFronzo, R.A.; Ferrannini, E.; Groop, L.; Henry, R.R.; Herman, W.H.; Juul Holst, J.; Hu, F.B.; Kanh, C.R.; Raz, I.; Shulman, G.I.; et al. Type 2 diabetes mellitus. *Nat. Rev. Dis. Prim.* **2015**, *1*, 15019. [CrossRef]
14. Cruz, M.; Valladares-Salgado, A.; Garcia-Mena, J.; Ross, K.; Edwards, M.; Angeles-Martinez, J.; Ortega-Camarillo, C.; de la Escobedo Peña, J.; Burguete-Garcia, A.I.; Wachter-Rodarte, N.; et al. Candidate gene association study conditioning on individual ancestry in patients with type 2 diabetes and metabolic syndrome from Mexico City. *Diabetes Metab. Res. Rev.* **2010**, *26*, 261–270. [CrossRef] [PubMed]
15. Maniruzzaman, M.; Rahman, M.J.; Ahammed, B.; Abedin, M.M. Logistic Regression based Feature Selection and Classification of Diabetes Disease using Machine Learning Paradigm. In Proceedings of the 7th International Conference on Data Science and SDGs, Rajshahi, Bangladesh, 18–19 December 2019.
16. Guerrero Flores, M.H.; Galván Tejada, C.E.; Chávez Lamas, N.M.; Galván Tejada, J.; Gamboa Rosales, H.; Celaya Padilla, J.; García Hernández, A.; Valladares Salgado, A.; Cruz, M. Implementación de Algoritmos de Inteligencia Artificial para la Identificación de Pacientes Diabéticos Utilizando los Niveles de Lípidos en Sangre. Available online: <http://ricaxcan.uaz.edu.mx/jspui/handle/20.500.11845/1943> (accessed on 7 December 2020).
17. Almatrooshi, F.; Alhamadi, S.; Salloum, S.A.; Akour, I.; Shaalan, K. A Recommendation System for Diabetes Detection and Treatment. In Proceedings of the IEEE 2020 International Conference on Communications, Computing, Cybersecurity, and Informatics (CCCI), Sharjah, United Arab Emirates, 3–5 November 2020; pp. 1–6.
18. Koren, G.; Nordon, G.; Radinsky, K.; Shalev, V. Identification of repurposable drugs with beneficial effects on glucose control in type 2 diabetes using machine learning. *Pharmacol. Res. Perspect.* **2019**, *7*, e00529. [CrossRef]



19. Alcalá-Rmz, V.; Zanella-Calzada, L.A.; Galván-Tejada, C.E.; García-Hernández, A.; Cruz, M.; Valladares-Salgado, A.; Galván-Tejada, J.I.; Gamboa-Rosales, H. Identification of diabetic patients through clinical and para-clinical features in Mexico: An approach using deep neural networks. *Int. J. Environ. Res. Public Health* **2019**, *16*, 381. [CrossRef]
20. Hosseini, M.M.; Zargoush, M.; Alemi, F.; Kheirbek, R.E. Leveraging machine learning and big data for optimizing medication prescriptions in complex diseases: A case study in diabetes management. *J. Big Data* **2020**, *7*, 1–24. [CrossRef]
21. Wu, X.W.; Yang, H.B.; Yuan, R.; Long, E.W.; Tong, R.S. Predictive models of medication non-adherence risks of patients with T2D based on multiple machine learning algorithms. *BMJ Open Diabetes Res. Care* **2020**, *8*, e001055. [CrossRef]
22. Kowsher, M.; Turaba, M.Y.; Sajed, T.; Rahman, M.M. Prognosis and Treatment Prediction of Type-2 Diabetes Using Deep Neural Network and Machine Learning Classifiers. In Proceedings of the IEEE 2019 22nd International Conference on Computer and Information Technology (ICCIT), Dhaka, Bangladesh, 18–20 December 2019; pp. 1–6.
23. Wright, A.P.; Wright, A.T.; McCoy, A.B.; Sittig, D.F. The use of sequential pattern mining to predict next prescribed medications. *J. Biomed. Inform.* **2015**, *53*, 73–80. [CrossRef] [PubMed]
24. Oh, W.; Kim, E.; Castro, M.R.; Caraballo, P.J.; Kumar, V.; Steinbach, M.S.; Simon, G.J. Type 2 diabetes mellitus trajectories and associated risks. *Big Data* **2016**, *4*, 25–30. [CrossRef]
25. Tukey, J.W. *Exploratory Data Analysis*; Addison-Wesley Series in Behavioral Science Quantitative Methods: Reading, MA, USA, 1977; Volume 2.
26. Field, A. *Discovering Statistics Using SPSS: (and Sex and Drugs and Rock 'n' Roll)*; Sage: London, UK, 2009.
27. Cody, R. *SAS Statistics by Example*; SAS Institute: Cary, NC, USA, 2011.
28. Frigge, M.; Hoaglin, D.C.; Iglewicz, B. Some implementations of the boxplot. *Am. Stat.* **1989**, *43*, 50–54.
29. Google. Scikit-Learn. Available online: <https://scikit-learn.org/stable/> (accessed on 11 December 2020).
30. Chollet, F. Keras: Deep Learning Library for Theano and Tensorflow. Available online: <https://keras.io/k> (accessed on 22 June 2020).
31. Google. Tensorflow. Available online: <https://www.tensorflow.org/> (accessed on 22 June 2020).
32. Agatonovic-Kustrin, S.; Beresford, R. Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research. *J. Pharm. Biomed. Anal.* **2000**, *22*, 717–727. [CrossRef]
33. Lomuscui, A.; Maganti, L. An approach to reachability analysis for feed-forward relu neural networks. *arXiv* **2017**, arXiv:1706.07351.
34. Carlini, N.; Wanger, D. Towards evaluating the robustness of neural network. In Proceedings of the 2017 IEEE Symposium on Security and Privacy (SP), San Jose, CA, USA, 22–26 May 2017.
35. Hilbe, J.M. *Logistic Regression Models*; CRC Press: Boca Raton, FL, USA, 2009.
36. Khairunnahar, L.; Hasib, M.A.; Rezanur, R.H.B.; Islam, M.R.; Hosain, M.K. Classification of malignant and benign tissue with logistic regression. *Inform. Med. Unlocked* **2019**, *16*, 100189. [CrossRef]
37. Yang, N. KNN Algorithm Simulation Based on Quantum Information. In Proceedings of the Student-Faculty Research Day Conference (CSIS), New York City, NY, USA, 3 May 2019.
38. Shah, K.; Patel, H.; Sanghvi, D.; Shah, M. A comparative analysis of logistic regression, random Forest and KNN models for the text classification. *Augment. Hum. Res.* **2020**, *5*, 1–16. [CrossRef]
39. Liu, L.; Su, J.; Liu, X.; Chen, R.; Huang, K.; Deng, R.H.; Wang, X. Toward highly secure yet efficient KNN classification scheme on outsourced cloud data. *IEEE Internet Things J.* **2019**, *6*, 9841–9852. [CrossRef]
40. Kowsher, M.; Tithi, F.S.; Rabeya, T.; Afrin, F.; Huda, M.N. Type 2 Diabetics Treatment and Medication Detection with Machine Learning Classifier Algorithm. In *Proceedings of International Joint Conference on Computational Intelligence*; Springer: Singapore, 2020; pp. 519–531.
41. Alam, F.; Mehmood, R.; Katib, I. Comparison of decision trees and deep learning for object classification in autonomous driving. In *Smart Infrastructure and Applications*; Springer: Cham, Switzerland, 2020; pp. 135–158.
42. Assegie, T.A.; Nair, P.S. Handwritten digits recognition with decision tree classification: A machine learning approach. *Int. J. Electr. Comput. Eng.* **2019**, *9*, 4446. [CrossRef]
43. Speiser, J.L.; Miller, M.E.; Tooze, J.; Ip, E. A comparison of random forest variable selection methods for classification prediction modeling. *Expert Syst. Appl.* **2019**, *134*, 93–101. [CrossRef] [PubMed]
44. Google. Python. Available online: <https://www.python.org/doc/essays/blurb/> (accessed on 11 December 2020).
45. Hanley, J.A.; McNeil, B.J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **1982**, *143*. [CrossRef] [PubMed]
46. Levetan, C. Oral antidiabetic agents in type 2 diabetes. *Curr. Med. Res. Opin.* **2007**, *23*, 945–952. [CrossRef]
47. Barr, M.M.; Aslibekyan, S.; Ashraf, A.P. Glycemic control and lipid outcomes in children and adolescents with type 2 diabetes. *PLoS ONE* **2019**, *14*, e0219144.