IEEE

mipro 2020

ISSN 1847-3946

organizer upro

43rd international convention

September 28 - October 2, 2020, Opatija, Croatia

Lampadem tradere

mipro - innovative promotional partnership

mipro proceedings

MIPRO 2020

43rd International Convention

September 28, 2020 – October 2, 2020
Opatija, Croatia

Proceedings

Conferences:

Microelectronics, Electronics and Electronic Technology /MEET

Data Science and Biomedical Engineering /DS-BE

Telecommunications & Information /CTI

Computers in Education /CE

Computers in Technical Systems /CTS

Intelligent Systems /CIS

Robotics Technologies and Applications/RTA

Information Systems Security /ISS

Business Intelligence Systems /miproBIS

Digital Economy and Digital Society /DE-DS

Information and Communication Technology Law /ICTLAW

Engineering Education /EE

Software and Systems Engineering /SSE

MIPRO Junior - Student Papers /SP

Smart, Sustainable And Resilient Cities And Infrastructure /SSRCI

Optoelectronics and Photonics /OPHO

Dew Computing /DEWCOM

Edited by:
Karolj Skala

International Program Committee

Karolj Skala, General Chair (Croatia)
Lejla Abazi-Bexheti (North Macedonia)
Enis Afgan (United States),
Miimu Airaksinen (Finland),
Saša Aksentijević (Croatia),
Slaviša Aleksić (Germany),
Slavko Amon (Slovenia),
Krešo Antonović (Croatia),
Michael E. Auer (Austria),
Viktor Avbelj (Slovenia),
Dubravko Babić (Croatia),
Snježana Babić (Croatia),
Tadej Bajd (Slovenia),
Ante Bakić (Croatia),
Marko Banek (Croatia),
Mirta Baranović (Croatia),
Bartosz Bebel (Poland),
Nina Begičević Ređep (Croatia),
Ladjel Bellatreche (France),
Adrian Boukalov (Belgium),
Ricardo Branco (Portugal),
Ljiljana Brkić (Croatia),
Marian Bubak (Poland),
Andrea Budin (Croatia),
Željko Butković (Croatia),
Patrizio Campisi (Italy),
Željka Car (Croatia),
Jesús Carretero Pérez (Spain),
Bojan Cukic (United States),
Alfredo Cuzzocrea (Italy),
Duško Čakara (Croatia),
Stipo Čelar (Croatia),
Marina Čičin-Šain (Croatia),
Dragan Čišić (Croatia),
Davor Davidović (Croatia),
Vlado Delić (Serbia),
Radoslav Delina (Slovakia),
Matjaž Depolli (Slovenia),
Saša Dešić (Croatia),
Dražen Dragičević (Croatia),
Todd Eavis (Canada),
Maurizio Ferrari (Italy),
Tiziana Ferrari (Netherlands),
Nikola Filip Fijan (Croatia),
Renato Filjar (Croatia),
Tihana Galinac Grbac (Croatia),
Enrico Gallinucci (Italy),
Dragan Gamberger (Croatia),
Paolo Garza (Italy),
Tom Gavazzi (Croatia),
Gordan Gledec (Croatia),
Matteo Golfarelli (Italy),
Stjepan Golubić (Croatia),
Vera Gradišnik (Croatia),
Simeon Grazio (Croatia),
Andrej Grgurić (Croatia),
Stjepan Groš (Croatia),
Nina Gumzej (Croatia),
Marjan Gusev (North Macedonia),
Jaak Henno (Estonia),
Bojan Hlača (Croatia),
Željko Hocenski (Croatia),
Tatjana Holjevac (Croatia),
Vlasta Hudek (Croatia),
Darko Huljenić (Croatia),
Robert Inkret (Croatia),
Ivo Ipšić (Croatia),
Mile Ivanda (Croatia),
Marina Ivašić-Kos (Croatia),
Hannu Jaakkola (Finland),
Tomislav Jagušt (Croatia),
Darko Jardas (Croatia),
Vojko Jazbinšek (Slovenia),
Leonardo Jelenković (Croatia),
Bojan Jerbić (Croatia),
Dragan Jevtić (Croatia),
Alen Jugović (Croatia),
Admela Jukan (Germany),
Oliver Jukić (Croatia),
Irena Jurdana (Croatia),
Ozren Jureković (Croatia),
Marko Jurić (Croatia),
Đani Juričić (Slovenia),
Nikola Kadoić (Croatia),
Jurij Matija Kališnik (Slovenia),
Ivan Kaštelan (Serbia),
Zlatko Katalenić (Slovenia),
Ana Katalinić Mucalo (Croatia),
Tihomir Katulić (Croatia),
Pekka Kess (Finland),
Tonimir Kišasondi (Croatia),
Zalika Klemenc-Ketiš (Slovenia),
Mario Konecki (Croatia),
Marko Koričić (Croatia),

Gregor Kosec (Slovenia),
Igor Kotenko (Russia),
Božidar Kovačić (Croatia),
Miklos Kozlovszky (Hungary),
Danica Kragić Jensfelt (Sweden),
Goran Krajačić (Croatia),
Dieter Kranzlmüller (Germany),
Marjan Krašna (Slovenia),
Srećko Krile (Croatia),
Lene Krøl Andersen (Denmark),
Benjamin Kušen (Croatia),
Marko Lacković (Croatia),
Erich Leitgeb (Austria),
Jadran Lenarčič (Slovenia),
Tomislav Lipić (Croatia),
Hrvoje Lisičar (Croatia),
Dražen Lučić (Croatia),
Duško Lukač (Germany),
Igor Ljubi (Croatia),
Zongmin Ma (China),
Goran Marković (Croatia),
Leslie Martinich (United States),
Ludek Matyska (Czech Republic),
Mladen Mauher (Croatia),
Igor Mekterović (Croatia),
Željka Mihajlović (Croatia),
Branko Mikac (Croatia),
Anđelko Milardović (Croatia),
Hrvoje Mlinarić (Croatia),
Thor Moen (Norway),
Lueny Morell (Puerto Rico),
Gorana Mudronja (Croatia),
Neeta Nain (India),
Jadranko F. Novak (Croatia),
Dario Ogrizović (Croatia),
Predrag Pale (Croatia),
Dana Paľová (Slovakia),
Nikola Pavešić (Slovenia),
Mile Pavlić (Croatia),
Branimir Pejčinović (United States),
Ana Perić Hadžić (Croatia),
Dana Petcu (Romania),
Juraj Petrović (Croatia),
Duc Truong Pham (UK),
Damir Pintar (Croatia),
Vincenzo Piuri (Italy),
Tonka Poplas Susič (Slovenia),
Andreja Pucihar (Slovenia),
Aleksandra Rashkovska (Slovenia),
Robert Repnik (Slovenia),
Libuša Révészová (Slovakia),
Slobodan Ribarić (Croatia),
Vittorio Rosato (Italy),
Dubravko Sabolić (Croatia),
Ioan Sacala (Romania),
Davor Salamon (Croatia),
Jörg Schulze (Germany),
Bruno Siciliano (Italy),
Sandro Skansi (Croatia),
Zoran Skočir (Croatia),
Ivanka Sluganović (Croatia),
Mladen Sokele (Croatia),
Ana Sović Kržić (Croatia),
Mario Spremić (Croatia),
Vlado Sruk (Croatia),
Uroš Janez Stanič (Slovenia),
Vjeran Strahonja (Croatia),
Tomislav Suligoj (Croatia),
Aleksandar Szabo (Croatia),
Dina Šimunić (Croatia),
Frano Škopljanac-Mačina (Croatia),
Dejan Škvorc (Croatia),
Zorislav Šojat (Croatia),
Andreja Špernjak (Slovenia),
Vitomir Štruc (Slovenia),
Velimir Švedek (Croatia),
Darko Švelec (Croatia),
Viktor Švigelj (Slovenia),
Zheng-Hua Tan (Denmark),
Nikola Tanković (Croatia),
Antonio Teixeira (Portugal),
Edvard Tijan (Croatia),
Paul Timmers (UK),
A Min Tjoa (Austria),
Ivan Tomašić (Sweden),
Roman Trobec (Slovenia),
Tibor Vámos (Hungary),
Mladen Varga (Croatia),
Lucija Vejmelka (Croatia),
Matjaž Veselko (Slovenia),
Linda Vicković (Croatia),
Marijana Vidas-Bubanja (Serbia),
Davor Vinko (Croatia),
Goran Vojković (Croatia),
Mihaela Vranić (Croatia),
Miroslav Vrankić (Croatia),
Boris Vrdoljak (Croatia),
Slavomir Vukmirović (Croatia),
Yingwei Wang (Canada),
Mario Weber (Croatia),
Roman Wyrzykowski (Poland)

organized by
MIPRO Croatian Society

technical cosponsorship
IEEE Region 8
IEEE Croatia Section
IEEE Croatia Section Computer Chapter
IEEE Croatia Section Electron Devices/Solid-State Circuits Joint Chapter
IEEE Croatia Section Education Chapter
IEEE Croatia Section Communications Chapter
EAI European Alliance of Innovation

under the auspices of
Ministry of Science and Education of the Republic of Croatia
Ministry of the Sea, Transport and Infrastructure of the Republic of Croatia
Ministry of Economy, Entrepreneurship and Crafts of the Republic of Croatia
Ministry of Public Administration of the Republic of Croatia
Ministry of Regional Development and EU Funds of the Republic of Croatia
Ministry of Environment and Energy of the Republic of Croatia
Ministry of Demography, Family, Youth and Social Policy of the Republic of Croatia
Ministry of Agriculture of the Republic of Croatia
Primorje-Gorski kotar County
City of Rijeka
City of Opatija
Croatian Regulatory Authority for Network Industries - HAKOM
Croatian Power Exchange - CROPEX

patrons
University of Zagreb
University of Rijeka
Juraj Dobrila University of Pula
Ruđer Bošković Institute, Zagreb
University of Zagreb, Faculty of Electrical Engineering and Computing
University of Zagreb, Faculty of Organization and Informatics, Varaždin
University of Rijeka, Faculty of Maritime Studies
University of Rijeka, Faculty of Engineering
University of Rijeka, Faculty of Economics and Business
Zagreb University of Applied Sciences
Croatian Academy of Engineering - HATZ
Croatian Regulatory Authority for Network Industries - HAKOM
Ericsson Nikola Tesla, Zagreb
T-Croatian Telecom, Zagreb
Končar - Electrical Industries, Zagreb
HEP - Croatian Electricity Company, Zagreb
A1 Hrvatska, Zagreb

sponsors
HEP - Croatian Electricity Company Zagreb
Končar-Electrical Industries Zagreb
Storm Computers Zagreb
InfoDom Zagreb
A1 Hrvatska Zagreb
Mjerne tehnologije Zagreb
Selmet Zagreb
Institute SDT Ljubljana
Nomen Rijeka

All papers are published in their original form

For Publisher:

Karolj Skala

Publisher:

Croatian Society for Information, Communication and
Electronic Technology – MIPRO
Office: Kružna 8/II, P. O. Box 303, HR-51001 Rijeka, Croatia
Phone/Fax: (+385) 51 423 984

Printed by:

GRAFIK, Rijeka

ISSN 1847-3946

CONTENTS

MICROELECTRONICS, ELECTRONICS AND ELECTRONIC TECHNOLOGY

DATA SCIENCE AND BIOMEDICAL ENGINEERING

INVITED PAPER

PAPERS

DATA SCIENCE

BIOMEDICAL ENGINEERING

TELECOMMUNICATIONS & INFORMATION

INVITED PAPERS

PAPERS

TOWARDS 5G NETWORKS

INTERNET DEVELOPMENT & ENERGY EFFICIENCY

SMART CITIES & CHALLENGES

COMPUTING IN TELECOMMUNICATIONS

MACHINE LEARNING APPLICATIONS

SIGNAL PROCESSING

BUSINESS PROCESSES & DATABASES

IMPLEMENTATIONS OF STANDARDS & BEST PRACTICES

COMPUTERS IN EDUCATION

PAPERS

COMPUTERS IN TECHNICAL SYSTEMS

PAPERS

ROBOTICS AND AUTOMATION

INTERNET BASED SYSTEMS, APPLICATIONS AND TECHNOLOGIES

SOFTWARE DESIGN AND AUTOMATION IN INDUSTRY

PROFESSIONAL PAPER IN CROATIAN

INTELLIGENT SYSTEMS

PAPERS

ROBOTICS TECHNOLOGIES AND APPLICATIONS

PAPERS

INFORMATION SYSTEMS SECURITY

PAPERS

BUSINESS INTELLIGENCE SYSTEMS

INVITED PAPER

PAPERS

DIGITAL ECONOMY AND DIGITAL SOCIETY

PAPERS

INFORMATION AND COMMUNICATION TECHNOLOGY LAW

PAPERS

ENGINEERING EDUCATION

PAPERS

SOFTWARE AND SYSTEMS ENGINEERING

PAPERS

SOFTWARE SYSTEM ARCHITECTURES

QUALITY ASPECTS

TECHNOLOGIES

APPLICATIONS

PROFESSIONAL PAPER IN CROATIAN

MIPRO JUNIOR - STUDENT PAPERS

PAPERS

SMART, SUSTAINABLE AND RESILIENT CITIES AND INFRASTRUCTURE

PAPERS

OPTOELECTRONICS AND PHOTONICS

PAPERS

DEW COMPUTING

PAPERS

Analysis of Diabetes Dataset

L. Beqiri *, A. Velinov *, B. Fetaji **, L. Loku **, A. Buçuku *, Z. Zdravev *
* University Goce Delcev, Faculty of Computer Science, Stip, Macedonia
** Mother Teresa University, Skopje, Macedonia
e-mail address: lavdim.beqiri@gmail.com, aleksandar.velinov@ugd.edu.mk, bekim.fetaji@unt.edu.mk, lindita.loku@unt.edu.mk, agon.bucuku@gmail.com, zoran.zdravev@ugd.edu.mk

***Abstract* - The focus of the research study was analysis of diabetes dataset and how it will perform if we try to do a prediction of diabetes with different machine learning algorithms. We used the original dataset from the National Institute of Diabetes, and Digestive and Kidney Diseases. The dataset can be used to predict whether or not a patient has diabetes, based on certain diagnostics. For analysis we used Amazon Web Services. We used AWS S3 service to store our dataset, and Amazon Sagemaker to perform an analysis. For the given dataset we applied three classification models: Logistic Regression Model, K-nearest Neighbors and Support Vector Machines. For each of the models we also performed a performance measurement. We also compared all the results we got and according to the results, Support Vector Machines has the best performance. Insights and recommendations are provided.**

Keywords – Diabetes Dataset, Machine Learning Algorithms, Amazon Sagemaker, Logistic Regression Model, K-nearest Neighbors, Support Vector Machines.

I. Introduction

The exponential growth of data has created a new area of interest in technology and business called “Big Data". To be classified as Big Data, a data set or business problem must have data that is so vast, fast or complex that it becomes impossible to store, process, and analyze using traditional data storage and analytics applications [1]. Volume is only one of the criteria because the need for real time processing of the data (also called data in motion) or the need for integrating structured and unstructured data may qualify the problem as Big Data problem [2]. The three characteristics of Big Data are: data itself, analytics and analytics result. Big Data technologies as a new generation of technologies and architectures, designed to economically extract value from very large volumes of a wide variety of data by enabling high-velocity capture, discovery, and/or analysis. Referring to the book by Kalyvas and Overly, the “Big Data” refers to datasets whose size is beyond the ability of typical database software tool to capture, store, manage, and analyze. This definition is intentionally subjective and incorporates a moving definition of how big dataset needs to be in order to be considered Big Data, don’t define Big Data in terms of being larger than a certain number of terabytes [3]. Looking at the trends of technological growth and data growth, he thought that terminology would change in terms of the number of sizes that determine what Big Data is. Big Data is all about analytics on a broader spectrum of data, and therefore represents an opportunity to create even more differentiation among industry peers. Real value can only emerge from a consumable analytics platform that saves you from having to build applications from scratch [4]. In Big Data, we can use the modeling process of large datasets, which uses statistical and artificial intelligence methods. Data mining makes the information structure readable for use in the future. This is done through the process of analyzing the content of the model, and relationships between variables, validating the findings and applying patterns. The use of classification is to extract the model and to describe the classes. For better understanding, the data at large is used for analysis help. Classification predicts categorical, labels, models continuous valued functions. Classification organize and categorize data in distinct classes.

The focus of this paper is to apply different classification algorithms. For the given dataset we applied three classification models: Logistic Regression Model, K-nearest Neighbors and Support Vector Machines.

Using Big Data should allow us to understand the cause of the disease and to help develop the latest therapies. Personalized medicine has great potential to make treatments more effective by reducing side effects, but it requires access for researchers to Big Data in order for it to progress.

According to the World Health Organization (WHO), diabetes is a chronic disease that occurs either when the pancreas does not produce enough insulin or when the body cannot effectively use the insulin it produces. [5]. There are three main types of diabetes mellitus. Type 1 and type 2 diabetes both occur when the body cannot properly store and use glucose, which is primary energy source [6]. Sugar, or glucose, collects in the blood and does not reach the cells that need it, which can lead to serious complications. Type 1 diabetes usually begins in childhood or early adulthood; but its onset can occur in adults. In type 1 diabetes, a person with a genetic predisposition who is exposed to a precipitating event, such as a viral infection, experiences autoimmune destruction of the beta cells. Type 2 diabetes is more likely to appear as people age, but many children are now starting to develop it. In this type, the pancreas produces insulin, but the body cannot use it effectively. There has been a substantial increase in the number of cases of type 2 diabetes diagnosed in young children. [7].

II. RELATED WORK

Analysis of dataset is widely applied and used in various fields, such as weather forecasting, GPS, target marketing, engineering, and medical diagnosis. By analyzing a given dataset, we can extract knowledge from it. In this section, we will present some related work which used medical datasets.

An analysis of diabetes dataset was conducted by Alahmar, Mohammed and Benlamri [8]. The dataset they used includes total 9 features which have numerical and nominal attributes. They targeted to make a classifier system by using several ML techniques. In that analysis, they found 79.13% accuracy using Support Vector Machine (SVM).

Similarly, Mir and Dhage applied ML techniques to classify diabetes dataset [9]. They applied a few machine learning (ML) algorithms, in particular: Naive Bayes, SVM, Random Forest (RF) and Simple CART. Among these algorithms, they found 76.5% accuracy using RF in terms of classification of diabetes dataset.

Verma and Mishra conducted a study to identify diabetes cases by using several ML algorithms: Naive Bayes, J48, Sequential Minimal Optimization (SMO), MLP and Reduced Error Pruning Tree (REP-tree) algorithms [10]. In their study, they found the best accuracy of 76.80% using SMO algorithm.

III. CLASSIFICATION

Classification is a technique for identifying and grouping data in such a way that based on a value of the target attribute, the entire dataset can be qualified to belong to a class. This is one of the techniques used in data mining to identify the data behavior patterns [11].

Classification is two-step process: I) Learning or training step where data is analyzed by a classification algorithm. II) Testing step where data is used for classification and to estimate the accuracy of the classification [12].

A. Model 1: Logistic Regression Model

Logistic Regression is the most common method used for binary classification problems (problems with a 1 or 0 outcome) [13].

The function used in Logistic regression is the logistic function. The logistic function is an S-shaped curve that can take any number and map it into a value between 0 and 1. The equation of the Logistic Function is as follows:

$$\frac{1}{(1 + e^{-value})}$$

Where e is the base of the natural logarithms (Euler's number or the exp() function) and value is the actual numerical value that you want to transform. In Figure 1 we can see a plot of numbers on the x-axis transformed into the range 0 and 1 on the y-axis using the logistic function.

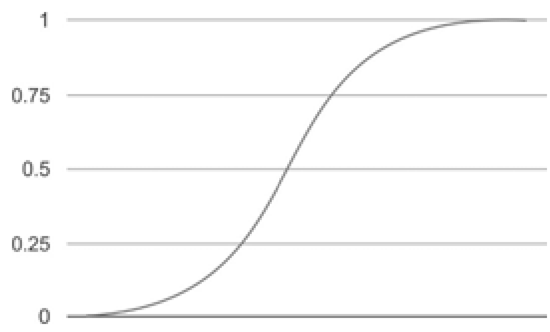

Figure 1. Logistic Function

How is the logistic function used in logistic regression? Input values are combined linearly using coefficient values to predict an output value. How this differs from linear regression is that it is used to model a binary value (0 or 1) instead of a numeric value.

Below is the logistic regression equation:

$$probability = \frac{1}{(1 + e^{-(b_0 + b_1(x))})}$$

b_0 is an intercept term and b1 is a coefficient for the input value (x). Each variable in our input data has a b coefficient that is calculated using the training dataset.

As an example, if we are modelling the probability of a patient having diabetes from a single factor (say blood-glucose level). A logistic regression model could be written as the probability of having diabetes given a person's height blood-glucose level (x).

The above equation can be re-written as follows:

$$ln\left(\frac{probability}{1 - probability}\right) = b_0 + b_1 x$$

The calculation on the right is a linear equation (like linear regression), and the input on the left is a log of a ratio including the probability.

This ratio $\frac{probability}{1-probability}$ is referred to as "**the odds**" of the default class. We could write the above equation as follows:

$$ln(odds) = b_0 + b_1 x$$

usually referred to as ***log odds***

Because we are interested in using the logistic function for a classification problem (deciding if something is 1 or 0), we don't use the probabilities directly. We can convert the probabilities to a binary class value, for example:

- If probability < 0.5 assign: 0
- If probability >= 0.5 assign: 1

B. Model 2: K-nearest Neighbors Algorithm

In order to understand the algorithm, it is useful to take a graphical example. Say we have a distribution of 1's and 0's on a plane as shown below. Now say we have a red square on the same plane and don't know whether it belongs to class 1 or 0.

Figure 2. K nearest neighbours

In the K nearest Neighbors algorithm (Figure 2) the “K” is the amount of near-by points we want to check [14]. Let’s take an example of K=3, we would make a circle with the red square at the center and increase the size of the circle until it surrounds 3 data points (Figure 3).

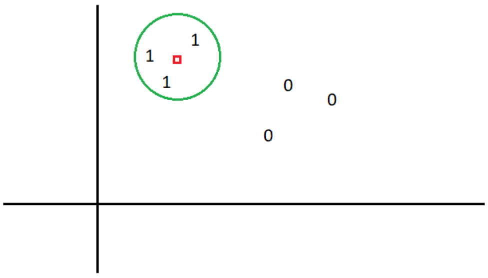

Figure 3. K nearest neighbours (3 points enclosed of class 1)

As we can see, all 3 points which have been enclosed are of class 1. Intuitively we can deduce that our unknown point (the red square) is of class 1.

C. Model 3: Support Vector Machine Algorithm

A Support Vector Machine is another algorithm that can be used in classification problems.

SVMs use the concept of finding a dividing boundary that splits the dataset into two seperate classes. This dividing boundary is referred to as a hyperplane. This is shown graphically in Figure 4.

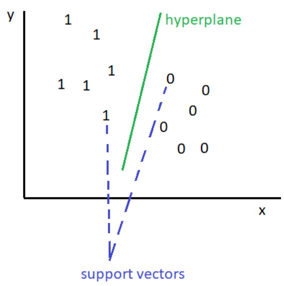

Figure 4. Support Vector Machine

- Support Vectors - The points of a data set that are closest to the dividing boundary as usually referred to as the "support vectors". These "support vectors" are considered to be the critical elements of a data set. If any of these points were removed, the position of the hyperplane would change significantly.
- What is a hyperplane? - Once the hyperplane has been set, if we introduce new testing data, the further the data point is from the hyperplane, the more confident we are that the point belongs to the class represented by that side of the hyperplane.
- How do we find the right hyperplane? - The margin is defined as: The distance between the hyperplane and the nearest data point. To achieve the best chance of new data being classified correctly, choosing a hyperplane with the greatest possible margin between the hyperplane and any point in the training set (Figure 5).

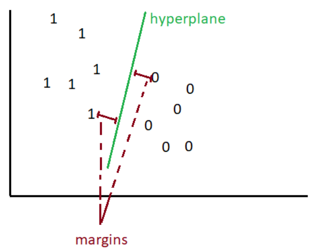

Figure 5. Support Vector Machine (hyperplane, margins)

IV. METHODOLOGY

Our research dataset is divided into two parts, two-thirds of the data is used as a training set, and one-third of the dataset is defined as a testing set to evaluate the performance of several classifiers. All classifiers were fitted to the same training and testing data. The specific process is shown in Figure 6.

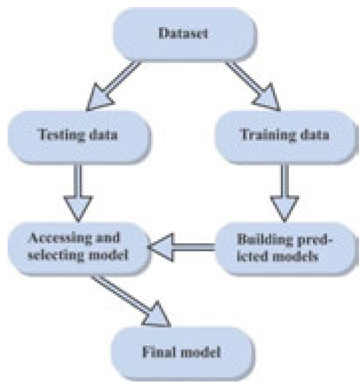

Figure 6. Methodology (flowchart)

A. Dataset

In order to analyze the factors contributing to the presence of diabetes, we sourced a dataset (Figure 7). This original dataset is from the National Institute of Diabetes and Digestive and Kidney Diseases. The dataset can be used to predict whether or not a patient has diabetes, based on certain diagnostics e.g. glucose, insulin levels in the dataset (Table I).

A	B	C	D	E	F	G	H	I
Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	**Outcome**
6	148	72	35	0	33.6	0.627	50	**1**
1	85	66	29	0	26.6	0.351	31	**0**
8	183	64	0	0	23.3	0.672	32	**1**
1	89	66	23	94	28.1	0.167	21	**0**
0	137	40	35	168	43.1	2.288	33	**1**
5	116	74	0	0	25.6	0.201	30	**0**
3	78	50	32	88	31	0.248	26	**1**
10	115	0	0	0	35.3	0.134	29	**0**
2	197	70	45	543	30.5	0.158	53	**1**
8	125	96	0	0	0	0.232	54	**1**
4	110	92	0	0	37.6	0.191	30	**0**

Figure 7. The dataset (contains 768 lines)

The dataset consists of 8 medical independent variables (pregnancies, glucose, blood pressure, skin thickness, insulin, body mass index – BMI, diabetes pedigree function and age) and one target dependent variable (outcome). "Outcome" is a binary target variable that has a value of 1 for diabetes and 0 for no diabetes. 268 of the 768 are 1, the others are 0. All records for patients in the dataset are females greater than 21 years old of Pima Indian heritage [15].

TABLE I. INDEPENDENT VARIABLES

1.	**Pregnancies**	Number of times pregnant
2.	**Glucose**	Glucose concentration a 2 hours in an oral glucose tolerance test
3.	**BloodPressure**	Diastolic blood pressure (mm Hg)
4.	**SkinThickness**	Triceps skin fold thickness (mm)
5.	**Insulin**	2-Hour serum insulin (mu U/ml)
6.	**BMI**	Body mass index (weight in kg/(height in m)^2)
7.	**DiabetesPedigreeFunction**	Diabetes pedigree function
8.	**Age**	Age (years)

B. Tools used

We used the S3 service from AWS that allow creating a bucket to store the dataset. In order to analyze the dataset we used Amazon SageMaker Machine Learning service. The tools used are Jupyter notebooks using Python code.

V. RESULTS OF CLASSIFICATION ALGORITHMS USING AWS

In order to see if our model gives good results at identifying 1's and 0's we check the confusion matrix (Table II and Table III).

TABLE II. CONFUSION MATRIX

True Negative	False Negative
False Positive	True Positive

The Receiver Operator Characteristic (ROC) curve is a graphical output which is used to analyse the accuracy of how good the model is at separating 1's and 0's. The higher the Area Under the Curve (AUC) value, the better the model is at predicting classification problems (predicting a 1 or 0 outcome). The area under the curve represents how 0s as 0s and 1s as 1s.

An excellent model has an AUC near to the 1. When the AUC is 0.5, it means that the model has no ability to separate 1's and 0's.

TABLE III. CONFUSION MATRIX (SIMPLER TERM)

Model predicts a 0: actually a 0 in the dataset	Model predicts a 0: actually a 1 in the dataset
Model predicts a 1: actually a 0 in the dataset	Model predicts a 1: actually a 1 in the dataset

Numerically, the area under the curve value equal to 0.9 represents a 90% probability that the model can distinguish between 1 and 0. The further the ROC curve is from the center straight line, the better in terms of accuracy. The area under this curve is a numerical measure of the test accuracy.

A. Model 1: Logistic Regression Model

The results of diabetes dataset for Logistic Regression Model classifier using AWS are shown in Table IV.

TABLE IV. MODEL 1 RESULTS

True Negative: 445	False Negative: 55
False Positive: 112	True Positive: 156

A more visual way to measure the performance of a binary classifier is the ROC curve (Figure 8). It is created by plotting the true positive rate (TPR) (or recall) against the false positive rate (FPR).

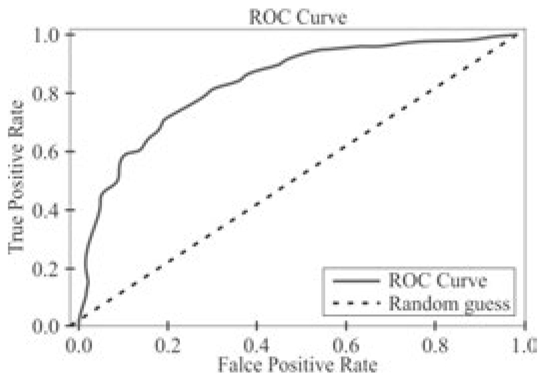

Figure 8. Model 1 - ROC Curve

As we can see the curve is close to the left-hand border and the top border. This indicates that it is quite good at predicting 1's as 1's and 0's as 0's.

We get the value for AUC of **0.83** which is good.

B. Model 2: K-nearest Neighbours Algorithm

The results of diabetes dataset for K-nearest Neighbors Model classifier using AWS are shown in Table V.

TABLE V. MODEL 2 RESULTS

True Negative: 442	False Negative: 58
False Positive: 93	True Positive: 175

Figure 9 shows the ROC curve for the K-nearest Neighbors Model.

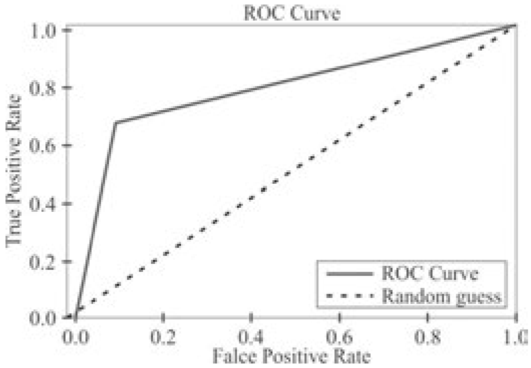

Figure 9. Model 2 - ROC Curve

At a first glance, it does not seem as accurate as the Logistic Regression model. We look at the AUC to check this numerically.

We get a value of **0.768** compared to our result for Logistic regression which yielded a value of 0.839 so this algorithm doesn't perform as well.

C. Model 3: Support Vector Machine Algorithm

The results of diabetes dataset for Support Vector Machine Model classifier using AWS are shown in Table VI.

TABLE VI. MODEL 3 RESULTS

True Negative: 500	False Negative: 0
False Positive: 0	True Positive: 268

The ROC curve for the model is shown in Figure 10. As for previous models, we will analyze the ROC curve and look at the AUC measurement.

We can see that the curve follows the left hand border and the top very closely which indicates excellent accuracy.

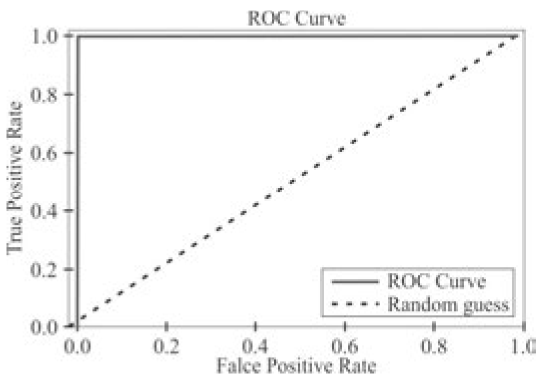

Figure 10. Model 3 – ROC Curve

We get a value of **1.0** which indicates perfect prediction.

VI. DISCUSSION

The results of applying different classification algorithms on diabetes dataset are shown in Table VII, which consists of correctly classified and incorrectly classified instances.

From this study we can see the performance of the three predictive models. For the testing dataset, the final comparative analysis results demonstrated that the Support Vector Machine algorithm proved the best, with an accuracy of 100%, and the results of the Confusion Matrix were True Negative: 500, False Negative: 0, False Positive: 0, and True Positive: 268. The Logistic Regression Model algorithm came out to be the second best with a classification accuracy of 83%, and the results of the Confusion Matrix were True Negative: 445, False Negative: 55, False Positive: 112, and True Positive: 156. Finally, the K-nearest Neighbors algorithm came last with an accuracy of 76.8%, and the results of the Confusion Matrix were True Negative: 442, False Negative: 58, False Positive: 93, and True Positive: 175.

In the results, the area under the receiver operating Characteristic (ROC) curve, Area Under The Curve (AUC) value of the SVM, Logistic Regression Model, and K-nearest Neighbors Algorithm were 1.000, 0.830, 0.768, respectively. The area under ROC curve for SVM method is 100%, showing a high reliability of classification capability among all the methods. Overall, the methodology presented in this paper has shown the good classification performance of health management control satisfaction of patients with diabetes.

VII. CONCLUSION

In this paper, we have classified the data of the diabetes dataset from the National Institute of Diabetes. We applied three classification techniques: Logistic Regression Model, K-nearest Neighbors Algorithm and Support Vector Machine. According to the results, Support Vector Machine is the best algorithm for classification. The dataset that we used in this study is small. In the following research work, we plan to apply algorithms to classify a larger dataset using three different platforms: Amazon Web Services, Microsoft Azure and Google Cloud. In this way we can get statistics, determine the time spent on analysis and compare algorithms.

TABLE VII. RESULTS OF DIFFERENT CLASSIFICATION ALGORITHMS

Algorithms	Value of AUC	Confusion Matrix	
Logistic Regression Model	0.830	True Negative: 445	False Negative: 55
		False Positive: 112	True Positive: 156
K-nearest Neighbours Algorithm	0.768	True Negative: 442	False Negative: 58
		False Positive: 93	True Positive: 175
Support Vector Machine Algorithm	1.00	True Negative: 500	False Negative: 0
		False Positive: 0	True Positive: 268

REFERENCES

[1] C. Razbonyalı and E. Güvenoğlu, "Traditional Data Storage Methods and the Big Data Concepts", International Research

Journal of Engineering and Technology (IRJET), vol. 03, pp. 2556-2561, June 2016.

[2] Y. Riahi, "Big Data and Big Data Analytics: Concepts, Types and Technologies", International Journal of Research and Engineering, vol. 5, pp. 524-528, September 2018.

[3] J. R. Kalyvas, M. R. Overly, Big Data A Business and Legal Guide. CRC Press, Auerbach, 2015.

[4] P. Zikopoulos, D. Deroos, K. Parasuraman. T. Deutsch, D. Corrigan, J. Giles, Harness the Power of Big Data, The McGraw-Hill Companies, 2013.

[5] G. Shahi, B. Luu (Editors), Emerging Trends in Global Health, Global Health Review and GBI Books, 2008.

[6] K. Moscou, K. R. Snipe, Pharmacology for pharmacy technicians, Elsevier, 2018.

[7] K. Cooper, K. Gonsell, Foundations and Adulht Health Nursning, Mosby, 2015.

[8] A. Alahmar, E. Mohammed., R. Benlamri, "Application of data mining techniques to predict the length of stay of hospitalized patients with diabetes", 4th International Conference on Big Data Innovations and Applications (Innovate-Data), Barcelona, Spain, pp. 38-43, 2018.

[9] A. Mir, S. N. Dhage, "Diabetes disease prediction using machine learning on big data of healthcare", Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), Pune, India, pp. 1-6, 2018.

[10] D. Verma, N. Mishra, "Analysis and predicion of breast cancer and diabetes disease datasets using data mining classification techniques", International Conference on Intelligent Sustainable Systems (ICISS), Palladam, India, 533-538, 2017.

[11] S. Gollapudi, Getting Started with Greenplum for Big Data Analytics, Packet Publishing Ltd. 2013.

[12] J. Han, M. Kamber, Data Mining; Concepts and Techniques, Morgan Kaufmann Publishers, 2000.

[13] R. P. Anderson, R. Jin, and G. L. Grunkemeier, "Understanding logistic regression analysis in clinical reports: an introduction", The Annals of thoracic surgery, vol. 75, pp. 753-757, March 2003.

[14] P. Cunningham and S. Delany, k-Nearest neighbour classifiers. Mult Classif Syst, 2007.

[15] (2017) Pima Indians Diabetes Database, [Online]. Available: https://www.kaggle.com/uciml/pima-indians-diabetes-database