JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Diabetes prediction using Various Machine Learning Algorithms

Chintan Rana, Sneha Unnarkar, Krishna Patel, Prof. Sudha Patel.

ABSTRACT:

Diabetes is a disease that is caused by a high glucose level in the human body. Diabetes should not be overlooked; if left untreated, it can lead to serious complications such as heart disease, renal disease, high blood pressure, vision loss, and damage to other organs in the body. If diabetes is detected early enough, it can be managed. To reach this purpose, we will use several machine learning techniques to do early diabetes prediction in a human body or a patient for a higher accuracy. Techniques for machine learning by developing models using patient datasets, you can improve your prediction results. On a dataset, we will utilize Machine Learning Classification and in this study. K-Nearest Neighbor (KNN), Logistic Regression (LR), Decision Tree (DT).Every model has a distinct level of precision.

Keywords: KNN , Decision Tree, Logistic regression, ML.

1 INTRODUCTION:

Diabetes is one of the world's most deadly diseases. Diabetes can be caused by elevated blood glucose levels, for example. It affects the hormone insulin, causing crab metabolism and blood sugar levels to drop. Diabetes occurs when the body does not create enough insulin. Diabetes affects 430 million people globally, with the majority residing in low- and middle-income countries, according to the World Health Organization. This figure might rise to 510 billion by 2035. Despite this, diabetes is common in some nations, including Canada, China, and India. Because India's population has topped 100 million, the country's diabetes rate has increased to 40 million people. Diabetes is one of the main causes of death in the globe.. Early detection of diseases such as diabetes can be controlled and human lives saved. To do so, this study looks into diabetes prediction using a variety of diabetes-related characteristics.we predict diabetes using several Machine Learning classification. Machine learning is a technique for explicitly training computers or machines. By constructing multiple classification from acquired datasets, various Machine Learning Techniques deliver efficient results for collecting knowledge. This type of information can be used to predict diabetes. Various Machine Learning approaches are capable of prediction, however selecting the optimum methodology is difficult. As a result, we use common classification on the dataset to make predictions.

1.1 Machine Learning Algorithms

As per Wikipedia, "Machine learning is the scientific study of algorithms and statistical models that computer systems use to perform a specific task without using explicit instructions, relying on patterns and inference instead. It is seen as a subset of artificial intelligence". In this paper various supervised learning algorithms are used for employee attrition prediction. Following algorithms were applied to predict employee attrition.

A. Logistic Regression

Logistic regression continues to be one of the most widely used methods in data mining in general and binary data classification in particular.It is useful when you wish to perform binary classification. It performs even better if you remove the irrelevant columns from the dataset. Removing irrelevant data from the dataset gives the higher accuracy in the case of logistic regression.

B. KNN (K Nearest Neighbor)

The KNN algorithm is very simple and very effective. KNN means it finds the similar objects or say instances from the given dataset. K is the number of instances.

C. Decision Tree

Decision Trees are an important type of algorithm for predictive modeling machine learning. It is similar to the concept of data structure. Input variable is root and leaf node contains Y or N in case of binary classification.

2 Literature Review

Predicting diabetes onset: an ensemble supervised learning strategy was described by Nonso-Nnamoko et al. [4]. For the ensembles, five widely used classifiers are used, and their outputs are aggregated using a meta-classifier. The findings are reported and compared to other studies in the literature that used the same dataset. Itis demonstrated that diabetes onset prediction can be done more accurately utilising the proposed strategy. Diabetes Prediction Using Machine Learning Techniques, given by Tejas N. Joshi et al. [3], tries to predict diabetes using three different supervised machine learning methods:, Logistic regression, and KNN. This project suggests an excellent method for detecting diabetes illness earlier.Dheeraj Shetty et al. [7] presented diabetes illness prediction using data mining to create the Intelligent Diabetes Disease Prediction System, which provides diabetes malady analysis using a database of diabetes patients.

In this method, KNN (K-Nearest Neighbor) algorithms are used to a diabetic patient database, which is then analyzed using numerous diabetes variables to predict diabetes disease. In their proposed study on diabetes prediction using machine learning algorithms in healthcare, they used three different machine learning algorithms. The algorithms' performance and accuracy are compared and rated. The study's analysis of multiple machine learning techniques reveals which algorithm is most commonly employed to predict diabetes. Researchers are interested in diabetes prediction in order to train an algorithm to correctly classify a dataset and determine if a patient is diabetic or not. According to previous study, the classification technique has not improved considerably. Because diabetes prediction is such an important topic in computers, a system is needed to handle the issues that have been raised based on previous research.

3.Implementation

3.1Dataset Description

For implementation of various machine learning algorithms, a dataset was needed. Here in this paper we have used this URL: https://www.kaggle.com/saurabh00007/diabetescsv. The name of dataset is Diabetes.csv. This dataset consisting of 769 rows and 8 columns.

```
diabetes.head()
```

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1

[Fig1. Dataset columns and its data type]

3.2 Introduction to Google Collab

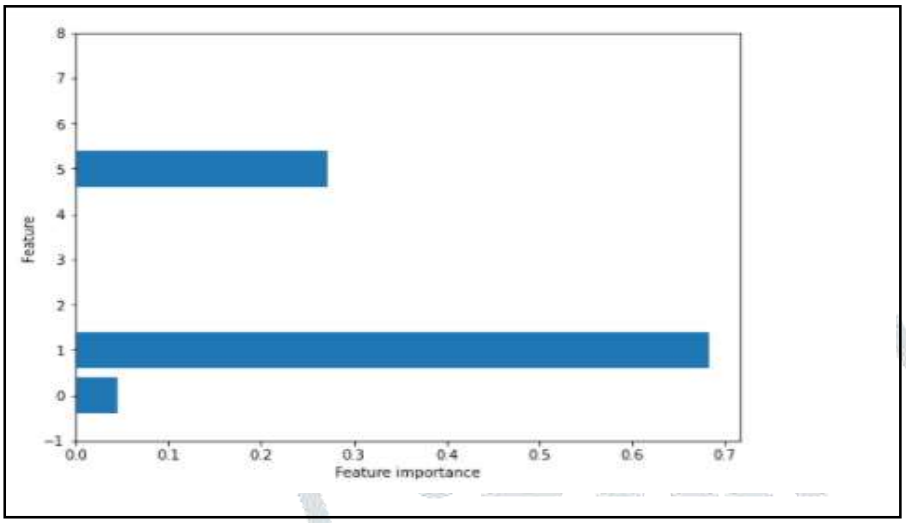
Google collab was utilized to develop multiple machine learning algorithms on the HR dataset. Google is adamant about AI research. Google spent many years developing TensorFlow, an AI framework, and Collaboratory, a development platform. TensorFlow is now open-source, and Google has made Collaboratory available to the public for free since 2017. Google Collab, or just Collab, has replaced Collaboratory. Another appealing feature that google provides to developers is the utilization of GPU. Collab supports GPU and is completely free. One of the motivations for making it freely available to the public could be to make its software a standard in academics for teaching machine learning and data science. It may also have the long-term goal of establishing a customer base for Google Cloud APIs, which are sold on a per-use basis. Regardless of the causes, the introduction of Collab has made machine learning application learning and development easier [14].

3.3 Implementation Steps

Implementation steps are explained using following diagram. Diagram also explains the process which is going to occur in each step. The Fig.2 shows both: process name and process details.

3.4 Identifying feature importance for Diabetes prediction

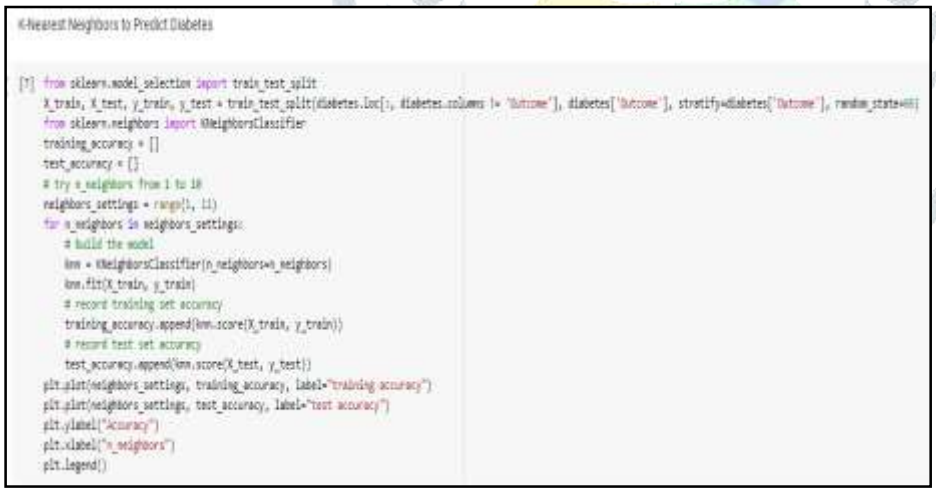
Feature of any project simply means the attribute. This section deals with the data that actually which features are more correlated for diabetes prediction.



[Fig.2 Various features histogram for diabetes importance]

3.5 Implementing ML algorithm

In this paper, logistic regression, Decision tree classifier and K-Nearest neigh bor (KNN) were applied to compare their results and finding out the best suitable algo rithm for Diabetes prediction. With the help of SCI-KITlearn library, various ML algorithms.



[Figure 3 : implementation of KNN]

Decision Tree Classifier

```
[9] from sklearn.tree import DecisionTreeClassifier
    tree = DecisionTreeClassifier(random_state=0)
    tree.fit(X_train, y_train)
    print("Accuracy on training set: {:.3f}".format(tree.score(X_train, y_train)))
    print("Accuracy on test set: {:.3f}".format(tree.score(X_test, y_test)))
```

[Figure 4 Implementation of Decision Tree]

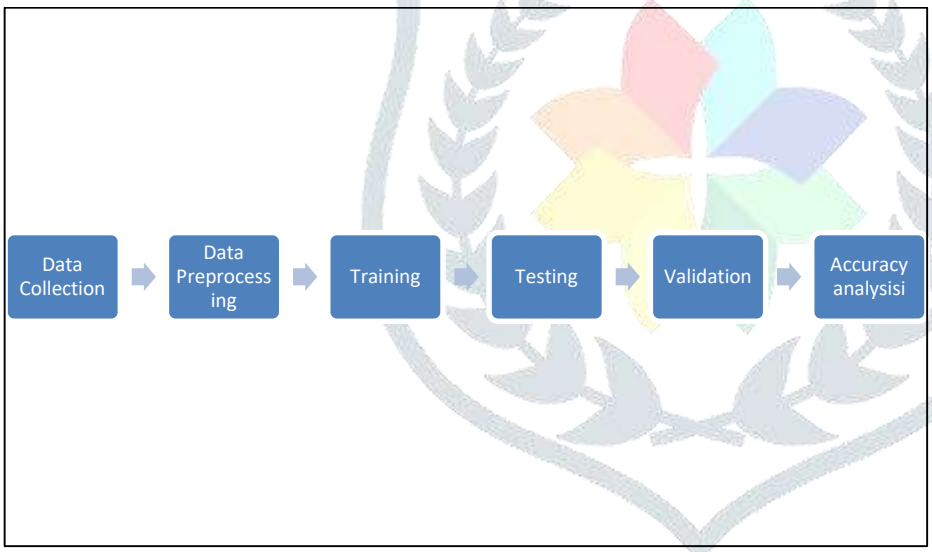
LOGISTIC REGRESSION

```
coeff = list(diabetesCheck.coef_[0])
labels = list(trainData.columns)
features = pd.DataFrame()
features['Features'] = labels
features['importance'] = coeff
features.sort_values(by=['importance'], ascending=True, inplace=True)
features['positive'] = features['importance'] > 0
features.set_index('Features', inplace=True)
features.importance.plot(kind='barh', figsize=(11, 6),color = features.positive.map({True: 'blue', False: 'red'}))
plt.xlabel('Importance')
```

[Figure 5 Implementation of logistic regression]

3.6 Machine learning Model

This is most important phase which includes model building for prediction of diabetes. In this we have implemented various machine learning algorithms which are discussed above for diabetes prediction.



[Figure 6 : ML model]

IMPLEMENTATION STEPS

Procedure of Proposed Methodology

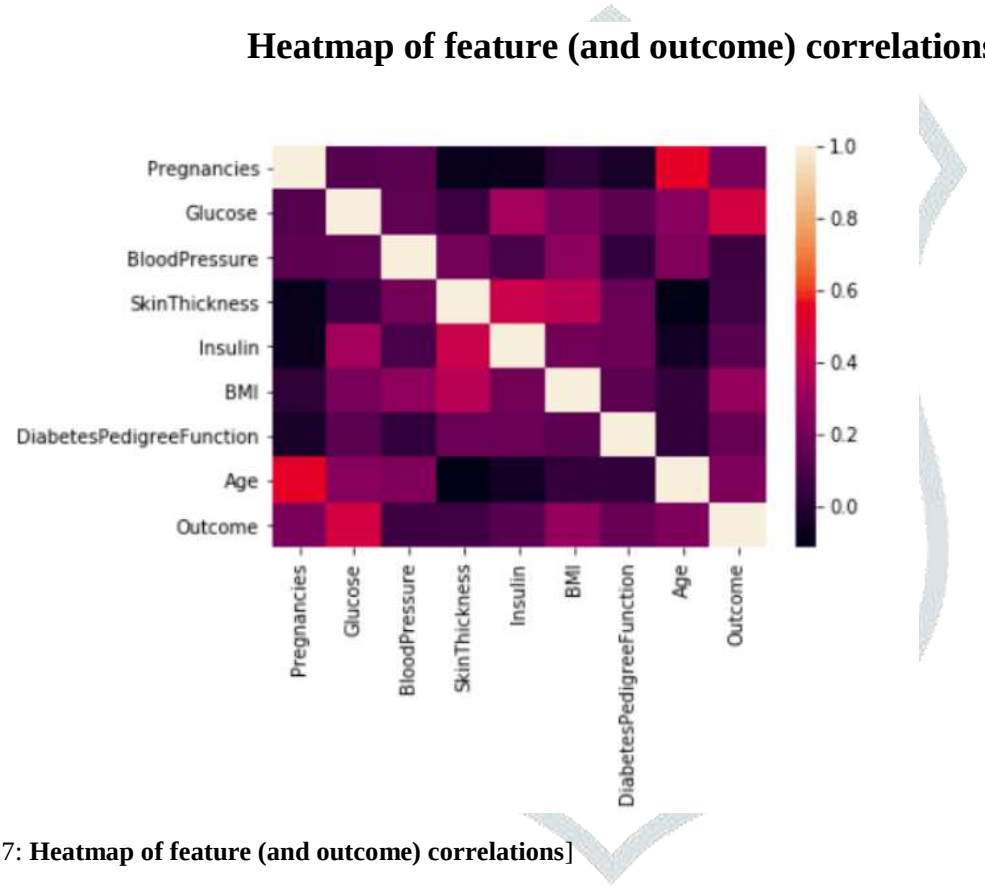
1: Import required libraries and dataset.
2: Pre-Process for remove missing data
3. Select 80 % for training data and select 20% for testing data
4: Choose the machine learning algorithm like- K-Nearest Neighbour.
Decision Tree, Logistic regression algorithm.
5: Build the classifier model for selected machine learning algorithm based on training set.
6: Test the Classifier model for the selected machine learning algorithm based on test set.
7: Perform Comparison Evaluation of the experimental performance results obtained for each classifier.
8: After analyzing based on various measures conclude the best performing algorithm.

4.Result Analysis

In this work different process were taken. The proposed approach uses different classification and implemented using python. These methods are standard Machine Learning methods used to obtain the best accuracy from data. In this work we see that KNN achieves better compared to others. Over-all we have used best Machine Learning techniques for prediction and to achieve high performance accuracy. Figure shows the result of these Machine Learning methods.

4.2Feature correlation Analysis

Here feature played very important role in in this diabetes prediction for machine learning The sum of the importance of all feature playing major role for diabetes have been described, where X-axis represents the importance of all feature and Y-Axis the names of the features.



[Figure 7: **Heatmap of feature (and outcome) correlations**]

MAIN FEATURES OF DIABETES PREDICTIONS:

1. Pregnancies
2. Glucose
3. Blood-Pressure
4. SkinThikness
5. Insulin
6. BMI
7. DiabetesPedigreeFuction
8. Age

5 Best algorithm Analysis

[Table 1: Various Algorithm analysis for Diabetes prediction]

No	Algorithm	Accuracy on training	Accuracy on testing
1	Logistic Regression	0.72	0.77
2	KNN	0.79	0.78
3	Decision Tree Classifier	1.00	0.714

6 Conclusion :

The main focus of this project was to implement Diabetes Prediction Using Machine Learning Methods .The proposed approach uses various classification algorithms on KNN, Decision Tree and Logistic Regression are used. 78% classification accuracy had been achieved. The Experimental results can be asst health care to take early prediction and make early decision to cure diabetes and save humans life. In this project KNN have higher testing accuracy.

7 References:

1] Debadri Dutta, Debpriyo Paul, Parthajeet Ghosh, "Analyzing Feature Im portance's for Diabetes Prediction using Machine Learning". IEEE, pp 942-928, 2018.

[2] Md. Faisal Faruque, Asaduzzaman, Iqbal H. Sarker, "Performance Analysis of Machine Learning Techniques to Predict Diabetes Mellitus". Interna tional Conference on Electrical, Computer and Communication Engineer ing (ECCE), 7-9 February, 2019.

[3] Tejas N. Joshi, Prof. Pramila M. Chawan, "Diabetes Prediction Using Machine Learning Techniques".Int. Journal of Engineering Research and Ap

[4] Nonso Nnamoko, Abir Hussain, David England, "Predicting Diabetes Onset: an Ensemble Supervised Learning Approach ". IEEE Congress on Evolutionary Computation (CEC), 2018.

[7] Dheeraj Shetty, Kishor Rit, Sohail Shaikh, Nikita Patil, "Diabetes Disease Prediction Using Data Mining ".International Conference on Innova tions in Information, Embedded and Communication Systems (ICIIECS), 2017.

[8] Nahla B., Andrew et al,"Intelligible support vector machines for diagno sis of diabetes mellitus. Information Technology in Biomedicine", IEEE Transactions. 14, (July. 2010), 1114-20.

[9] A.K., Dewangan, and P., Agrawal, "Classification of Diabetes Mellitus Using Machine Learning Techniques," International Journal of Engineer ing and Applied Sciences, vol. 2, 2015.

[10]A comprehensive review of machine learning techniques on diabetes de taction Tahira Sharma1 and Manan Shah2*

[11] N.P. Bhensadadiya, D.Bosamiya,"Survey On Various Intelligent Traffic Management Schemes For Emergency Vehicles", International Journal on Recent and Innovation Trends in Computing and Communication.

[12] Ashish Bhagchandani, Dulari Bhatt on "Machine Learning Model For Predicting Social Media Influence On Sports", Proceedings of TheIRES
International Conference, Abu Dhabi, UAE, 27th – 28th April, 2020

[13] Bhatt D, Patel C, Talsania H, Patel J, Vaghela R, Pandya S, Modi K, Ghayvat H. CNN Variants for Computer Vision: History, Architecture, Appli cation, Challenges and Future Scope. *Electronics*. 2021; 10(20):2470. https://doi.org/10.3390/electronics10202470

[14] https://colab.research.google.com/?utm_source=scs-index

[15] Umair Muneer Butt, Sukumar Letchmunan, Mubashir Ali, Fadratul Hafinaz Hassan, Anees Baqir, Hafiz Husnain Raza Sherazi, "Machine Learning Based Diabetes Classification and Prediction for Healthcare Applications", *Journal of HealthcareEngineering, vol. 2021, Article ID 9930985, 17 pages, 2021. https://doi.org/10.1155/2021/9930985.*

[14] Gauri D. Kalyankar, Shivananda R. Poojara and Nagaraj V. Dharwadkar," Predictive Analysis of Diabetic Patient Data Using Machine
Learning and Hadoop", International Conference On I-SMAC,978-1-5090-3243-3,2017.

[15] Ayush Anand and Divya Shakti," Prediction of Diabetes Based on Personal Lifestyle Indicators", 1st International Conference on Next
Generation Computing Technologies, 978-1-4673-6809-4, September 2015.

[16] B. Nithya and Dr. V. Ilango," Predictive Analytics in Health Care Using Machine Learning Tools and Techniques", International Conference
on Intelligent Computing and Control Systems, 978-1-5386-2745-7,2017.

[17] Dr Saravana kumar N M, Eswari T, Sampath P and Lavanya S," Predictive Methodology for Diabetic Data Analysis in Big Data", 2nd

International Symposium on Big Data and Cloud Computing,2015.
[18] Aiswarya Iyer, S. Jeyalatha and Ronak Sumbaly," Diagnosis of Diabetes Using Classification Mining Techniques", International Journal of
Data Mining & Knowledge Management Process (IJDKP) Vol.5, No.1, January 2015.
[19] P. Suresh Kumar and S. Pranavi "Performance Analysis of Machine Learning Algorithms on Diabetes Dataset using Big Data Analytics",
International Conference on Infocom Technologies and Unmanned Systems, 978-1-5386-0514-1, Dec. 18-20, 2017.
[20] Mani Butwall and Shraddha Kumar," A Data Mining Approach for the Diagnosis of Diabetes Mellitus using Random Forest Classifier",
International Journal of Computer Applications, Volume 120 - Number 8,2015.
[21] K. Rajesh and V. Sangeetha, "Application of Data Mining Methods and Techniques for Diabetes Diagnosis", International Journal of
Engineering and Innovative Technology (IJEIT) Volume 2, Issue 3, September 2012.
[22]Humar Kahramanli and Novruz Allahverdi,"Design of a Hybrid System for the Diabetes and Heart Disease", Expert Systems with
Applications: An International Journal, Volume 35 Issue 1-2, July, 2008.
[23] B.M. Patil, R.C. Joshi and Durga Toshniwal,"Association Rule for Classification of Type-2 Diabetic Patients", ICMLC '10 Proceedings of
the 2010 Second International Conference on Machine Learning and Computing, February 09 - 11, 2010.
[24] Dost Muhammad Khan1, Nawaz Mohamudally2, "An Integration of K-means and Decision Tree (ID3) towards a more Efficient Data
Mining Algorithm ", Journal Of Computing, Volume 3, Issue 12, December 2011.
[25] Gauri D. Kalyankar, Shivananda R. Poojara and Nagaraj V. Dharwadkar," Predictive Analysis of Diabetic Patient Data Using Machine
Learning and Hadoop", International Conference On I-SMAC,978-1-5090-3243-3,2017.
[26] Ayush Anand and Divya Shakti," Prediction of Diabetes Based on Personal Lifestyle Indicators", 1st International Conference on Next
Generation Computing Technologies, 978-1-4673-6809-4, September 2015.
[27] B. Nithya and Dr. V. Ilango," Predictive Analytics in Health Care Using Machine Learning Tools and Techniques", International Conference
on Intelligent Computing and Control Systems, 978-1-5386-2745-7,2017.
[28] Dr Saravana kumar N M, Eswari T, Sampath P and Lavanya S," Predictive Methodology for Diabetic Data Analysis in Big Data", 2nd
International Symposium on Big Data and Cloud Computing,2015.
[29] Aiswarya Iyer, S. Jeyalatha and Ronak Sumbaly," Diagnosis of Diabetes Using Classification Mining Techniques", International Journal of
Data Mining & Knowledge Management Process (IJDKP) Vol.5, No.1, January 2015.
[30] P. Suresh Kumar and S. Pranavi "Performance Analysis of Machine Learning Algorithms on Diabetes Dataset using Big Data Analytics",
International Conference on Infocom Technologies and Unmanned Systems, 978-1-5386-0514-1, Dec. 18-20, 2017.
[31] Mani Butwall and Shraddha Kumar," A Data Mining Approach for the Diagnosis of Diabetes Mellitus using Random Forest Classifier",
International Journal of Computer Applications, Volume 120 - Number 8,2015.
[32] K. Rajesh and V. Sangeetha, "Application of Data Mining Methods and Techniques for Diabetes Diagnosis", International Journal of
Engineering and Innovative Technology (IJEIT) Volume 2, Issue 3, September 2012.
[33]Humar Kahramanli and Novruz Allahverdi,"Design of a Hybrid System for the Diabetes and Heart Disease", Expert Systems with
Applications: An International Journal, Volume 35 Issue 1-2, July, 2008.
[34] B.M. Patil, R.C. Joshi and Durga Toshniwal,"Association Rule for Classification of Type-2 Diabetic Patients", ICMLC '10 Proceedings of
the 2010 Second International Conference on Machine Learning and Computing, February 09 - 11, 2010.
[35] Dost Muhammad Khan1, Nawaz Mohamudally2, "An Integration of K-means and Decision Tree (ID3) towards a more Efficient Data
Mining Algorithm ", Journal Of Computing, Volume 3, Issue 12, December 2011.
[36]. Komi, Zhai. 2017. Application of Data
Mining Methods in Diabetes Prediction
[37]. Analysis of Various Data Mining Techniques
to Predict Diabetes Mellitus, Omar Kassem
Diabetes Prediction using Machine Learning Techniques.
Khalil Aissa Boudjella, 2016 Sixth