I. Introduction
Diabetes is a chronic condition that occurs when the human body cannot produce enough use of insulin [15]. The high blood sugar is due to two reasons. The first is either the pancreas does not produce enough insulin; or another reason is cells do not respond to the insulin that is produced [16].

When the glucose level in body is not metabolized properly, it keeps on circulating in the blood and causes damage to various tissues. High blood glucose level in diabetes is associated with long-term damage and dysfunction, and failure of various organs, especially the kidneys, eyes, nerves, heart, and blood vessels. Diabetes can mainly be of 3 types: type-1 diabetes, type-2 diabetes, and gestational diabetes. Type-1 diabetes results from nonproduction of insulin. Type-2 diabetes results from the development of resistance of insulin, as a result of which insulin produced is not able to metabolize the sugar levels properly.

Gestational diabetes occurs in pregnant women, who develop a high blood glucose level during pregnancy. It may be preceded by the development of type-2 diabetes [15]. Recognizing a type of diabetes to an individual is mostly based on the situation at the time of diagnosis. It is observed that most of the diabetic patients do not just fix into a single category [17] [18].

Data mining is the process of analyzing data from different perspectives and summarizing it into useful information that can be used for industrial, medical and scientific purposes [16]. Data mining methods is one of the best ways to utilize large volumes of available diabetes-related data for extracting knowledge and making decisions [19]. At present the results of early prediction of diabetes are not highly accurate. Therefore, there is a need to develop a diagnosis system for diabetes with better accuracy.

II. Data Mining In Diabetic Mellitus
Data mining involves computational process, machine learning, statistical techniques, classification, clustering and discovering patterns processes to extract useful information from large volume of database [20].

Medical data mining is used in the knowledge acquisition and analyses the information obtained from research reports, medical reports, flow charts, evidence tables, and transform these mounds of data into useful information for decision making. The present work has taken up to analyze the obtained data of...
diabetic patients by various data mining classification algorithms which can be helpful for medical analysts or practitioners for accurate diabetes diagnosis [16].

Our current work mainly focuses on type-2 diabetic patients. The patient information of type-2 diabetes perspective used to find out classification algorithm's accuracy and error rate (ER) [21]. This has done using Weka comprehensive classification techniques like linear regression model, Gaussian Process, Multilayer Perceptron, Simple Linear Regression and SMOReg.

This paper presents an analytical study on the existing techniques available for diabetes mellitus. The characteristic of the approaches are investigated to develop a better approach for the early and efficient diagnosis of the disease[18].

This study implement classification techniques in data mining to assist medical institutions, medical research centers and laboratory with predicting the human beings diabetic mellitus status. If the persons are predicted to have a chance to affect by the diabetic mellitus, then extra exertion can be made to improve their health conditions and suggesting the necessary steps to be taken to protect their health [22][23]. There are numerous factors causing diabetics in men and women are getting treated.

III. Literature Review

DeeptiSisodiaa, Dilip Singh Sisodia studied to design a model which can prognosticate the likelihood of diabetes in patients with maximum accuracy. Therefore three machine learning classification algorithms namely Decision Tree, SVM and Naive Bayes are used. The performances of all the three algorithms are evaluated on various measures like Precision, Accuracy, F-Measure, and Recall. Accuracy is measured over correctly and incorrectly classified instances. Results obtained show Naive Bayes performs with the highest accuracy of 76.30% compare with other algorithms [1].

Rafiqul Islam, ObaidurRahman, find out the risk factors of Type 2 diabetic patients in Bangladesh. To fulfill this objective, chi square test and logistic regression analysis have been used. It is found that diabetes affects more in the age 35 years and over which is 89.7% in which 68.3% have type 2 diabetes. Again, 79.3% of diabetic patients have type 2 diabetes in which females (43.7%) are more affected than males (35.7%). It has been found that age, controlling diabetic through exercise, controlling diabetic through taking medicine and living house of the respondents are significantly associated with the type 2 diabetes of diabetic patients [2].

S.UmmugulthumNatchiar, S.Baulkani, observed that various computational intelligent techniques are used for diabetes classification. Among those rules based classification algorithms are highly used by the researchers for medical diagnosis. Since these decision rules are easily interpretable and understandable. They suggested that for decision rule generation rough set theory can also be used to handle noisy, missing and uncertainty data [3].

Koushik Chandra Howlader, Md. ShahriareSatu y, AvijitBarua, Mohammad Ali Moni, evaluated classification outcomes of decision tree classifiers and found the best decision tree model from them. In this work, CDT unpruned tree shows highest accuracy, precision, recall, f-measure, second highest AUROC and lowest RMSE than other models. Then, they extracted possible rules and significant features from this model and plasma glucose, plasma glucose 2hr after glucose and HDL-cholesterol have been found as the most significant features to predict the severity of Diabetes Mellitus [4].

S Alby, BL Shivakumar, tried to develop a method for the prediction of type 2 diabetes using adaptive neuro-fuzzy interface system (ANFIS) with genetic algorithms (GA). A comparative study has also been done with the result of our previous work in which General Regression Neural Network (GRNN) is applied. ANFIS with GA was applied for predicting diabetes disease and the most accurate result was obtained compared to our previous work in which we have tried the same using GRNN. The results strongly suggest that ANFIS with GA can aid in the field of diabetes disease. Optimization technique like Genetic Algorithm has much more influence on the accuracy of classification techniques. A combination of ANFIS and GA gave better results than the results achieved by using GRNN [5].

Tejas N. Joshi, Prof. Pramila M. Chawan, developed a system which can perform early prediction of diabetes for a patient with a higher accuracy by combining the results of different machine learning techniques. To predict diabetes via three different supervised machine learning methods including: SVM, Logistic regression, ANN. They propose an effective technique for earlier detection of the diabetes disease. The technique may also help researchers to develop an accurate and effective tool that will reach at the table of clinicians to help them make better decision about the disease status [6].

ChaithraNandMadhu B. analyses the various data mining techniques introduced in recent years to design a predictive model for cardiovascular diseases from the data obtained by transthoracic echo cardiographs.
This study investigates three different classification models: J48 Decision Tree, Naive Bayes and Neural Network on cardiovascular disease prediction and the same has been justified with the results of different experiments conducted and the performance of the models was evaluated using the standard metrics of Accuracy, Precision, Recall and F-measure [7].

Chandeep Kaur and Olufemi Muibi Omisakin, evaluates different data mining methods capable of analyzing big clinical data. To extract information from clinical trial data mining methods to improve on them analytically, and above all to make a recommendation for the best possible data mining and analytical technique. This research was done as the number of cases of diabetes, especially Type 2 diabetes has increased recently as Type 2 diabetes accounts for nearly 90% of all cases of diabetes. This research evaluated a lot of data mining methods, but it is concluded that Big Data proves to be very efficient and effective as it can handle structured and unstructured data like medical records; diagnostic images etc. and interpret its data during the analysis step [8].

M.N. Sohail, Renjiadong, M.M. Uba, M. Irshad, Musavir Bilal, Usman Akbar, Tahir Rizwan examined machine learning classification and logistic regression modeling used to forecast the outcome of dataset through classification. They compared the classification analysis of previous researchers work by weka classification of experimenter part and auto-weka guide for better predictions. They establish a forecast prediction model on base of linear regression with the classification method by using weka performed on the real-life diabetes data set in order to acknowledge an education session for the medical specialists for the initial care of patients. Our results are based on the past researchers experience algorithms to perform the classification and we have pulled the relaxation accordance to free room of improvement in accuracy ratios and forecast analysis [9].

P. Suresh Kumar and V. Umatejaswi, tries to diagnose diabetes using k-means algorithm is used for gestational diabetes, cluster-1 for type-1 diabetes (juvenile diabetes), cluster-2 for type-2 diabetes. From this study, it was found out that data mining techniques can be used for predicting the type and risk levels of diabetes. Thus the proposed model helps in improving the diagnosis of the diseases which indeed helps in early cure of disease in the patients [10].

Mr. R. Sengamuthu, Mrs. R. Abirami, Mr. D. Karthik, explores the early prediction of diabetes using various data mining techniques. The analysis proves that Modified J48 Classifier provide the highest accuracy than other techniques. In the analysis of data mining techniques and tools Modified J48 Classifier gives 99.87% of highest accuracy using WEKA & MATLAB tool [11].

S.Sibi, S.Pushpalatha, provide a detailed survey on the use of data mining techniques in the field of diabetics. They also provide a survey on the recent and most useful research in diabetics using data mining techniques. Predicting diabetics using the risk factor also have certain drawbacks as the symptoms related to diabetics also is common with other diseases and so there is a need for a better data mining model that can effectively identify the occurrence of diabetics. This can be done by building an effective model that takes into consideration all the factors related to diabetics disorder. This survey provides a path for future research in application of data mining in diagnosis of diabetics [12].

D. Jeevanandhini, E. Gokul Raj, V. Dinesh Kumar, N. Sasipriyaa, compared the performance of the algorithm using the data mining techniques. K-means clustering algorithm and Classification algorithm like SVM, KNN, J48, Random forest. They obtained that SVM has higher accuracy comparing to other three algorithms [13].

IV. Methodologies

Figure 1 represents framework of working methodology. The methodology work starts with problem definition followed by data collection and pre-processing steps. The pre-processing part includes data selection and data transformation. After pre-processing it precedes with classification techniques with pruning which finally leads to discovering knowledge which is to be benefitted. [22]
In order to research the high-risk group of Data Mining, we need to utilize advanced information technology. Therefore, data mining, also known as knowledge Discovery Database (KDD), is defined as the computational process of discovering patterns in large datasets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database system.

Table 4.1: Descriptive Statics

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Use</th>
<th>Formula</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic Regression</td>
<td>LR is useful for situations in which can be able to predict the presence or absence of a characteristic or outcome based on values of set of predictor variables.</td>
<td>( H(Y = 1) = \frac{1}{1 + e^{-(\alpha + \beta_1BPAD + \beta_2BWAN + \beta_3BWAN^2 + \beta_4BMAN^2 + 0.5552)}} )</td>
<td>-1.7122 * BPAD + 1.3842 * BWAN + -1.1221 * BWAN2 + 0.0109 * BMAN2 + 0.5552</td>
</tr>
<tr>
<td>GAUSSIAN PROCESS CLASSIFIER</td>
<td>It is a collection of random variables, any infinite number of which have a joint Gaussian distribution [6]. The Gaussian distribution is over vector, whereas the Gaussian process is over functions.</td>
<td>Inverted Covariance Matrix * Target-value Vector: Lowest Value = -0.6409469815752027 Highest Value = 0.37328157457484096</td>
<td>Inverted Covariance Matrix * Target-value Vector: Lowest Value = -0.6409469815752027 Highest Value = 0.37328157457484096</td>
</tr>
<tr>
<td>Multilayer Perception</td>
<td>Multilayer Perceptron (MLP) is gained their popularity due to it is a simple architecture but a powerful problem-solving ability.</td>
<td>( E(w;l) = \frac{1}{2} \sum_{i=1}^{d} \left( y(w(x_i); w) - d_i \right)^2 )</td>
<td>Threshold and Node 5 grouping dataset</td>
</tr>
<tr>
<td>Simple Linear Regression</td>
<td>It is a model that assumes a linear relationship between the input variables (x) and the single output variable (y). When there is a single input variable (x), the method is referred to as <strong>simple linear regression</strong>.</td>
<td>( y_i = \alpha + \beta * X_i + \varepsilon_i )</td>
<td>Linear regression on BWAN 0.33 * BWAN + 0.56</td>
</tr>
</tbody>
</table>
SMOreg

SMO implements the sequential minimal optimization algorithm for training a support vector classifier using polynomial or Gaussian kernels (Platt, J. 1998 and S. Keerthiet. al., 2001). Missing values are replaced globally, nominal attributes are transformed into binary ones, and attributes are normalized by default.

Predicted price = +0.2209 * (normalized) SEX - 0.3164 * (normalized) GRADE + 0.3937

Number of kernel evaluations: 5050 (87.422% cached)

V. Results and Analysis
Authors have used five different algorithms: linear regression model, Gaussian Process, Multilayer Perception, Simple Linear Regression and SMOreg. These processes were selected for inclusions in this study because of their popularity in the recently published literature as well as their good performance even in fewer amounts of training data. Below given is a short description about the selected algorithms for study.

We used different data mining classification algorithms for making predictions in the domain of medical diagnosis and analyze their efficiency. For analyzing the efficiency of the system, the total data set collected was divided into two parts viz training data and test data [15].

Authors takeMetadata Created on March 30, 2017 and Metadata Updated Date on April 11, 2018 from Allegheny County / City of Pittsburgh / Western PA Regional Data Center publisher and maintain by Lynda Jones. Source Schema Version of dataset has 1.1 version in Health Category.

Table 5.1 Used dataset Categories

<table>
<thead>
<tr>
<th>Code</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>CT</td>
<td>Census Tract</td>
</tr>
<tr>
<td>BPAD</td>
<td>Total number of members, who are Allegheny County residents and who were effectively enrolled for at least 90 continuous days during the 2015 calendar year (Jan 1 - Dec 31).</td>
</tr>
<tr>
<td>BPAN</td>
<td>(Diagnosed): Total number of members who met the denominator criteria and are diagnosed with Type 2 diabetes mellitus.</td>
</tr>
<tr>
<td>BPAN 2</td>
<td>(Diagnosed &amp; Medicated): Total number of members who met the denominator criteria, who are diagnosed with Type 2 diabetes mellitus, and had at least one Type 2 diabetes related medication claim.</td>
</tr>
<tr>
<td>BWAD</td>
<td>Total number of women members, who are Allegheny County residents and who were effectively enrolled for at least 90 continuous days during the 2015 calendar year (Jan 1 - Dec 31).</td>
</tr>
<tr>
<td>BWAN</td>
<td>(Diagnosed): Total number of women members who met the denominator criteria and are diagnosed with Type 2 diabetes mellitus.</td>
</tr>
<tr>
<td>BWAN 2</td>
<td>(Diagnosed &amp; Medicated): Total number of women members who met the denominator criteria, who are diagnosed with Type 2 diabetes mellitus, and had at least one Type 2 diabetes related medication claim.</td>
</tr>
<tr>
<td>BMAD</td>
<td>Total number of male members, who are Allegheny County residents and who were effectively enrolled for at least 90 continuous days during the 2015 calendar year (Jan 1 - Dec 31).</td>
</tr>
<tr>
<td>BMAN</td>
<td>(Diagnosed): Total number of male members who met the denominator criteria and are diagnosed with Type 2 diabetes mellitus.</td>
</tr>
<tr>
<td>BMAN 2</td>
<td>(Diagnosed &amp; Medicated): Total number of male members who met the denominator criteria, who are diagnosed with Type 2 diabetes mellitus, and had at least one Type 2 diabetes related medication claim.</td>
</tr>
</tbody>
</table>

Table 5.1 was identifying the dataset columns which can identify all the values of data description about diabetes medication. These categories of data help to analyses and Determination of health care diseaselike diabetes medication. On Basis of these dataset, we made one graphical representation, to identify the different ration of women and male category diagnosed with Type 2 diabetes mellitus.
These influencing factors are categorized as input variables. For this work, recent real world data is collected from Allegheny County / City of Pittsburgh / Western PA Regional Data Center. This data is then filtered out using manual techniques like Stratified Remove Folds, Attribute Selection, Add Classification and Unfiltered Data. Authors analyzed this filter to identifying different values with its parameters like Minimum value, Maximum value, Mean and Standard Deviation value.

Authors apply function type classification data mining techniques like Linear Regression Model, Gaussian Processes, Multilayer Perceptron, Simple Linear Regression and SMOreg with 100 Instances, 10 attributes with 10-fold cross-validation in Test mode.

Table 5.2: Descriptive Statics

<table>
<thead>
<tr>
<th>Classifier function</th>
<th>Time taken to build model</th>
<th>Correlation coefficient</th>
<th>Mean absolute error</th>
<th>Root mean squared error</th>
<th>Relative absolute error</th>
<th>Root relative squared error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear Regression Model</td>
<td>0.02 seconds</td>
<td>0.3957</td>
<td>0.231</td>
<td>0.2764</td>
<td>87.927 %</td>
<td>90.5877 %</td>
</tr>
<tr>
<td>Gaussian Processes</td>
<td>0.05 seconds</td>
<td>0.3126</td>
<td>0.2394</td>
<td>0.2899</td>
<td>91.122 %</td>
<td>95.0031 %</td>
</tr>
<tr>
<td>Multilayer Perceptron</td>
<td>0.47 seconds</td>
<td>0.2207</td>
<td>0.2711</td>
<td>0.3229</td>
<td>103.1621 %</td>
<td>105.8187 %</td>
</tr>
<tr>
<td>Simple Linear Regression</td>
<td>0.01 seconds</td>
<td>0.1545</td>
<td>0.2514</td>
<td>0.2984</td>
<td>95.665 %</td>
<td>97.7956 %</td>
</tr>
<tr>
<td>SMOreg</td>
<td>0.05 seconds</td>
<td>0.3216</td>
<td>0.2214</td>
<td>0.3069</td>
<td>84.2451 %</td>
<td>100.5754 %</td>
</tr>
</tbody>
</table>
We can identify from this algorithm with different categories. Simple Linear Regression is taken to build a model in 0.01 seconds, Correlation coefficient with low values 0.1545, Mean absolute error in 0.2514, Root mean squared error in 0.2984, Relative absolute error in 95.665% and Root relative squared error in 97.7956%. So from this analysis we analyzed that Simple Linear Regression generates the best result for this type of Dataset.

The results suggest that these algorithms can perform good prediction with least error in the field of medical diagnosis and can be used to build various tools for prognosis and thus can supplement medical doctors in performing expert diagnosis. A limitation of this study is that the data set under consideration limits to physiological parameters only, which can be increased to include clinical parameters also, thus the comparative study of prediction using clinical parameters can be carried out [15].

VI. Conclusion and Future Task
In this paper, function type classification techniques are used for prediction on the training dataset of 390 values, to predict and analyzed disease’s performance. In this study, based on selected health care related input variables collected from real world. Among all data mining classifiers Simple Linear Regression performs best with High accuracy and therefore SLR proves to be potentially effective and efficient classifier algorithm. Also comparison of all 5 classifiers with the help of WEKA experiments is also performed, in this case also SLR proves to be the best in higher than other classifiers. For the future research work, it is suggested that for decision rule generation rough set theory can also be used to handle noisy, missing and uncertainty data.

References
5. S Alby, BL Shivakumar, A prediction model for type 2 diabetes using adaptive neuro-fuzzy interface System, Biomedical Research 2018; Special Issue:S69-S74, ISSN 0970-938X
8. ChandeepKaur and OlufemiMuibiOmisakin, Data Mining Methods to Improve Clinical Trials in Diabetic Patients, IMedPub Journals, Annals of Clinical and Laboratory Research, ISSN 2386-5180, Vol.6 No.4:266 2018


11. Mr. R. Sengamuthu, Mrs. R. Abrami 2, Mr. D. Karthik, VARIOUS DATA MINING TECHNIQUES ANALYSIS TO PREDICT DIABETES MELLITUS, International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056 Volume: 05 Issue: 05 May-2018 www.irjet.net p-ISSN: 2395-0072


13. D. Jeevanandhini, E. Gokul Raj, V. Dinesh Kumar, N. Sasipriya, Prediction of Type2 Diabetes Mellitus Based on Data Mining, International Journal of Engineering Research & Technology (IJERT) ISSN: 2278-0181


15. AbidSarwar and Vinod Sharm, Comparative analysis of machine learning techniques in prognois of type II diabetes, ISSN 0951-5666, Volume 29, Number 1, 16 April 2013, pp. 123–129.


