A Machine Learning Framework to Identify the Causes of HbA1c in Patients With Type 2 Diabetes Mellitus

Anar Taghiyev*, Alpaslan A. Altun*, Novruz Allahverdi, Sona Caglar*****

[*]Department of Computer Engineering, Selcuk University, Konya, Turkey. (E-mail: anart@selcuk.edu.tr)
[**]Department of Computer Engineering, KTO Karatay University, Konya, Turkey
[***]Department of Health, Aksaray, Ministry of Health, Turkey

Abstract: In this study, the effects of blood glucose levels on hemoglobin A1c (HbA1c) were investigated. For this reason, a classification model was developed by carrying out a logistic regression analysis based on machine learning and data mining methods. The purpose of using logistic regression analysis in this study was to establish a method of creating a statistical model that is most suitable and reasonable for determining the relationship between dependent and independent variables. This model shows how effective the factors that cause an increase in the HbA1c level. It can be planned to verify this method on more Electronic Heath Records databases to address the learning method of information in the local health sector with the help of data mining and machine learning methods and different clinical problems for future work.

Keywords: big data; data mining; machine learning; classification; regression analysis; diabetes mellitus.

1. INTRODUCTION

With growing the amount of data, it has become increasingly difficult for people to understand big data (Philip Chen and Zhang, 2014). The concept of machine learning and data mining were developed as solutions to this problem. Nowadays, the vast majority of data mining and machine learning models can also be used in medical research (Kavakiotis et al., 2017; Wu et al., 2018; Cox et al., 2005).

In recent years, the proliferation of diabetes mellitus (World Health Organization 2016: Global Report on Diabetes) has led to the utilization of data mining in various studies on this disease (Li et al., 2018). For instance, data mining has shown that epigenetic (non-genetic) changes in an organism influence the development of type 2 diabetes (Wren et al., 2005).

In this study, a classification model was developed using logistic regression analysis, based on Spark machine learning library (i.e., Spark MLlib) (Meng et al., 2015; Guller, 2015), which was used to create a classification model for the determination of the probability (odds ratio) factors for the predictive analysis of hemoglobin A1c (HbA1c) (Soltani and A. Jafarian, 2016; Nitin, 2010) which is a dependent attribute in the current dataset. Classification, which is the process of dividing input data into categories, is the general task of machine learning. A classification algorithm defines a principle of assigning "labels" to the input data. Logistic regression is one of the algorithms used for classification (Lin et al., 2014). It is widely used for multi-class and binary classification of input data in the Spark Application Programming Interface (API). The logistic regression process creates a logistic function that can be used to predict the probability (odds ratio) of an input vector belonging to a particular group.

Diabetes mellitus is a disease that causes acute metabolic and chronic degenerative complications, with disrupted carbohydrate, protein, and lipid metabolism characterized by hyperglycemia, which leads to an absolute or relative insufficiency of insulin or pancreatic insulin secretion. Currently, diabetes causes increasingly serious health problems throughout the world.

Uncontrolled diabetes leads to hyperglycemia, which causes complications that affect all systems of an organism, primarily the cardiovascular system, eyes, kidneys, and the nervous system, over time (Marks and Raskin, 2000; Grundy et al., 2002). It is known that the management of hyperglycemia significantly impacts patient morbidity and mortality (Umpierrez et al., 2002; Egi et al., 2016; Wang et al., 2016). There are two common tests for the glycemic control of patients with diabetes mellitus: (a) measuring the amount of blood glucose (in mg/dL) and (b) measuring the level of HbA1c (in %). In general, the goal of treatment is to keep the HbA1c concentration less than 7%, but controlling the treatment regimen is recommended if the concentration of HbA1c is greater than 8% (American Diabetes Association: Tests of Glycemia in Diabetes, 2003). The purpose of this study is to develop a new classification model in the current dataset that helps analyze the factors causing increases in the value of HbA1c. It should be noted that for the patients with HbA1c<7% have well-controlled diabetes (*WCD*) and for those with HbA1c≥7% have poorly controlled diabetes (*PCD*) (Baum et al., 2017; Riveline et al., 2012; McCoy et al., 2017).

Several attributes that can influence the HbA1c output values can be considered as statistically independent features, as an

assumption of the logistic regression model. Regression is widely used as a statistical method in data mining (e.g., the Waikato Environment for Knowledge Analysis, WEKA) (Koliopoulos et al., 2015; Wang, 2006) for determination the relationship between two variable clusters in dataset. The logistic regression model estimates the relationship between the continuous independent and dependent variables using a two-dimensional linear surface (e.g., a plane or a hyperplane). The independent values in the model can be continuous and categorical variables, and the binary interactions of the independent variables can be included as general variables.

If the dependent variable is categorical, using the logit regression model would be much more convenient than using the standard methods. The development of the classification model by the logistic regression method is similar for both machine learning and data mining. Defining the relationship between the array of dependent and independent variables constitutes the bases for the creation of the most suitable and reasonable model. The logistic distribution function is used to describe the logistic regression model (Hilbe, 2009) with dependent and independent variables (Y and X, respectively):

$$P_i = E(Y = 1 \mid (X_{i1}, X_{i2}, \ldots, X_{ik})) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_k X_{ik})}} \quad (1)$$

In the logistic distribution function (P_i), the probability of the feature that will make a specific choice value i of the independent variable X_i (the probability that Y will be 1 or 0 for the i^{th} sample); e is the base of the natural logarithm (approximate value e = 2.72 equals), where P_i is a nonlinear function with respect to both values: the independent variables ((X_{i1},X_{i2},…,X_{ik}) are the *first*, *second,* and k^{-th} values of the independent features) and model β-parameters (β_0,β_1,β_2,…,β_k) are the regression coefficients for the corresponding variables in the model). The cumulative logistic probability distribution function shows the "S" curve, where the lower and upper bounds are zero and one, respectively. The model, which expressed as equation (1) can be linearized using the appropriate transformations, even if it is nonlinear. If in the model (1), (β_0+$\beta_1 X_{i1}$+ $\beta_2 X_{i2}$…+$\beta_k X_{ik}$) to interchange by (Z_i), so function (2) is gotten:

$$P_i = \frac{1}{1 + e^{-Z_i}} \quad (2)$$

If P_i is the case's occurrence probability, so (1-P_i) is the case's non-occurrence probability. By dividing the case's occurrence probability to the case's non-occurrence probability, function (3) is received:

$$\frac{P_i}{1 - P_i} = e^{Z_i} \quad (3)$$

If apply natural logarithm to both sides, so function (4) is obtained:

$$\ln\left(\frac{P_i}{1-P_i}\right) = \ln(e^{Z_i}) \Rightarrow L_i = \ln\left(\frac{P_i}{1-P_i}\right) = Z_i \Rightarrow$$
$$\Rightarrow Z_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_k X_{ik} \quad (4)$$

Therefore, the non-linear logistic regression model is linearized based on both its parameters and variables. *(L)* is called the logit model (Gujarati, 2003). In our case, independent variables, X_i=(X_{i1},X_{i2},…,X_{i15})=[$race_i$, $gender_i$, age_i, $max\_glu\_serum_i$, $metformin_i$, $repaglinide_i$, $chlorpropamide_i$, $glimepiride_i$, $glipizide_i$, $glyburide_i$, $pioglitazone_i$, $rosiglitazone_i$, $insulin_i$, $change_i$, $diabedesMed_i$], and Y=[*A1c result*] is dependent variable.

Logistic regression analysis considers the dataset as a homogeneous whole. Therefore, the reliability of the parameters predicted by these methods and generalization at the universe have been discussed in studies (Larose, 2007). The regression analysis (logistic), which examines the relationship between the independent and dependent variables, is a test statistic that can be applied after some assumptions (e.g., linearity, normality, homogeneity and summability) are fulfilled. If the assumptions are not fulfilled, optimization of the dataset may be attempted. It can be done either by the logarithmic transformations of the original values in the dataset or by square root transformation methods (Chen and Kou, 2000; Efe et al., 2000). One of the important concepts in logistic regression is the "odds ratio". The "odds" or "odds ratio" is used to interpret the coefficients. In other words, the "odds ratio" is a measure of the relationship between impact and outcome. Odds ratios in the logistic regression analysis are the most important coefficients that represent the odds that an outcome will be obtained at a particular impact, compared to the odds of the outcome occurring in the absence of such an impact.

The rest of this paper is organized as follows. Section 2 introduces the materials and methods that were used to develop a classification model using logistic regression analysis, and the workflow of data preparation and evaluation is also provided. Section 3 presents the results and evaluations of the experiment. Finally, Section 4 presents the conclusion and elaborates upon future work.

2. MATERIALS AND METHODS

Clinical datasets and databases contain very valuable data, but these data also contain missing, incomplete values and inconsistent, complex records (Cios and Moore, 2002). In big data research, it is important to develop new models that can find and extract new useful knowledge from big data. Data mining and machine learning methods are increasingly used in diabetes research (Marinov et al., 2011; Mani et al., 2012, Huang et al., 2007). In this study, independent features were selected by taking recommendations of specialist doctors and we have been analyzed the factors influencing the level of HbA1c. In other words, the classification model was created by carrying out a logistic regression analysis.

In our study, was used the dataset which extracted from the database of Cerner Health Facts (Cerner Corporation, Kansas City, MO, USA), a national data warehouse that collects comprehensive clinical records from 130 US hospitals. This dataset is available online as Supplementary Material at http://dx.doi.org/10.1155/2014/781670 (Strack et al., 2014) and the UCI Machine Learning Repository (Frank and Asuncion, 2010). The examined dataset consists of 50

attributes and 101,766 samples. This dataset contains missing, unnecessary, and noisy data as would be expected from real-world data. It also includes various features that have many missing values that cannot be directly corrected. These features were found to be useless for the analysis. One of these features is patient weight; 98% of the weight values in this dataset are missing. Only 332 out of 101,766 samples, which are 2% of the total sample size, contained weight values. The HbA1c test result values were 7% or higher for 253 samples and less than 7% for 79 samples.

In the study, since the developed classification model was based on supervised learning, a filtering strategy was applied by taking the specialist doctor's opinion to determine the factors causing the increase of HbA1c in the dataset, and the classification model was developed by performing a logistic regression analysis on a total of 16 features; factors causing the increase of HbA1c have been identified. Thus, a data preparation and evaluation workflow (see Fig. 1) was developed in order to improve the classification model features that can be included in more *PCD* samples related to high *PCD* values, by determining the factors that increase the value of HbA1c in accordance with the purpose of the study.

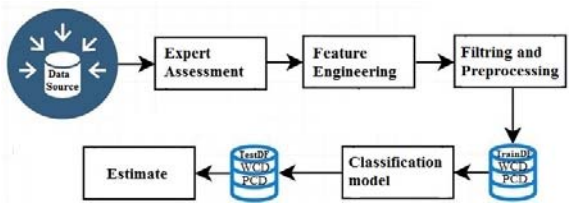

Fig. 1. Workflow of the data preparation and evaluation process.

In the samples examined, patients were categorized by race into "*Caucasian*", "*Asian*", "*African American*" or "*Hispanic*" and patients with missing data in the "*Race*" feature were categorized as "*Other*". In the dataset, under the "*Gender*" attribute, patients with unidentified gender (i.e., "*unknown*") were not included in the classification model because of the lack of HbA1c results only the relationship between HbA1c and "*female*" or "*male*" was analyzed in this study.

The patients were divided into three "*Ages*" groups such that the first group included patients under 30 years old (i.e., between 0 and 30), the second group included patients between 30 and 60 years old, and the last group included patients between 60 and 100 years old. After cleaning the dataset by removing the missing and unidentified values of independent features, HbA1c was examined as a dependent variable. First, the HbA1c result features were classified into two groups based on recommendations of specialist doctors:

(a) patients with HbA1c results and (b) patients without HbA1c results.

Analyzing patients without HbA1c test results would have been meaningless in the current study. Therefore, information about patients without HbA1c test results was not included in the classification model, and only patients with HbA1c test results were analyzed.

At the second stage, patients with HbA1c test results were also classified into two groups; that is, the target dependent variable (i.e., the HbA1c feature) was only considered in two groups:

1. The level of HbA1c was measured, and the patients have test results less than 7% (i.e., they have *WCD*).
2. The HbA1c level was measured, and the patients have test results 7% or higher (i.e., they have *PCD*).

The relationship between the drugs used in the treatment of diabetes, other independent features, and the dependent feature - HbA1c result were analyzed and added to the classification model. Fifteen independent features (including the drugs used in the treatment of diabetes) were determined and analyzed by logistic regression analysis. The probability (odds ratio) of the factors causing the increase of HbA1c is presented in the next section. The attributes studied are summarized in Table (Strack et al., 2014; Frank and Asuncion, 2010).

Table 1. List of attributes and descriptions.

Attribute	Description of values	
race	Caucasian Asian African American Hispanic and Other	
gender	Male Female	
age (year)	Grouped in three parts: [0-30), that is 30.0 [30-60), that is 60.0 [60-100), that is 100.0	
max_glu_serum (mg/dl)	Indicates the range of the result or if the test was not taken: None (If not measured); >200; >300; Norm	
metformin	Steady	(Dosage did not change)
	No	(Drug was not prescribed)
	Up	(Dosage was increased)
	Down	(Dosage was decreased)
repaglinide	Steady	(Dosage did not change)
	No	(Drug was not prescribed)
	Up	(Dosage was increased)
	Down	(Dosage was decreased)
chlorpropamide	Steady	(Dosage did not change)
	No	(Drug was not prescribed)
glimepiride	Steady	(Dosage did not change)
	No	(Drug was not prescribed)
	Up	(Dosage was increased)
	Down	(Dosage was decreased)
glipizide	Steady	(Dosage did not change)
	No	(Drug was not prescribed)
	Up	(Dosage was increased)
	Down	(Dosage was decreased)
glyburide	Steady	(Dosage did not change)
	No	(Drug was not prescribed)
	Up	(Dosage was increased)
	Down	(Dosage was decreased)
pioglitazone	Steady	(Dosage did not change)
	No	(Drug was not prescribed)
	Up	(Dosage was increased)
	Down	(Dosage was decreased)
rosiglitazone	Steady	(Dosage did not change)
	No	(Drug was not prescribed)

	Up	(Dosage was increased)
	Down	(Dosage was decreased)
insulin	Steady	(Dosage did not change)
	No	(Drug was not prescribed)
	Up	(Dosage was increased)
	Down	(Dosage was decreased)
change	Indicates if there was a change in diabetic medications (either dosage or generic name): "Change" and "No change"	
diabetesMed	Indicates if there was any diabetic medication prescribed: "Yes" and "No"	
A1c result	If the result (Prediction): HbA1c value 7% or higher, PCD [0.0] HbA1c value less than 7%, WCD [1.0]	

After preanalysis and preprocessing, the number of attributes and samples was reduced to 16 (see Table 1) and 17,018 respectively. Therefore, each patient feature that could affect the HbA1c test result may be assumed to be statistically independent, as part of the logistic regression classification model. It was aimed to develop a classification model by using logistic regression analysis to determine the factors that cause the increase of HbA1c by controlling the relationship between HbA1c test result, which is accepted as a dependent feature, and the 15 independent features above determined.

3. RESULTS AND DISCUSSION

3.1. Experimental Setup

At the first stage of establishing the classification model of logistic regression analysis, it is necessary to determine what should be done with the 15 independent features of patients with drugs used for the treatment of diabetes mellitus. Once this is completed, the factors causing an increase in HbA1c can be estimated. Therefore, a logistic regression classification model was created in the Scala language (Scala version 2.11.8) using the parallel Apache Spark (spark-2.3.1-bin-hadoop2.7.tgz) machine learning packages (Meng et al., 2016) to estimate the factors influencing the HbA1c result by 15 independent features.

The dataset in ".csv" format was loaded into DataFrame via Scala using the library "apache.spark.sql.DataFrame". Information about the sample numbers (data.count) and attributes (data.columns) and information regarding the data structure have also been checked in the data frame (DataFrame) (Algorithm 1).

Some features must be converted to a more appropriate double type using org.apache.spark.sql.types. DoubleType and org.apache.spark.ml.feature. {StringIndexer} libraries and categorical (nominal, numeric variables) features to predict the target dependent variable HbA1c. In order to do this, the HbA1c result dependent feature was indexed; that is the *PCD* results were converted to [0.0] and the *WCD* results were converted to [1.0] (Algorithm 2).

Machine learning algorithms do not work directly with categorical values. Therefore, the OneHotEncoding (org.apache.spark.ml.feature.OneHotEncoder) library, which represents categorical variables as binary vectors, was necessary to convert categorical values into categorical numbers (Algorithm 3).

```
scala> val data = spark .read .option("header", "true") .csv("HbA1C.csv")
data: org.apache.spark.sql.DataFrame = [race: string, gender: string ... 3 more fields]
scala> data.count
res0: Long = 17018
scala> data.columns
res1: Array[String] = Array(race, gender, age, max_glu_serum, metformin, repaglinide,
chlorpropamide, glimepiride, glipizide, glyburide, pioglitazone, rosiglitazone, insulin,
change, diabetesMed, A1Cresult)
scala> data.printSchema
root
 |-- race: string (nullable = true)
 |-- gender: string (nullable = true)
 |-- age: string (nullable = true)
 |-- max_glu_serum: string (nullable = true)
 |-- metformin: string (nullable = true)
 |-- repaglinide: string (nullable = true)
 |-- chlorpropamide: string (nullable = true)
 |-- glimepiride: string (nullable = true)
 |-- glipizide: string (nullable = true)
 |-- glyburide: string (nullable = true)
 |-- pioglitazone: string (nullable = true)
 |-- rosiglitazone: string (nullable = true)
 |-- insulin: string (nullable = true)
 |-- change: string (nullable = true)
 |-- diabetesMed: string (nullable = true)
 |-- A1Cresult: string (nullable = true)
scala> data.take(5)
res3: Array[org.apache.spark.sql.Row] =
Array([Caucasian,Male,100,None,Steady,No,No,No,No,Steady,No,No,No,Ch,Yes,PCD],
[Caucasian,Female,100,None,No,No,No,No,No,No,No,No,Up,Ch,Yes,PCD],
[Other,Female,60,None,No,No,No,No,No,No,No,No,Up,Ch,Yes,PCD],
[Caucasian,Male,100,None,No,No,No,No,No,No,No,No,Steady,No,Yes,WCD],
[Caucasian,Female,100,None,No,No,No,No,No,No,No,No,No,No,No,WCD])
```

Algorithm 1. DataFrame creation and data structure.

```
scala> val labelIndexer = new StringIndexer() .setInputCol("A1Cresult")
.setOutputCol("A1Cresult_index") .fit(dataDF)
labelIndexer: org.apache.spark.ml.feature.StringIndexerModel = strIdx_7329066a7c3b
scala> labelIndexer .transform(dataDF) .columns
res6: Array[String] = Array(race, gender, max_glu_serum, metformin, repaglinide,
chlorpropamide, glimepiride, glipizide, glyburide, pioglitazone, rosiglitazone, insulin,
change, diabetesMed, A1Cresult, age, A1Cresult_index)
scala> labelIndexer.transform(dataDF).take(4)
res7: Array[org.apache.spark.sql.Row] =
Array([Caucasian,Male,None,Steady,No,No,No,No,Steady,No,No,No,Ch,Yes,PCD,100.0,0.
0], [Caucasian,Female,None,No,No,No,No,No,No,No,No,Up,Ch,Yes,PCD,100.0,0.0],
[Other,Female,None,No,No,No,No,No,No,No,No,Up,Ch,Yes,PCD,60.0,0.0],
[Caucasian,Male,None,No,No,No,No,No,No,No,No,Steady,No,Yes,WCD,100.0,1.0])
```

Algorithm 2. Indexing of dependent features.

```
scala> val raceIndexer = new StringIndexer() .setInputCol("race") .setOutputCol("race" +
"_index") .setHandleInvalid("skip") .fit(dataDF)
raceIndexer: org.apache.spark.ml.feature.StringIndexerModel = strIdx_892fae569753
scala> val raceOneHotEncoder = new OneHotEncoder() .setInputCol("race" + "_index")
.setOutputCol("race" + "_vec")
raceOneHotEncoder: org.apache.spark.ml.feature.OneHotEncoder = oneHot_d057984f09c9
scala> val genderIndexer = new StringIndexer() .setInputCol("gender")
.setOutputCol("gender" + "_index") .setHandleInvalid("skip") .fit(dataDF)
genderIndexer: org.apache.spark.ml.feature.StringIndexerModel = strIdx_929bd6ad77f7
scala> val genderOneHotEncoder = new OneHotEncoder() .setInputCol("gender" +
"_index") .setOutputCol("gender" + "_vec")
genderOneHotEncoder: org.apache.spark.ml.feature.OneHotEncoder =
oneHot_d535e580780b
.
.
.
scala> val test1 = raceIndexer.transform(dataDF)
test1: org.apache.spark.sql.DataFrame = [race: string, gender: string ... 15 more fields]
scala> val test2 = raceOneHotEncoder.transform(test1)
test2: org.apache.spark.sql.DataFrame = [race: string, gender: string ... 16 more fields]
```

Algorithm 3. Example of data conversion.

The index and vector columns were created after the transformation of independent categorical variables, and the org.apache.spark.ml.feature.VectorAssembler library was used to combine all features in a vector (Algorithm 4).

```
scala> val featuresAssembler = new VectorAssembler() .setInputCols(Array("age",
"race_vec", "gender_vec", "max_glu_serum_vec", "metformin_vec", "repaglinide_vec",
"chlorpropamide_vec", "glimepiride_vec", "glipizide_vec", "glyburide_vec",
"pioglitazone_vec", "rosiglitazone_vec", "insulin_vec", "change_vec", "diabetesMed_vec"))
.setOutputCol("features")
featuresAssembler: org.apache.spark.ml.feature.VectorAssembler =
vecAssembler_302ff0e3aa6a
scala> val assembled = featuresAssembler.transform(test28)
assembled: org.apache.spark.sql.DataFrame = [race: string, gender: string ... 43 more fields]
scala> assembled.columns
res19: Array[String] = Array(race, gender, max_glu_serum, metformin, repaglinide,
chlorpropamide, glimepiride, glipizide, glyburide, pioglitazone, rosiglitazone, insulin,
change, diabetesMed, A1Cresult, age, race_index, race_vec, gender_index, gender_vec,
max_glu_serum_index, max_glu_serum_vec, metformin_index, metformin_vec,
repaglinide_index, repaglinide_vec, chlorpropamide_index, chlorpropamide_vec,
glimepiride_index, glimepiride_vec, glipizide_index, glipizide_vec, glyburide_index,
glyburide_vec, pioglitazone_index, pioglitazone_vec, rosiglitazone_index,
rosiglitazone_vec, insulin_index, insulin_vec, change_index, change_vec,
diabetesMed_index, diabetesMed_vec, features)
```

Algorithm 4. Putting all features together in a vector.

At this stage, the last-created features were normalized using the org.apache.spark.ml.feature.MinMaxScaler and org.apache.spark.ml.linalg.Vectors libraries. Each feature was converted to the range [0,1] for resizing, after the summary statistics were calculated (Algorithm 5).

```
scala> val scaler = new MinMaxScaler() .setInputCol("features")
.setOutputCol("scaledFeatures")
scaler: org.apache.spark.ml.feature.MinMaxScaler = minMaxScal_ce98e8a896f7
scala> val scalerModel = scaler.fit(assembled)
scalerModel: org.apache.spark.ml.feature.MinMaxScalerModel =
minMaxScal_ce98e8a896f7
scala> val scaledData = scalerModel.transform(assembled)
scaledData: org.apache.spark.sql.DataFrame = [race: string, gender: string ... 44 more fields]
scala> scaledData.columns
res21: Array[String] = Array(race, gender, max_glu_serum, metformin, repaglinide,
chlorpropamide, glimepiride, glipizide, glyburide, pioglitazone, rosiglitazone, insulin,
change, diabetesMed, A1Cresult, age, race_index, race_vec, gender_index, gender_vec,
max_glu_serum_index, max_glu_serum_vec, metformin_index, metformin_vec,
repaglinide_index, repaglinide_vec, chlorpropamide_index, chlorpropamide_vec,
glimepiride_index, glimepiride_vec, glipizide_index, glipizide_vec, glyburide_index,
glyburide_vec, pioglitazone_index, pioglitazone_vec, rosiglitazone_index,
rosiglitazone_vec, insulin_index, insulin_vec, change_index, change_vec,
diabetesMed_index, diabetesMed_vec, features, scaledFeatures)
```

Algorithm 5. Scaling all features.

After preparing the training sample, (i.e., modifiers for data processing), it is necessary to create an estimation algorithm (odds ratio) which will give the training model on the output, so to define the logistic regression model (Dreiseitl et al., 2002) using the library of org.apache.spark.ml.classification.LogisticRegression, LogisticRegressionModel} and to set the result (outcome) column so that the logistic regression estimator (odds ratio) can be defined. In the study, the column "A1Cresult index" was processed as the "'label" column. The featured column must be scaled, and the maximum number of iterations was set to 20. After setting the "RegParam" parameter to 0.05, how to set the "ElasticNetParam" parameter and the kind of regularization, it was set to 0.0 so that L2 regularization was superior needed were investigated (Li et al., 2016). After making the necessary adjustments using the estimator and org.apache.spark.ml.{Pipeline,PipelineModel} library the pipeline was created as shown in Algorithm 6.

```
scala> val lr = new LogisticRegression() .setLabelCol("A1Cresult_index")
.setFeaturesCol("scaledFeatures") .setMaxIter(20) .setRegParam(0.05)
.setElasticNetParam(0.0)
lr: org.apache.spark.ml.classification.LogisticRegression = logreg_19478955fb0e
scala> val pipeline = new Pipeline().setStages(Array(raceIndexer, raceOneHotEncoder,
genderIndexer, genderOneHotEncoder, max_glu_serumIndexer,
max_glu_serumOneHotEncoder, metforminIndexer, metforminOneHotEncoder,
repaglinideIndexer, repaglinideOneHotEncoder, chlorpropamideIndexer,
chlorpropamideOneHotEncoder, glimepirideIndexer, glimepirideOneHotEncoder,
glipizideIndexer, glipizideOneHotEncoder, glyburideIndexer, glyburideOneHotEncoder,
pioglitazoneIndexer, pioglitazoneOneHotEncoder, rosiglitazoneIndexer,
rosiglitazoneOneHotEncoder, insulinIndexer, insulinOneHotEncoder, changeIndexer,
changeOneHotEncoder, diabetesMedIndexer, diabetesMedOneHotEncoder) ++
Array(featuresAssembler, scaler, labelIndexer, lr))
pipeline: org.apache.spark.ml.Pipeline = pipeline_87f28daa6d1e
```

Algorithm 6. Setting up the logistic regression model and creating the pipeline.

In this study, the pipeline contained a binary vector (i.e., raceIndexer, raceOneHotEncoder, genderIndexer, and genderOneHotEncoder) and categorical variables (featuresAssembler, scaler, labelIndexer, and lr). The dataset needed to be split into "train" and "test" datasets. That is why two data frames were created for the dataset that uploaded to the "DataFrame", with rates of 65% and 35% for the "train" and "test" data, respectively, using the random selection method (Algorithm 7).

```
scala> val Array(trainDF, testDF) = dataDF.randomSplit(Array(0.65, 0.35))
trainDF: org.apache.spark.sql.Dataset[org.apache.spark.sql.Row] = [race: string, gender: string ...
14 more fields]
testDF: org.apache.spark.sql.Dataset[org.apache.spark.sql.Row] = [race: string, gender: string ...
14 more fields]
scala> trainDF.count
res23: Long = 11075
scala> trainDF.take(3)
res24: Array[org.apache.spark.sql.Row] =
Array([AfricanAmerican,Female,>200,No,No,No,No,No,No,No,No,No,No,Yes,WCD,100.0],
[AfricanAmerican,Female,>300,No,No,No,No,No,No,No,No,No,No,No,PCD,30.0],
[AfricanAmerican,Female,>300,No,No,No,No,No,No,No,No,No,No,No,PCD,60.0])
scala> testDF.count
res25: Long = 5943
scala> testDF.take(3)
res26: Array[org.apache.spark.sql.Row] =
Array([AfricanAmerican,Female,>200,No,No,No,No,No,No,No,No,No,No,No,PCD,100.0],
[AfricanAmerican,Female,>200,No,No,No,No,Steady,No,No,No,No,No,Yes,PCD,60.0],
[AfricanAmerican,Female,>200,Steady,No,No,No,No,No,No,No,No,No,Yes,WCD,60.0])
scala> val model = pipeline.fit(trainDF)
model: org.apache.spark.ml.PipelineModel = pipeline_87f28daa6d1e
```

Algorithm 7. Splitting of the dataset for training and testing.

The results obtained using the classification model in the test dataset is shown in Algorithm 8.

```
scala> val testPreds = model.transform(testDF)
testPreds: org.apache.spark.sql.DataFrame = [race: string, gender: string ... 48 more fields]
scala> testPreds.columns
res35: Array[String] = Array(race, gender, max_glu_serum, metformin, repaglinide, chlorpropamide,
glimepiride, glipizide, glyburide, pioglitazone, rosiglitazone, insulin, change, diabetesMed, A1Cresult, age,
race_index, race_vec, gender_index, gender_vec, max_glu_serum_index, max_glu_serum_vec,
metformin_index, metformin_vec, repaglinide_index, repaglinide_vec, chlorpropamide_index,
chlorpropamide_vec, glimepiride_index, glimepiride_vec, glipizide_index, glipizide_vec, glyburide_index,
glyburide_vec, pioglitazone_index, pioglitazone_vec, rosiglitazone_index, rosiglitazone_vec, insulin_index,
insulin_vec, change_index, change_vec, diabetesMed_index, diabetesMed_vec, features, scaledFeatures,
A1Cresult_index, rawPrediction, probability, prediction)
scala> labelIndexer.labels
res37: Array[String] = Array(PCD, WCD)
scala> testPreds.select("A1Cresult", "probability", "prediction").take(4)
res38: Array[org.apache.spark.sql.Row] = Array([WCD,[0.0550784984420189, 0.9449215015579811],1.0],
[PCD,[0.9881064722794497,0.011893527720550 3],0.0],
[PCD,[0.9764829180274204,0.0235170819725796],0.0],
[PCD,[0.9898491325470162,0.0101508674529838],0.0])
```

Algorithm 8. Results of probability and prediction parameters.

Descriptions of the probability and prediction parameters are shown in Fig. 2 and for instance, *PCD* result [0.9898491325470162, 0.0101508674529838], 0.0] is shown in Fig. 2(a), but *WCD* result [0.0550784984420189, 0.9449215015579811], 1.0] in Fig. 2(b). The probability is a vector, where 98.984% indicates the probability index of "0" and 1.0150% indicates the probability index of "1" [see Fig. 2(a)]. Thus, the higher prediction value (98.984%) of the "0" index indicates that the *PCD* value [0.0] will be on the output [see Fig. 2(a)]. In the other example [see Fig. 2(b)], 5.5078% indicates the probability index of "0" and 94,492% indicates the probability index of "1" and the higher prediction value (94,492%) "1" index, means that the *WCD* value [1.0] will be on the output [see Fig. 2(b)].

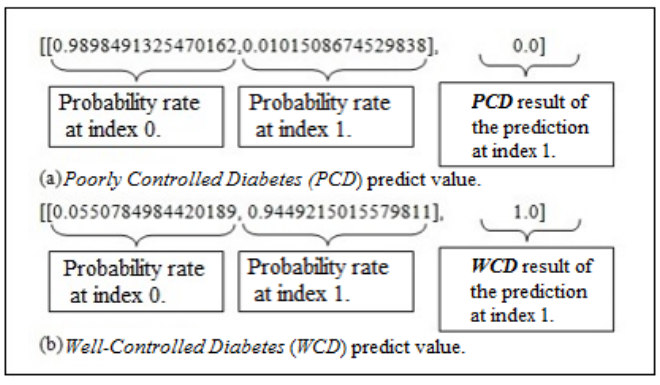

Fig. 2. Description of the prediction parameters of (a) PCD and (b)WCD.

Due to the dependent feature (the result of HbA1c), the label indexes used for the *PCD* and *WCD* parameters were determined as [0.0] and [1.0] values, respectively. This means that the [0.0] value mapped to the *PCD* prediction parameters and the [1.0] value mapped to the *WCD* prediction parameters.

3.2. Performance of the Classification Model

The performance of the classification model was evaluated on the test dataset (Gu and Li, 2013; Lustgarten et al., 2008). Spark ML packages were used for evaluation and the "areaUnderROC" value was obtained using the org.apache.spark.ml.evaluation.{Binary ClassificationEvaluator} library (Algorithm 9).

```
scala> val evaluator = new BinaryClassificationEvaluator().setLabelCol("A1Cresult_index")
.setRawPredictionCol("probability").setMetricName("areaUnderROC")
evaluator: org.apache.spark.ml.evaluation.BinaryClassificationEvaluator = binEval_068fad501a68
scala> evaluator.evaluate(testPreds)
res42: Double = 0.9980203970132945
```

Algorithm 9. Evaluation of the classification model.

The following results were obtained from other metrics: TP = 3598, TN = 2255, FP = 81, FN = 9, precision (positive predictive value) = 0.9779, recall (true positive rate) = 0.9975, $F_{measure}$= 0.9876, accuracy = 0.985, sensitivity = 0.9975 and specificity = 0.965.

Using the classification model of logistic regression in data mining (i.e., in WEKA) (Koliopoulos et al., 2015), the odds ratios of the factors that affect the dependent factors are shown in Table 2. The correlation between several actual and unrealized cases was found upon studying the odds coefficients (Can et al., 2018; Özdil et al., 2010; Girginer et al., 2008). Thus, the results of the regression analysis are obtained by the methods of machine learning and data mining (see Table 2).

Table 2. Analysis results of the factors influencing HbA1c.

Attribute name	Value	Estimate (Odds Ratios)
race	Caucasian	1.0382
	Other	1.3415
	African American	0.8501
	Hispanic	1.1491
	Asian	1.1431
gender	Female	0.8911
age	[60-100)	0.8046
	[30-60)	1.0842
	[0-30)	1.997
max_glu_serum	None	0.4407
	>200	0.8232
	>300	14.5117
	Norm	0.402
metformin	Steady	1.0136
	No	0.9139
	Up	1.9309
	Down	0.9913
repaglinide	No	0.8109
	Steady	1.0694
	Up	3.8397
	Down	0.691
glimepiride	No	0.7999
	Steady	1.1572
	Down	0.9896
	Up	2.5232
glipizide	No	0.8604
	Steady	1.1474
	Up	1.3482
	Down	0.9709
pioglitazone	No	1.0526
	Steady	0.9538
	Down	1.0326
	Up	0.8554
rosiglitazone	No	0.9677
	Steady	1.0152
	Up	2.0176
	Down	0.7056
insulin	No	0.7427
	Up	1.5134
	Steady	0.9428
	Down	1.2255
change	No	0.9316
diabetesMed	No	0.7345

According to the results shown in Table 2, odds ratios of more than 1 affect the dependent feature and increase the probability of getting the *PCD* result of HbA1c. Moreover, coefficients that are less than 1 affect the dependent feature and decrease the superiority of the *PCD* result probability.

First, for the independent values of "*Race*", with the exception "*African American*" value, the odds ratios for values "*Caucasian*", "*Other"*, "*Hispanic*", and "*Asian*") were more than 1. However, the odds ratio for "*African American*" was less than 1. This means that the HbA1c *PCD* result probability in "*Caucasian*", "*Other*", "*Hispanic*", and "*Asian*" patients increased 1.0382, 1.3415, 1.1491, and 1.1431 times, respectively. However, the HbA1c *PCD* result (output) probability in "*African American*" patients decreased 0.8501 times. In other words, it was determined that the HbA1c *PCD* result in the "*Other*" subgroup was rather high here (i.e., among "*Race*" values). It was determined that the HbA1c *PCD* result was low in "*African American*" subgroup of "*Race*". The HbA1c *PCD* result slightly increased in patients belonging to the "*Caucasian*", "*Other*", "*Hispanic*", and "*Asian*" subgroups.

The odds ratio for the "*Female*" independent value was less than 1. That is, the HbA1c *PCD* result probability in the "*Female*" values decreased 0.8911 times. This indicates that the level of HbA1c in women is low.

Among the values of the independent feature "*Age*" " (Wu et al., 2017; Yang et al., 2015), the odds ratios for the subgroups (values) apart from the [60-100] subgroup (value), were more than 1. However, the odds ratio for the [60-100] subgroup of "*Age*" was less than 1. This indicates that the HbA1c *PCD* result probability of patients from the [30-60) and [0-30) subgroups of "*Age*" increased 1.0842 and 1.997 times, respectively. However, the HbA1c *PCD* result (output) probability in patients from the [60-100] group decreased 0.8046 times. In other words, the level of HbA1c in the [0-30) subgroup was rather high among the values of the "Age" independent feature.

Previous studies have shown that age, gender and race have no clinically significant effect on HbA1c. However, there is no such consensus on the influence of age. Some researchers found a ~0.1% increase in results every 10 years after 30, whereas other researchers reported insignificant or no increase.

In this study, among values of "*max_glu_serum*" independent feature, except "*>300 mg/dL*" value, the odds ratios for other subgroups ("*None*", "*>200 mg/dL*", and "*Norm*" values) are less than 1. The odds ratio for the subgroup with "*>300 mg/dL*" value of "*max_glu_serum*" independent feature is more than 1. This indicates that probability of the HbA1c *PCD* result at patients with "*None*", "*>200 mg/dL*" and "*Norm*" values of "*max_glu_serum*" independent feature decreased 0.4407, 0.8232, and 0.402 times, respectively. However, the odds ratio of *PCD* output at patients from "*max_glu_serum*" independent feature with *>300 mg/dL* value increased 14.5117 times. The HbA1c level was rather high when the blood glucose level was over 300 mg/dL.

When the relationship between the drugs used in diabetes mellitus and HbA1c was investigated, it was determined that the odds ratios for "*Steady*" and "*Up*" subgroups (values) of the "*metformin*" independent value were more than 1. This indicates that the HbA1c *PCD* result probabilities, for patients using metformin at maintained and increased dosages, have increased 1.0136 and 1.9309 times, respectively. Thus, patients with elevated blood sugar levels may continue to use metformin and increase the dosage, so that elevated blood sugar levels can increase the level of HbA1c. The odds ratios for the "*Down*" and "*No*" subgroups of the metformin independent feature were less than 1. This indicates that the odds ratios of *PCD,* for patients who use metformin in small dosages and those who do not use metformin at all, have decreased 0.9913 and 0.9139 times, respectively. In other words, patients with normal blood glucose levels may have decreased HbA1c levels in those who have not used metformin or have a reduced dose and therefore have a normal blood glucose level.

Similar results were obtained when oral antidiabetic drugs such as repaglinide, glimepiride, glipizide, and rosiglitazone were used. The odds ratios for "*Steady*" and "*Up*" subgroups were more than 1, and the odds ratios for the "*Down*" and "*No*" subgroups were less than 1, as summarized in Table 2. Most of all, when patients used oral antidiabetic drugs such as repaglinide, glimepiride, and rosiglitazone in high dosages, the HbA1c *PCD* result probability increased 3.8397, 2.5232, and 2.0176 times, respectively. In the case of elevated blood glucose levels, the dosage of antidiabetic medications such as repaglinide, glimepiride, and rosiglitazone also increased; increasing the blood sugar level may increase the HbA1c level.

The odds ratios for the "*Down*" and "*No*" subgroups of "*pioglitazone*" independent value were more than 1. This indicates that the HbA1c *PCD* result probability increased 1.0326 and 1.0526 times for patients who use pioglitazone in small dosages and those who do not use it at all, respectively. The odds ratios for the "*Steady*" and "*Up*" subgroups of the pioglitazone independent feature were less than 1. This indicates that the HbA1c *PCD* result probabilities, for patients using pioglitazone at maintained and increased dosages, were decreased 0.9538 and 0.8554 times, respectively, that is, when the dose of pioglitazone is not changed and increased in individuals, HbA1c level may be decreased.

When the relationship between the "*insulin*" used in diabetes mellitus and HbA1c was investigated (Bergenstal et al., 2012; Baldwin et al., 2005), it was determined that the odds ratios for the "*Up*" and "*Down*" subgroups of the "*insulin*" independent value were more than 1. This indicates that the HbA1c *PCD* result probabilities, for patients using insulin in decreased and increased dosages, have increased 1.2255 and 1.5134 times, respectively. In other words, when the insulin dosage was decreased, the blood glucose level was increased and therefore the HbA1c level was increased. The insulin dosage was increased for patients whose blood glucose level was increased, at which time the HbA1c level could also be increased. The odds ratios for the "*Steady*" and "*No*" subgroups of the "*insulin*" independent value were less than 1. This indicates that the HbA1c *PCD* result probabilities, for patients not using insulin and those using insulin in maintained dosages, decreased 0.9428 and 0.7427 times, respectively. In other words, patients with stable blood glucose levels did not use insulin or maintained their insulin dosage which may lead to the levels of HbA1c being reduced in these patients.

The odds ratios for the "c*hange*" and "*diabetesMed*" independent features were less than 1. This indicates that the HbA1c *PCD* result probabilities decreased 0.9316 and 0.7345 times, for patients maintaining their dosage of diabetic medications and those not using diabetic medications.

From the analysis result, it is obvious that when the blood glucose level was over 300 mg/dL, the odds ratio of *PCD* output increased 14.5117 times. Therefore, the blood glucose level can be considered as a major factor affecting the PCD output. Each 1% change in HbA1c corresponds to ~35mg/dL as the mean plasma glucose value (Rohlfing et al., 2002). Previous studies on diabetes control and complications have shown a very high relationship between the blood glucose profile and multiple average HbA1c values measured in a laboratory environment over a one-year. HbA1c is a reliable indicator of average glycemia over a long period of time (Bailey et al., 2016). All diabetic patients should be regularly tested, initially for glycemic control and later for HbA1c.

4. CONCLUSIONS

In this study, a classification model was created by carrying out a logistic regression analysis. Regression is a statistical method that is used to determine the relationship between the values of independent and dependent features in the final dataset. Independent features had continuous and categorical values in the classification model, and the binary interactions of the independent values were included as the general variable.

The aim of using logistic regression analysis is to develop the most appropriate and reasonable model for describing the relationship between dependent and independent features.

According to the obtained results, a new classification model that analyzes the causes of HbA1c was successfully developed. Moreover, for patients with blood glucose levels above 300 mg/dL, the odds ratio of the *PCD* output increased 14.5 times. Therefore, the blood glucose level can be considered as the factor with the most significant effect. The classification model developed by logistic regression analysis showed a more accurate and effective approach to identifying the *PCD* and *WCD* cases (output) of HbA1c.

In the future, the logistic regression method will be verified using local Electronic Health Records Databases in order to solve different local clinical problems.

ACKNOWLEDGEMENTS

A. Taghiyev and A. A. Altun thank the the Selcuk University Scientific Research Projects Coordination Office (BAP) for its support (Project No: 18201154). N. Allahverdi thanks the KTO Karatay University for its support. S. Caglar thanks the Aksaray Health Department for its support.

CONFLICTS OF INTEREST

The authors declare no conflicts of interest regarding the publication of this paper.

REFERENCES

American Diabetes Association: Tests of Glycemia in Diabetes, (2003). *Diabetes Care,* vol.26, Supp 1, pp.106-108, doi: 10.2337/diacare.26. 2007.S106.

Bailey T.S. et al., (2016). American Association of clinical endocrinologists and American college of endocrinology 2016 outpatient glucose monitoring consensus statement, *Endocrine Practice*, vol.22, no.2, pp.231-261, doi:10.4158/EP151124.CS.

Baldwin D., Villanueva G., McNutt R., and Bhatnagar S., (2005). Eliminating inpatient sliding-scale insulin: a reeducation project with medical house staff, *Diabetes Care*, vol.28, no.5, pp. 1008-1011.

Baum A., Scarpa J., Bruzelius E., et al., (2017). Targeting weight loss interventions to reduce cardiovascular complications of type 2 diabetes: a machine learning-based post-hoc analysis of heterogeneous treatment effects in the Look AHEAD trial, *The Lancet Diabetes & Endocrinology*, vol.5, no.10, pp.808-815, doi: 10.1016/S2213-8587(17)30176-6.

Bergenstal R. M., Fahrbach J. L., Iorga S. R., et al., (2012). Preadmission glycemic control and changes to diabetes mellitus treatment regimen after hospitalization, *Endocrine Practice*, vol.18, no.3, pp.371–375.

Can Ş., Özdil T., and Yılmaz C., (2018). Üniversite öğrencilerinin ders başarısını etkileyen faktörlerin lojistik regresyon analizi ile tahminlenmesi, *Int.Review of Economics and Management*, vol.6, no.1, pp.28-49, doi: 10.18825/iremjournal.349984.

Chen Z., and Kou L., (2000). A Note on the Estimation of the Multinominal Logit Model with Random Effects, *The American Statistician*, vol.55, no.2, pp.89-95, 2000.

Cios K. J., and Moore G. W., (2002). Uniqueness of medical data mining, *Artificial Intelligence in Medicine*, vol.26, no.1-2, pp.1–24.

Cox E., (2005). Chapter 3. Approaches to model building, Fuzzy in *the Morgan Kaufmann Series in Data Management Systems,* Modeling and Genetic Algorithms for Data Mining and Exploration, pp.37-64, doi: 10.1016/B978-012194275-5/50005-0.

Dreiseitl S., Machado L., (2002). Logistic regression and artificial neural network classification models: a methodology review, *Journal of Biomedical Informatics*, 35(5-6) 352-359, doi: 10.1016/S1532-0464(03)00034-0.

Efe E., Bek Y., and Şahin M., (2000). SPSS'te Çözümleri ile İstatistik Yöntemler II, *Kahramanmaraş Sütçü İmam Üniversitesi Rektörlüğü Yayın No: 73*, Ders Kitapları Yayın No:9, K.S.Ü. Basımevi, Kahramanmaraş, p.214.

Egi M., Krinsley J.S., et al., (2016). Pre-morbid glycemic control modifies the interaction between acute hypoglycemia and mortality, *Intensive Care Med,* vol. 42, pp. 562, doi: 10.1007/s00134-016-4216-8.

Frank A., and Asuncion A., (2010). UCI Machine Learning Repository, University of California, *School of Information and Computer Science*, 2010; "Diabetes 130-US hospitals for years 1999-2008 Data Set", [Online]. Available: https://archive.ics.uci.edu/ml/datasets/diabetes+130-us+hospitals+for+years+1999-2008#, [Accessed 2018].

Girginer N., and Cankuş B., (2008). Tramvay Yolcu Memnuniyetinin Lojistik Regresyon Analiziyle Ölçülmesi: Estram Örneği", C. Bayar Üniversitesi, Yönetim ve Ekonomi Dergisi, 15(1), pp.181-193.

Grundy S. M., Howard B., Smith S. Jr., et al., (2002). Prevention Conference VI: Diabetes and Cardiovascular Disease: executive summary: conference proceeding for healthcare professionals from a special writing group of the American Heart Association, *Circulation,* vol.105, no.18, pp.2231-2239.

Gujarati DN., (2003). Basic Econometrics. McGraw-Hill Companies. New York.

Gu L., and Li H., (2013). Memory or Time: Performance Evaluation for Iterative Operation on Hadoop and Spark, *2013 IEEE International Conference on High Performance Computing and Communications & 2013 IEEE International Conference on Embedded and Ubiquitious Computing*, Zhangjiajie, pp.721-727, doi: 10.1109/HPCC.and.EUC.2013.106.

Guller M., (2015). Big Data Analytics with Spark: A Practitioner's Guide to Using Spark for Large Scale Data Analysis, *Apress. https://books.google.de/ books?id= bNP8rQEACAAJ.*

Hilbe Y., (2009). Logistic Regression Models. New York: Chapman and Hall/CRC, doi: 10.1201/9781420075779.

Huang Y., McCullagh P., Black N., and Harper R., (2007). Feature selection and classification model construction on type 2 diabetic patients' data, *Artificial Intelligence in Medicine,* vol.41, no.3, pp.251-262.

Kavakiotis I., Tsave O., Salifoglou A., et al., (2017). Machine Learning and Data Mining Methods in Diabetes Research, *Computational and Structural Biotechnology Journal*, vol.15, pp.104-116, doi: 10.1016/j.csbj.2016.12.005.

Koliopoulos A-K., Yiapanis P., Tekiner F., et al., (2015). A Parallel Distributed Weka Framework for Big Data

Mining Using Spark, *2015 IEEE International Congress on Big Data, New York,* pp.9-16, doi: 10.1109/BigDataCongress.2015.12.

Larose D. T., (2007). Chapter 2 Regression Modelling, Chapter 4, Data Mining Methods and Models, Wiley-IEEE Press, pp. 35-169.

Lin C., Tsai C., Lee C. and Lin C., (2014). Large-scale logistic regression and linear support vector machines using spark," *2014 IEEE International Conference on Big Data (Big Data),* Washington, pp.519-528, doi: 10.1109/BigData.2014.7004269.

Li W., Liu H., Yang P., and Xie W., (2016). Supporting Regularized Logistic Regression Privately and Efficiently, *PloS ONE*, 11(6), doi: 10.1371/journal.pone.0156479.

Li Y., Li H., and Yao H., (2018). Analysis and Study of Diabetes Follow-Up Data Using a Data-Mining-Based Approach in New Urban Area of Urumqi, Xinjiang, China, 2016-2017, *Computational and Mathematical Methods in Medicine*, vol.2018, Article ID 7207151, 8 pages, doi: 10.1155/2018/7207151.

Lustgarten J. L., Gopalakrishnan V., Grover H., and Visweswaran S., (2008). Improving classification performance with discretization on biomedical datasets, *AMIA Annual Symposium proceedings*; pp.445-449.

Mani S., Chen Y., Elasy T., et al., (2012). Type 2 Diabetes Risk Forecasting from EMR Data using Machine Learning, *AMIA Annual Symposium Proceedings*, pp. 606-615.

Marinov M., Mosa A. S., Yoo I., and Boren S. A., (2011). Data-Mining Technologies for Diabetes: A Systematic Review", *Journal of Diabetes Science and Technology*; vol.5, no.6, pp.1549-1556, doi: 10.1177/193229681100500631.

Marks J. B., and Raskin P., (2000). Cardiovascular risk in diabetes: a brief review, *J. Diabetes Complications*, vol.14, no.2, pp.108-115.

McCoy R. G., Ngufor C., et al., (2017). Trajectories of Glycemic Change in a National Cohort of Adults with Previously Controlled Type 2 Diabetes, *Medical Care*, vol.55, no.11, pp.956-964, doi: 10.1097/MLR. 00000000 00000807.

Meng X., Bradley J. K., Yavuz B., et. al., (2015). Mllib: Machine learning in apache spark, *Journal of Machine Learning Research,* vol.17, pp.1-7, https://arxiv.org/abs/1505.06807.

Nitin S., (2010). HbA1c and factors other than diabetes mellitus affecting it, *Singapore medical journal*, vol.51, no.8, pp.616-622.

Özdil T., Urdaletova A., Yılmaz C., (2010). İktisadi ve İdari Bilimler Fakültesi Öğrencilerinin Ders Başarılarını Etkileyen Faktörlerin Lojistik Regresyon Analiziyle Araştırılması, *2. Uluslararası Balkanlarda Sosyal Bilimler Kongresi Bildiriler Kitabı,* Sakarya Üniversitesi, Priştine Üniversitesi, Bozok Üniversitesi, Kosova, vol.1, pp.823-842.

Philip Chen C., Zhang, C., (2014). Data-intensive applications, challenges, techniques and technologies: A survey on Big Data, *Information Sciences*, vol.275, pp.314–347, doi: 10.1016/j.ins.2014.01.015.

Riveline JP., Schaepelynck P., et al., (2012). Assessment of patient-led or physician-driven continuous glucose monitoring in patients with poorly controlled type 1 diabetes using basal-bolus insulin regimens: a 1-year multicenter study, *Diabetes Care*, vol.35, no.5, pp.965-971, doi: 10.2337/dc11-2021.

Rohlfing C. L., Wiedmeyer H. M., et al., (2002). Defining the relationship between plasma glucose and HbA1c: analysis of glucose profiles and HbA1c in Diabetes Control and Complications Trial, *Diabetes Care*, vol.25 no.2, pp.275-278.

Soltani Z., and Jafarian A., (2016). A new artificial neural networks approach for diagnosing diabetes disease type II, *International Journal of Advanced Computer Science and Applications,* vol.7, no.6, pp.89–94, doi: 10.14569/IJACSA.2016.070611.

Strack B., DeShazo J., Gennings C., et al., (2014). Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records, *BioMed Research International*, vol. 2014, Article ID 781670, 11 pages, 2014, doi: 10.1155/2014/781670.

Umpierrez G. E., Isaacs S. D., Bazargan N., et al., (2002). Hyperglycemia: an independent marker of in-hospital mortality in patients with undiagnosed diabetes", *Journal of Clinical Endocrinology and Metabolism*, vol.87, no.3, pp.978–982.

Wang J., (2006). Encyclopedia of Data Warehousing and Mining, *Information Science Reference*, pp. 49-140, 2006, doi: 10.4018/978-1-60566-010-3.

Wang Y. J., Seggelke S., et al., (2016). Impact Of Glucose Management Team On Outcomes Of Hospitalizaron In Patients With Type 2 Diabetes Admitted To The Medical Service. *Endocrine Practice*: December 2016, Vol. 22, No. 12, pp. 1401-1405.

World Health Organization 2016: GLOBAL REPORT ON DIABETES [Online] Available: *http://www.who.int/diabetes/global-report/en/* [Accessed 2018].

Wren J. D., and Garner H. R., (2005). Data mining analysis suggests an epigenetic pathogenesis for Type 2 Diabetes, *Journal of Biomedicine and Biotechnology*, vol.2005, no. 2, pp. 104-112, doi: 10.1155/JBB.2005.104.

Wu H., Yang S., Huang Z., et al., (2018). Type 2 diabetes mellitus prediction model based on data mining, *Informatics in Medicine Unlocked*, vol.10, pp.100-107, doi: 10.1016/j.imu.2017.12. 006.

Wu L. et al., (2017). Effect of age on the diagnostic efficiency of HbA1c for diabetes in a Chinese middle-aged and elderly population: The Shanghai Changfeng Study, *PloS ONE,* vol.12, no.9, doi: 10.1371/journal.pone.0184607.

Yang L., et al., (2015). The effectiveness of age on HbA1c as a criterion for the diagnosis of diabetes in Chinese different age subjects, *Clin Endocrinol*, 82, pp.205-212, doi:10.1111/cen.12494.