Journal of Network and Innovative Computing
ISSN 2160-2174 Volume 6 (2018) pp. 016-024

A Predictive Model for Diabetes Medications Management

Tigabu D. Akal[1] and Niketa Gandhi[2]

[1]School of Information Technology Doctorial Program
Addis Ababa University, Addis Ababa, Ethiopia
tigabu.dagne@aau.edu.et

[2]Machine Intelligence Research Labs (MIR Labs)
Scientific Network for Innovation and Research Excellence
Auburn, Washington 98071, USA
niketa@gmail.com

Abstract: The increasing usage of medical electronics in the healthcare industry leads to ease of accessing medical records. Accessing patients' record electronically will save time for health professionals, patients and other concerned bodies. But the information that can be accessed from the electronic medical records or health management information system could not tell the patterns or future existence of particular diseases. In this study experiments were conducted to develop a predictive model for diabetes management that helps health professionals and other units to determine the patterns of the diabetes diseases. The model can predict either the diabetes medication is needed or not for a particular patient based on different features of the investigation. For development of the model different classification algorithms were enforced. From four different experiments conducted in the study, J-48 decision tree with percentage splitting approach scored the best classification accuracy. Future research can be attempted using different approaches like Artificial Neural network, Fuzzy logic and Neuro-fuzzy algorithms in order to compare the results and to enhance the accuracy of the model.

Keywords: Big Data, Big Data Analytics, Data Mining, Decision Tree, Diabetes, Electronic Medical Records, KDD

1. Introduction

Big Data is one of the leading terms in the 21st century. This is due to digitized information accessibility, the internet of things (IoT) and bulky information availability on internet [1]. Different scholars have defined the term big data: the library review of Emerald defines: "*Big Data is the Information asset characterized by such a High Volume, Velocity and Variety to require specific Technology and Analytical Methods for its transformation into Value*"[2]. One of the hottest sub topics of big data is data mining (DM) which is focusing on generating both prevailing and interesting patterns from datasets. DM is the process of generating previously unseen patterns and approaches in the large database or data warehouse in order to build predictive models [2].

Diabetes is a disease that happens when the insulin production in the body is insufficient or the body is incapable to use the produced insulin in a good method, as a result, this leads to high blood glucose [23]. As discussed by Iyer et al. [23] the body cells break down the food into glucose and this glucose requirement to be transported to altogether the cells of the body.

The data mining revolution in the healthcare industry has been driven by the availability of data as collected and generated in electronic medical records (EMRs), health management information systems (HMISs), hospital information systems (HISs) and wearable sensors [3][31-33]. The development of HISs were addressed many significant points in the healthcare industry. Some of these were: changing the data handling approach from manual file handling to computerized, decentralized file handling system by creating both regional and global HISs, creating all-inclusive data handling approach by considering all stakeholders, focusing both on technical and non-technical features of HISs and the introduction of sensors based technologies for health controlling and monitoring [4]. In the meantime researchers, medical professionals, policy makers, non-governmental organizations, international donors and other stakeholders have a lot of expectations of information systems and technologies applications in the health care industries to exploit the maximum potential of HISs using some advanced approaches[5][6].

Even if HMISs are considered as one of the important development line in the case of shifting from paper-based to computer-based storage processing, it is difficult to introduce new treatment of diseases that address population health status. In most healthcare systems patients' data are available and recorded in the form of electronic health records (EHRs) and manual handling ways-using papers and pen. These records are in different format either structured data like patient

demographics, laboratory data and vital signs or unstructured data like physician notes and imaging studies [3]. Therefore capturing, storing, managing and analyzing patients' data using the traditional database and manual file handling approaches will not be a long term solutions.

2. Related Works

Different capability areas should be in place to use the challenges and strategic implications of big data revolution as one of the best opportunity for the creation of knowledge based society. Some of the big data analytics capabilities in the cases of healthcare industry are: "*analytical capability for patterns of care, unstructured data analytical capability, decision support capability, predictive capability, and traceability*" [7]. The research of big data as a source of innovation in the healthcare industry has been attracting many scholars in the area. A global positioning system (GPS) based mobile application development for diabetic patients that helps to track patients' health status by comparing previous history of patients record using big data approach [8]; a predictive algorithm for categorizing patients medical appointments, problems, follow-up and other communications [9][34-36]. As mentioned by the National Health Service (*NHS*) which is the public health services of England, Scotland and Wales with big data healthcare industries are entering to a new era and data-driven innovation is currently being used to improve population health [10]. "*Diabetes is often called a modern-society disease because widespread lack of regular exercise and rising obesity rates are some of the main contributing factors for it*" [12]. Researches has been attempted in the application of diabetes management using data mining approaches. For example, machine learning and DM approaches in the DM research are key ways to utilize huge volumes of existing diabetes related for extracting new knowledge [11] [12] [13] have been attempted. Most of the researches attempted so far are using a supervised approach that is the prediction is as per the exiting or known classes. This research paper applied a semi-supervised model which is considering both known classes and unknown classes of existing datasets. The research of data mining as source of innovation for the healthcare industry should be the interest of academics that enable to contribute to the field of big data. Academics have been moving to address the industry gaps in the application of big data in healthcare. There are many advantages mentioned in the application of DM in regard to healthcare industries. Some of these are:

- It is highly applicable for health organizations which are providing insurance for the people [21]
- It enables for patients in order to get appropriate and timely treatment. In the meantime it helps for healthcare professionals in order to provide timely and accurate decisions regarding patients' health status [22]
- It is very important for public agencies in order to interfere and provide accurate health statistics information for the population [21] [22]

Patil et al., [22] proposed involuntary detection of diabetic symptoms in retinal images by using a neural network method. The network is trained using algorithms for evaluating the optimal global threshold which can minimize pixel classification errors. The system developed by Patil et al., [22] are validated by mechanism of enough indexing approach and providing percentage measures in the detection of eye suspect regions based on neuro-fuzzy subsystem.

There are many data mining tools used by different researchers in order to extract and convert datasets for analysis purpose [21] [22]. As discussed by Oswal et al., and Patil et al., [21] [22] different analysis and conversions tools generate different results on the same datasets. The most commonly used software tool addressed by scholars in the field is the Classification and Regression Tree (CART). CART recursively partitions the input variable space to maximize purity in the terminal tree nodes [23].

As addressed by Velu et al., [24] three techniques have been employed for data analysis, these are: EM algorithm, H-means+ clustering and Genetic Algorithm (GA), for the classification of the diabetic patients. A study has been conducted by Sankaranarayanan et al., [25] intended to determine the hidden knowledge from a specific dataset to progress the quality of health care for diabetic patients. Fuzzy Ant Colony Optimization (ACO) was used on the Pima Indian Diabetes dataset to find set of rules for the diabetes diagnosis [26].

As addressed by Iyer et al., [23] the automatic diagnosis of diabetes is significant real-world medical problem. Discovery of diabetes in its early stages is the key for management. As discussed by Iyer et al., [23] decision trees and Naïve Bayes used to model actual diagnosis of diabetes for local and methodical behavior, along with presenting related work in the field. Investigational results demonstration the helpfulness of the proposed model.

2.1. Methods

The knowledge discovery in database (KDD) process is one of the known model applicable in the data mining researches. It is the model that is also selected for this study. The KDD approaches working on the principle of converting low level data into high level knowledge. The goal KDD and DM is to find interesting patterns or models that exist in databases but are hidden among the volumes of data [15]. The KDD process as described by Fayyad et al., [15] consists of five major phases. Data were collected then using appropriate algorithms then mined patterns were modeled. Figure 1 shows the KDD process model that was used in this study.

As shown in figure 1 the DM process consists of five steps [15]:

- **Data selection** – having two subcomponents: (a) developing an understanding of the application domain and (b) creating a target dataset from the universe of available data;

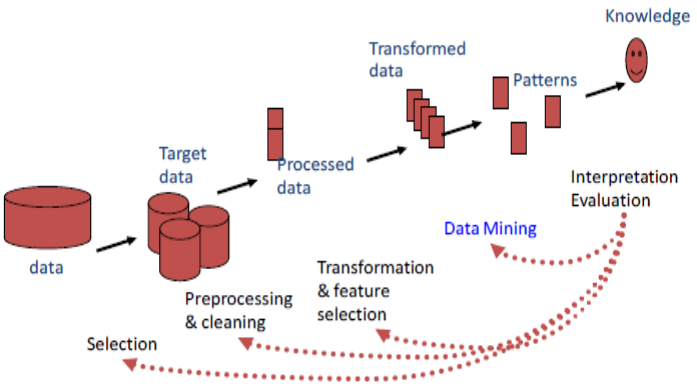

Figure 1: An overview of steps that compose the KDD process [15]

- **Preprocessing** – including data cleaning (such as dealing with missing data or errors) and deciding on methods for modeling information, accounting for noise, or dealing with change over time;
- **Transformation**–using methods such as dimensionality reduction to reduce data complexity by reducing the effective number of variables under consideration;
- **Data mining** – having three subcomponents: (a) choosing the data mining task (e.g., classification, clustering, summarization), (b) choosing the algorithms to be used in searching for patterns, (c) and the actual search for patterns (applying the algorithms);
- **Interpretation/evaluation**– having two subcomponents: (a) interpretation of mined patterns (potentially leading to a repeat of earlier steps), and (b) consolidating discovered knowledge, which can include summarization and reporting as well as incorporating the knowledge in a performance system.

2.1. Initial Data Selection

Dataset accessed from UCI machine learning repository which were recorded 130- US hospitals for years 1999-2008 [14]. In this step all 144,407 records were considered for further processing. The datasets includes around 50 features representing patient and hospital outcomes

2.2. Data Processing

Under the data processing basic operations like removing noise or outliers, deciding strategies for handling missing data fields, removing missing data fields and other operations has been applied [16]. As part of the data pre-processing task the following are applied:

- For some attribute or features, the values of the records were not given. These types of features were automatically discarded.
- Records which have missing values were removed.

2.3. Data Transformation

In this step based on the goal of the study dimensionality reduction or transformation methods to reduce the effective number of features or variables were applied. As part of the data transformation task the following were applied:

- For some attribute or features, the values of the features were the same for all records. These types of features have no impact on the final output of the model due to their similarity. Therefore, these kinds of features were automatically discarded.
- Some features are patient identification and encounter identification number. These were removed.
- After applying pre-processing and data transformation, 27,041 datasets were selected for this study: 66% of these datasets used to build the model and the remaining used to test the model.

2.4. Choosing Data mining tasks

In this step the DM methods used for the study were decided. DM methods have been successfully applied for solving classification problems in many applications [17]. In DM, algorithms (learners) try to automatically filter the knowledge from example data (datasets).This knowledge can be used to make predictions about original data in the future and to provide insight into the nature of the target concept(s). According to Pradeep [17] the example data typically consists of a number of input patterns or examples to be learned. DM systems typically attempt to discover regularities and relationships between features and classes in learning or training phase. For analyzing the data and classification of diabetes medication from a healthcare environment, the three machine learning algorithms [18] the J48 decision tree classifier, Naïve Bayes Classifier and simple k-means clustering were used in this study.

Classification and regression are two data analyzing family of methods which determine important data classes or may construct models which can predict future data trends. The classification model predicts the categorical values; the regression is used in the prediction of values showing continuity. For instance while the classification model is constructed to categorize whether the bank loan applications are safe or risky, the regression model may be constructed to predict the spending of clients buying computer products whose income and occupation are given [27][28] .

2.4.1. Decision Tree

Decision tree is a predictive modeling technique most often used for classification in DM. The Classification algorithm is inductively learned to construct a model from the pre-classified dataset. Each data item is defined by values of the attributes. Classification may be viewed as mapping from a set of attributes to a particular class. The Decision tree classifies the given data item using the values of its attributes. The decision tree is initially constructed from a set of pre-classified data. The main approach is to select the attributes, which best divides the data items into their classes [18]. In this study the J48 decision tree algorithms is used. It is an implementation of the C4.5 decision tree learner. This implementation produces decision tree models. It recursively splits a dataset according to tests on attribute values in order to separate the possible predictions. A decision-tree model is built

by analyzing the training data and the model is used to classify the trained data.

A decision tree is a classifier expressed as a recursive partition of the instance space [30]. Classification trees are frequently used in applied fields such as finance, marketing, engineering and medicine. The classification tree is useful as an exploratory technique. A decision tree may incorporate nominal or numeric or even both of attributes types. As discussed by Chaudhuri [30] the decision tree consists of nodes that form a rooted tree, meaning it is a directed tree with a node called "root" that has no incoming edges. All other nodes have exactly one incoming edge. Test nodes are those which have outgoing edges and the remaining nodes are the leaf nodes which are also referred to as the decision nodes. For each new sample (i.e., feature vector x), the 14 classification algorithm will search for the region along a path of nodes of the tree to which the feature vector x will be assigned. Each of the internal nodes splits the instance space into two or more sub-divisions. The split is based on a certain discrete function used as input. In the simplest and most frequent case, each test considers a single attribute, such that the instance space is partitioned according to the attribute's value. There are two possible types of divisions or partitions: Nominal partitions: a nominal attribute may lead to a split with as many branches as values there are for the attribute. The node of the J48 decision trees evaluates the existence and the significance of every individual feature. Considering a set A of case objects, J48 initially grows a tree and uses divide-and-conquer algorithm as follows: (i) if all the cases in A belong to the same class or if the set is a small one, the tree is leaf labeled with the most frequent occurring class in A. (ii) or, a test is selected based on a single attribute with two or more outcomes.

2.4.2: Naive Bayes

A popular supervised learning technique is the Bayesian statistical methods which allow taking into account prior knowledge when analyzing data [29]. Its popularity can be attributed to several reasons. It is fairly easy to understand and design the model; it does not employ complicated iterative parameter estimation schemes. Its simplicity makes it easier to extend it to large scale datasets. Another reason is that it is also easy to interpret the results. The end users do not require prior expert knowledge in the field which is how it derived the name Na¨ıve Bayes classifier. In the Bayesian approach, the objective is to find the most probable set of class labels given the data (feature) vector and a priori or prior probabilities for each class. It essentially reduces an n-dimensional multivariate problem to n-dimensional univariate estimation. The Naïve Bayes classifier which is based on probabilistic model for assigning the most likely class to given instance. Probabilistic model (approach) in classification field allows (model or looks for) the estimation of conditional probability of classes given instance, p(C/A1…, AN) where C∈ {C1…CM} the classes and Ai, i=1...N, a set of features describing dataset examples [19].

Given a valued example, the most appropriate class to be assigned to is the class with the upper a posterior probability,

Argmaxc p(C=c/A1=a1…, AN=aN)……. (1) Bayesian approach splits a posterior distribution into a priori distribution and likelihood,

Argmaxc p(C=c/A1=a1…, AN=aN) = Argmaxc α p (A1=a1… AN=aN/C=c) p(C=c)……. (2)

Where α is normalization factor to ensure that sums of conditional probabilities over class labels are equal to 1. The distribution of features given class label is more complex to estimate. Its estimation is exponential in attribute number and requires a complete training dataset with sufficient examples for each class. Such problem can be avoided, assumed the independence of features of given class, and likelihood estimation uses the following formula.

P (A1=a1… AN=aN /C=c) =Πi p (Ai=ai /C=c) … (3)

Depending on the precise nature of the probability model, Naive Bayes classifiers can be trained very efficiently in a supervised learning mode for this study.

2.5. Interpretation/evaluation

It is the final step in the selected KDD process model for this study. It includes two basic subcomponents: (a) interpretation of mined patterns (potentially leading to a repeat of earlier steps), and (b) consolidating discovered knowledge, which can included incorporating the knowledge in a performance system.

2.6. Architecture of the study

Supervised diabetes medication recommendation approaches use only labeled data for training. To label the data however are often difficult, expensive, or time consuming as they require the efforts of experienced domain experts. Semi-supervised learning addresses this problem by using large amount of unlabeled data, together with the labeled data, to build better classifiers. Semi-supervised learning requires less human effort. The architecture used for this study showed in figure 2. This architecture proposed by Pachghare et al., [20] for the semi-supervised approach for diabetes medication system. As showed in figure 2, labeled data used for training the system as supervised approach. After training, the system test using unlabeled data. The tested data will add to the training data so as to implement semi- supervised approach.

3. Experimentation

This section describes experimental study of the algorithms and procedures, which are described in the previous sections. In this study both labeled and unlabeled records are used. The dataset is in a spreadsheet (Excel) format save a lot of time in preprocessing. Rows are records of connection; columns are attributes of records (race, gender, age, admission_type_id, num_lab_procedures, metformin, etc…) and each cell should include one value only. Data mining software tool used was WEKA (version 3.9).

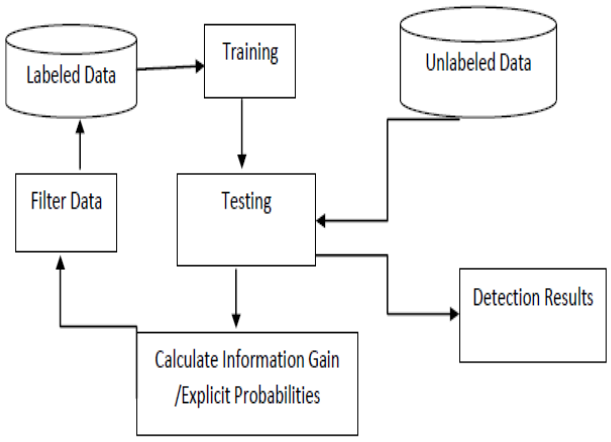

Figure 2: Architecture proposed for Semi-supervised Diabetes management [20]

3.1. Experimental Design

All experiments are performed in a computer with the configurations Intel(R) Core(TM) 7 CPU 2.16GHz, 16 GB RAM, and the operating system platform is Microsoft Windows 10. Weka and Tanagra are collections of machine learning algorithms for data mining tasks that contain facilities for data preprocessing, classification, regression, clustering, association rules, and visualization. Microsoft Excel is used to filter records. The following are the steps used in this study experimentation:

- In the beginning, in order to build the experiment, the researcher selected the data mining software and records used in the experiment.
- The selected records are changed from text format into Microsoft excel format. The Microsoft excel portion of the records contain both labeled and unlabeled records.
- To come up with cleaned datasets preprocessing tasks are undertaken for underling missing values, outliers and other issues.
- Train the classifier using WEKA data mining software. The records are included both labeled and unlabeled that is a semi-supervised modeling approach is followed.

At this step open the ARRF file using WEKA tool. Before running classification techniques different filtering techniques are done depending on the objective of the study. For feature selection there is a filtering technique which is called Attribute Selection under supervised approach or under select attribute menu. Evaluate the trained classifier using all attributes or some selected attributes by excluding unimportant features to achieve feature ranking and selection of relevant features. After selecting either all features or some selected features develop the classifier model for different predictive modeling techniques. Here, there were a number of experiments done by changing different test options and classifier techniques. The performance comparison between different experimentation was evaluated and discussed.

- From previous step, those training models which scores better classification accuracy has selected for this study.
- Lastly, the selected model for this study is tested by previously unseen records which were unlabeled that enable to determine the performance of the selected model.

3.2. Semi-Supervised Modeling

Supervised diabetes management modeling approaches use only labeled data for training. To label the data however are often difficult, expensive, or time consuming as they require the efforts of experienced human attention. Meanwhile unlabeled data may be relatively easy to collect, but there has been few ways to use them. In this study semi-supervised learning addressed this problem by using large amount of unlabeled data, together with the labeled data, to build better classifiers. Semi-supervised learning (SSL) addressed in this study by allowing the model to integrate part or all of the available unlabeled data in its supervised learning [20].The goal is to maximize the learning performance of the model through such newly-labeled examples while minimizing the work required of human effort. In semi-supervised learning the training data (observations, measurements, etc.) are accompanied by both labels and unlabeled records. Label records indicating the class of the observations. Those unlabeled records show that the class is empty. New data is classified based on the training set. Classification is one of the categories under Semi-supervised learning.

As discussed in this study under section 2.3, initially 27,041 datasets records which included only labeled records. Using the labeled records model was trained and tested.

3.2.1. Training the Classifier

As described earlier, the J48 algorithm is used for building the decision tree model. The training of the decision tree classification models of the experimentation is done by employing the 10-fold cross validation and the percentage split classification models. J48 algorithm contains some parameters that can be changed to further improve classification accuracy. Initially the classification model is built with the default parameter values of the J48 algorithm. Table 1 summarizes the default parameters with their values for the J48 decision tree algorithm.

Table 1: Some of the J48 algorithm parameters and their default values

Parameter	Description	Default Value
Confidenec Factor	The confidence factor used for pruning (smaller values incur more pruning)	0.25
minNumObj	The minimum number of instances per leaf	2
Unpruned	Weather pruning is performed	False
Subtreeraising	Weather sub tree information is hidden or expanded	True

As described before, the J48 algorithm is used for building the decision tree model. The training of the decision tree classification models of the experimentation is done by employing the 10-fold cross validation and the percentage split classification modes.

3.2.1.1. Experimentation I:

The first experimentation is performed with the default parameters. The default 10-fold cross validation test option is employed for training the classification model. Using these default parameters the classification model is developed with a J48 decision tree having 20 numbers of leaves and 27 tree size. The decision tree used 15 variables for generating the tree. Some of these are: Race, Gender, Age, admission_type_id, num_lab_procedures, metformin, change and diabetesMed. The decision tree has shown that the change varaible(Indicates if there was a change in diabetic medications (either dosage or genericname) that is its Values: "change" and "no change" is the most determining one. Table 2 depicted the resulting confusion matrix of this model:

Table 2: Semi-supervised classification accuracy using J48 algorithm parameters with 10- fold cross validation

Total number of instances (Training sets)	Correctly classified Instances	Incorrectly Classified Instances
27,041	26,589 (98.3%)	452 (1.7%)

As shown in the resulting confusion matrix, the J48 learning algorithm scored an accuracy of 98.3%. This result shows that out of the total training datasets 27,041 (98.3%) records are correctly classified, while 452 (1.7%) of the records are incorrectly classified. Resulting of confusion is presented in table 3 below:

Table 3: Detailed accuracy by classes

TP Rate	FP Rate	Precision	Recall	F-Measure	Class
1	0.02	0.94	1	0.97	No
0.98	0	1	0.98	0.99	Yes

3.2.1.2. Experimentation II:

The second experiment is performed, by changing the default testing option (the 10-fold cross validation). In this learning scheme a percentage split is used to partition the dataset into training and testing data. The purpose of using this parameter was to assess the performance of the learning scheme by increasing the proportion of testing dataset if it could achieve better classification accuracy than the first experimentation. Even if different experiments were tested using percentage splitting mechanisms the one with 75 % training the model and 25 % testing has scored better classification accuracy. The result of this learning scheme is summarized and presented in Table 4.

Table 4: classification accuracy using J48 algorithm parameters with percentage-split set to 75%

Total number of instances (testing sets)	Correctly classified Instances	Incorrectly Classified Instances
6,760	6,652 (98.4%)	108 (1.6%)

The size of the tree and the number of leaves produced from this training were 27 and 20 respectively. In this experiment out of the 27,041 total records 20,389 (75%) of the records are used for training purpose while 6,670 (25%) of the records are used for testing purpose. As we can see from the confusion matrix of the model developed with this proportion, out of the 6,670 testing records 98.4% of them are correctly classified and 108(1.4%) records are incorrectly classified. Resulting confusion matrix is presented in table 5. The above experiments showed that when the training data decreases the performance of the algorithm for predicting the newly coming instances also increase. Though this experiment is conducted by varying the value of the training and the testing datasets, the accuracy of the algorithm for predicting new instances in their respective class could be improve. This shows that the previous experiment conducted with default 10-fold cross validation is worse than the percentage splitting mechanism.

Table 4: Detailed accuracy by classes

TP Rate	FP Rate	Precision	Recall	F-Measure	Class
1	0.02	0.94	1	0.97	No
0.98	0	1	0.98	0.99	Yes

Therefore, from the above two experiments using decision tree, the second experiment using percentage-split has been chosen due to its better overall classification accuracy

3.2.1.3. Experimentation III & IV:

In this experiment the researchers attempted using the Naïve bayes classification approach in order to compare the previous classification accuracy of the model which was found using the decision tree- J48 method. Using both the cross-validation (10 folds) and the percentage splitting method (75% of the total dataset for training), the classification accuracy of the model is scored is 98% which is equal.

4. Discussion and Evaluation of the Discovered Knowledge

In this section the selected model from those 4 experiments conducted in this study were evaluated. From all experiments in this study, the J48 with percentage split (75% for training the model and 25% for testing it), achieved best classification accuracy which is 98.4%.

Resulting Decision tree

Test mode: split 75.0% train, remainder test
=== Classifier model (full training set) ===
J48 pruned tree

change = No

```
|  insulin = No
|  |  glyburide = No
|  |  |  glipizide = No
|  |  |  |  metformin = No
|  |  |  |  |  rosiglitazone = No
|  |  |  |  |  |  glimepiride = No: No (7738.0/452.0)
|  |  |  |  |  |  glimepiride = Steady: Yes (264.0)
|  |  |  |  |  |  glimepiride = Down: No (0.0)
|  |  |  |  |  |  glimepiride = Up: No (0.0)
|  |  |  |  |  rosiglitazone = Steady: Yes (311.0)
|  |  |  |  |  rosiglitazone = Up: No (0.0)
|  |  |  |  |  rosiglitazone = Down: No (0.0)
|  |  |  |  metformin = Steady: Yes (1109.0)
|  |  |  |  metformin = Up: No (0.0)
|  |  |  |  metformin = Down: No (0.0)
|  |  |  glipizide = Steady: Yes (1185.0)
|  |  |  glipizide = Up: No (0.0)
|  |  |  glipizide = Down: No (0.0)
|  |  glyburide = Steady: Yes (1392.0)
|  |  glyburide = Up: No (0.0)
|  |  glyburide = Down: No (0.0)
|  insulin = Up: Yes (0.0)
|  insulin = Steady: Yes (4811.0)
|  insulin = Down: Yes (0.0)
change = Ch: Yes (10231.0)
```

Some of the rules generated from the selected model are the following:

Rule 1: If change= No and insulin= No and glyburide=No and glipizide=No and metformin=No and rosiglitazone=No and glimepiride=No then No (the diabetes medication No)

Rule 2: If change= No and insulin= Up then Yes (the diabetes medication Yes)

Rule 3: If change= No and insulin= No and glyburide=No and glipizide=No and metformin=No and rosiglitazone=No and glimepiride=Steady then Yes (the diabetes medication Yes)

Rule 4: If change= No and insulin= No and glyburide=No and glipizide=No and metformin=No and rosiglitazone=No and glimepiride=Down then No (the diabetes medication No)

Rule 5: If change= No and insulin= No and glyburide=No and glipizide=No and metformin=No and rosiglitazone=No and glimepiride=Up then No (the diabetes medication No)

Rule 6: If change= No and insulin= No and glyburide=No and glipizide=No and metformin=No and rosiglitazone=Steady then Yes (the diabetes medication Yes)

Rule 7: If change= No and insulin= No and glyburide=No and glipizide=No and metformin=No and rosiglitazone=Steady then Yes (the diabetes medication Yes)

Rule 8: If change= No and insulin= No and glyburide=No and glipizide=No and metformin=No and rosiglitazone=Up then No (the diabetes medication No)

Rule 9: If change= No and insulin= No and glyburide=No and glipizide=No and metformin=No and rosiglitazone=Down then No (the diabetes medication No)

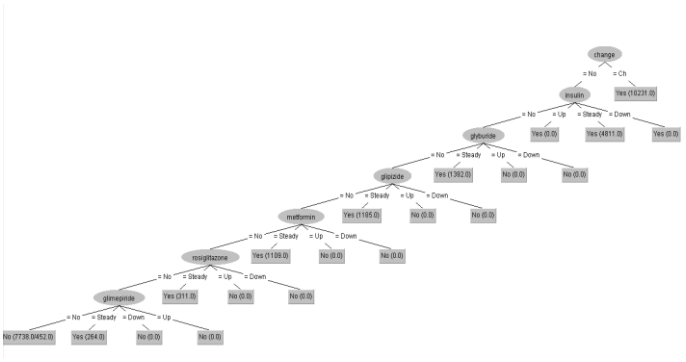

Figure 3: Resulting Decision tree from percentage splitting experiment

5. Conclusion

Data mining is one of the hottest research filed to address healthcare related researches. It helps to generate patterns (both interesting and prevailing rules) that contribute for health professionals, patients and other units in order to generate new patterns for medication issue. Both J48 decision tree algorithm and the Naïve Bayes simple algorithm have been tested as a classification approach for building a predictive model to diabetes patients' medication system. By changing the training test options and the default parameter values of these algorithms, different models have been created. The model created using percentage splitting using the J48 decision tree algorithm with the default parameter values showed the best prediction accuracy. In summary, the results from this study can contribute towards in improving the diabetes patients' for easy decision making for patients to order the medication. This study was carried out using classification algorithms such as J48 decision tree and Naïve Bayes algorithms. So further investigation needs to be done using other classification algorithms such as Neural Networks, Fuzzy logic, Support Vector Machine and Neuro-fuzzy logic plus using association rule discovery for optimize the predication accuracy.

6. References

[1] Paulo B. Goes "Big Data and IS Research", MIS Quarterly, Vol. 38 No. 3, 2014, pp. iii-viii.

[2] Andrea De Mauro, Marco Greco, Michele Grimaldi "A formal definition of Big Data based on its essential features", Library Review, Vol. 65 Issue: 3, 2016,122-135.

[3] Christopher Austin and Fred Kusumoto, "The application of Big Data in Medicine: current implications and future directions", Journal of Interventional Cardiac Electrophysiology, Springer, Vol. 47, Issue 1, 2016, pp 51–59.

[4] Reinhold Haux " Health information systems- past, present, future", International Journal of Medical Informatics, Elsevier, Vol. 75, 2006, pp.268—281.

[5] Chaudhry B, Wang J, Wu S, Maglione M, Mojica W, Roth E, Morton SC and Shekelle PG, " Systematic review: impact of health information technology on quality, efficiency, and costs of medical care", Ann. Intern. Med, Vol.144, 2006, pp. 742–752.

[6] Currie and Finnegan, "The policy-practice nexus of electronic health records adoption in the UK NHS: an institutional analysis", J. Enterp. Inform. Manage, Vol. 24, 2013, pp. 146–170.

[7] Yichuan Wang , LeeAnn Kung , Terry Anthony Byrd, " Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations", Technological Forecasting & Social Change, Elsevier, Vol.126 , 2016, pp. 3–13.

[8] Peter Groves, Basel Kayyali, David Knott and Steve Van Kuiken, "The Big data revolution in Healthcare", Center for US health System Reform Business Technology Office, Mckinsey Global Institute Analysis, 2013.

[9] Robert M. Cronin, Daniel Fabbria, Joshua C. Denny, S. Trent Rosenbloom, and Gretchen Purcell Jackson, "A comparison of rule-based and machine learning approaches for classifying patient portal messages", International Journal of Medical Informatics, Elsevier, Vol. 105 , 2017, pp. 110–120.

[10] Kate Laycock "Big Data in government: challenges and opportunities", 2017.

[11] Ioannis Kavakiotis, Olga Tsave, Athanasios Salifoglou, Nicos Maglaveras, Ioannis Vlahavas and Ioanna Chouvarda, "Machine Learning and Data Mining Methods in Diabetes Research", Computational and Structural Biotechnology Journal, 2017.

[12] Miroslav Marinov, Abu Saleh, Illhoi Yoo, and Suzanne Austin, "Data-Mining Technologies for Diabetes: A Systematic Review", Journal of Diabetes Science and Technology Volume 5, Issue 6, 2011.

[13] Shivakumar and Alby, "A Survey on Data Mining Technologies for Prediction and Diagnosis of Diabetes", International Conference on Intelligent Computing Applications, 2014.

[14] Beata Strack, Jonathan DeShazo, Chris Gennings, Juan Olmo, Sebastian Ventura, Krzysztof Cios, and John Clore , "Impact of HbA1 c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records," BioMed Research International, 2014.

[15] Fayyad U., Piatetsky G., and Smyth P.," The KDD process for Extracting Useful Knowledge from Volumes of Data", Communications of the ACM, Vol. 39, 1996, PP. 27-34.

[16] Meera G., Gandhi and Srivatsa S, "Adaptive Machine Learning Algorithm (AMLA) Using J48 Classifier", Advances in Computational Sciences and Technology, Vol. 3, 2010, PP. 291–304.

[17] Pradeep S. "Comparing the Effectiveness of Machine Learning Algorithms for Defect Prediction", International Journal of Information Technology and Knowledge Management, Vol.2, No.2, 2005, pp.481-483.

[18] Eibe F. and Witten I., "Data Mining–Practical Machine Learning Tools and Techniques," 2nd Edition, Elsevier, 2005.

[19] Shekhar R., Vir V. and Kiran S. "K-Means+ID3: A Novel Method for Supervised Anomaly Detection by Cascading K-Means Clustering and ID3 Decision Tree Learning Methods," IEEE Transactions on Knowledge and Data Engineering, Vol.19, No.3, 2007, PP. 345-354.

[20] Pachghare V., Vaibhav K., Parag K., "Performance Analysis of Supervised Intrusion Detection System", IJCA Special Issue on Network Security and Cryptography, NSC, 2011.

[21] Oswal S. and Shah G, "A study of Data Mining Techyniques on Healthcare Issues and its uses and Application on Health Sector", IJESC, Vol. 7, Issue No. 6, 2017.

[22] Patil D, Agrawal B, Andhalkar S, Biyani R, Gund M, and Wadhai V, "An Adaptive parameter free data mining approach for healthcare application", IJACSA, Vol.3, No.1, 2012.

[23] Iyer, A., Jeyalatha, S., and Sumbaly, R, "DIAGNOSIS OF DIABETES USINGCLASSIFICATION MINING TECHNIQUES, International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.5, No.1, January 2015.

[24] C. M. Velu and K. R. Kashwan, "Visual Data Mining Techniques for Classification of Diabetic Patients", 3rd IEEE International Advance Computing Conference (IACC), 2013.

[25] Sankaranarayanan.S and Dr Pramananda Perumal.T, "Predictive Approach for Diabetes Mellitus Disease through Data Mining Technologies", World Congress on Computing and Communication Technologies, 2014, pp. 231-233.

[26] Mostafa Fathi Ganji and Mohammad Saniee Abadeh, "Using fuzzy Ant Colony Optimization for Diagnosis of Diabetes Disease", Proceedings of ICEE 2010, May 11-13, 2010.

[27] Berson A., Smith S. and Thearling K., "Building Data Mining Applications for CRM", McGraw-Hill Professional Publishing, New York, USA, 2000

[28] Chaudhuri S. "Data Mining and Database Systems: Where is the Intersection?" IEEE Bulletin of the Technical Committee on Data Engineering, Vol. 21, No.1, PP. 4-8, 1998.

[29] D. J. Hand and K. Yu. "Idiots bayes not so stupid after all", 2001.

[30] Chaudhuri S. "Data Mining and Database Systems: Where is the Intersection?" IEEE Bulletin of the Technical Committee on Data Engineering, Vol. 21, No.1, PP. 4-8, 1998.

[31] Mohamed Taha, Hamed Nassar, Tarek Gharib and Ajith Abraham, An Efficient Algorithm for Incremental Mining of Temporal Association Rules, Data and Knowledge Engineering, Elsevier Science, Netherlands, 69:800-815, 2010.

[32] Swagatam Das, Sambarta Dasgupta , Arijit Biswas, Ajith Abraham and Amit Konar, On Stability of the Chemotactic Dynamics in Bacterial Foraging Optimization Algorithm, IEEE Transactions on Systems Man and Cybernetics - Part A, IEEE Press, USA, 39(3): 670-679, 2009.

[33] Tibebe Tesema, Ajith Abraham and Crina Grosan, Rule Mining and Classification of Road Accidents Using Adaptive Regression Trees, International Journal of Simulation Systems, Science & Technology, UK, 6(10-11):80-94, 2005.

[34] Hongbo Liu, Ajith Abraham and Maurice Clerc, Chaotic Dynamic Characteristics in Swarm Intelligence, Applied Soft Computing Journal, Elsevier Science, 7(3):1019-1026, 2007.

[35] Lizhi Peng, Bo Yang, Yuehui Chen and Ajith Abraham, Data Gravitation Based Classification, Information Sciences, Elsevier Science, Netherlands, 179(6):809-819, 2009.

[36] Hesam Izakian, Behrouz Ladani, Kamran Zamanifar and Ajith Abraham, A Particle Swarm Optimization Approach for Grid Job Scheduling, Third International Conference on Information Systems, Technology and Management (ICISTM-09), Communications in Computer and Information Science, Springer Verlag, Germany, ISBN 978-3-642-00404-9, pp. 100-109, 2009.