# Building Model For Predicting Readmission Rate of Diabetic Patients

Tanmay Kank
*Masters in Computer Science*
*Student Id:1110177*
Section: COMP 5411 FA
Lakehead University

Karan Sood
*Masters in Computer Science*
*Student Id:1099550*
Section: COMP 5411 FA
Lakehead University

Simranjeet Singh
*Masters in Computer Science*
*Student Id:1093604*
Section: COMP 5411 FA
Lakehead University

***Abstract*—**

*A. Background*

**Feature selection and predicting the models with these extracted features can aid us to deliver future directions to the hospitals to have a specific reliable treatment for non-ICU diabetic patients. These historic patterns of diabetic care which were given to individuals who were admitted in hospitals can help to improve the care and lower the expense of the subsequent patients.**

*B. Objective*

**The study aims at selecting features that can best predict the readmission rate of patients. We also want to see if HbA1c also happens to be one of the features, as hypothesized in previous work, that can affect the readmission.**

*C. Methods*

**For feature selection, the techniques that were used are information gain, Extratree classifier, and recursive feature elimination. The first technique, information gain is the filter approach and the other two Extratree classifiers, and recursive feature elimination is the wrapper approach. The algorithms that were implemented were Random Forests, it is a classification method that tries to find a sweet spot between under-fitting and overfitting by constructing a multitude of decision trees during the training phase. The second one is SVM, in which the biggest advantage is the kernel trick, which transforms our data and from these transformations finds the optimal boundaries between the outputs. At last, we have used the Knn model which classifies by using the k-nearest neighbor.**

*D. Results*

**We find 35 important features that are highly related to our target variable, which is the rate of readmission. The accuracies on these features for the model Random Forest, SVM, and Knn is 64 percent, 61 percent, and 57 percent respectively.**

***Index Terms*—**

## I. Introduction

With the development of living standards, diabetes has become increasingly common in people's daily life. Therefore, top studies related to accurately diagnose diabetes is worthy of studying. The management of hyperglycemia or HbA1c test in the hospitalized patient has a significant bearing on the outcome, in terms of both morbidity and mortality[1,2]. This test tells us how well we are controlling our diabetes. Currently, HbA1c is considered as a marker of whether a diabetic patient would be re-admitted to the hospital or not. We are trying to find the features besides with HbA1c that is affecting the readmission ratio.

There are formalized protocols developed in the intensive care unit (ICU) setting with rigorous glucose targets in many institutions. However, the same cannot be said for most non-ICU inpatient admissions. The present analysis of a dataset was undertaken to examine historical patterns of diabetes care in patients with diabetes patients admitted to a US hospital and to inform future directions which might lead to improvements in patient safety[10]. The greater attention to diabetes reflected in the determination of HbA1c along with other explored features that may improve patient outcomes and lower the cost of inpatient care.

It is increasingly recognized that the management of hyperglycemia in hospitalized patients has a significant bearing on the outcome, in terms of both morbidity and mortality [1, 2]. This recognition has led to the development of formalized protocols in the intensive care unit (ICU) setting with rigorous glucose targets in many institutions [3]. However, the same cannot be said for most non-ICU inpatient admissions. Rather, anecdotal evidence suggests that inpatient management is arbitrary and often leads to either no treatment at all or wide fluctuations in glucose when traditional management strategies are employed. Although data are few, recent controlled trials have demonstrated that protocol-driven inpatient strategies can be both effective and safe [4, 5]. As such, the implementation of protocols in the hospital setting is now recommended [6, 7]. At present, it was examined the use of HbA1c as a marker of attention to diabetes care in a large number of individuals identified as having a diagnosis of diabetes mellitus. So it was hypothesized that the measurement of 'only' HbA1c is associated with a reduction in readmission rates in individuals admitted to the hospital[10].

The multivariable logistic regression is used in previous work to fit the relationship between the measurement of HbA1c and early readmission while controlling for covariates in the experiment[1]. So, the results are shown based on only this model. Here we will try to determine other features that might also affect the readmission rate. We use feature selection methods such as ExtraTressClassifier, Information Gain and

Recursive Feature Elimination to determine the subset of features that can best predict the readmission. Then, we select the top 35 features because PCA tells us that we can capture 98 percent data variance using only these many features. We, then, implement models such as KNN, Random Forests and SVM to determine which model can make the best predictions.

So far here we have achieved 35 important features from the ExtraTree classifier feature selection method. Also, the model that gave the maximum accuracy on these features was the Random forest.

## II. Data

The dataset is obtained from different hospital physical examinations as well as integrated delivery networks in the USA [10]. The data set is an ade-identified abstract of the Health Facts database. The data was prepared by John Clore, Krzysztof J. Cios, Jon DeShazo, Beata Strack and was submitted on behalf of the Center for Clinical and Translational Research, Virginia Commonwealth University, a recipient of NIH CTSA grant UL1 TR00058 and a recipient of the CERNER data. The data has been collected over 10 years (1999-2008) and consists of data from clinical care of 130 various US hospitals and integrated delivery networks. The data is publicly available on the UCI Machine LearningRepostiory for the research purpose. The data set has been extracted from the original database by using the following criteria:

- It is a hospital admission
- Another entry in the list
- Any kind of diabetes if entered the system as a diagnosis.
- Length of stay at the hospital was at least for a day and a max of 14 days.
- Laboratory tests were performed during the stay.
- Medications were provided during the stay.

This segregation of dataset from the original database (Health Facts Database) was done due to incomplete, redundant and noisy information. The original database contained 41 tables in a fact-dimension schema, comprising of 117 features, 74,036,643 unique visits, and 17880231 patients. The data set that is used by us contains 50 features and 1,00,000 instances.

## III. Methods and Tools

### A. PREPROCESSING STEPS

From Domain Knowledge, we came to know that there are certain columns that doesn't impact the response variable. Two of such columns are "encounterid" and "patientnumber" since both of them don't have any kind of correlation between our class variable i.e. "readmitted". We removed them before tackling the relevant ones.

*1) MAPPING RELEVANT COLUMNS:* We used "Label Encoder" from "sklearn" to encode the values of the Attribute into our Dataset. There are a lot of columns that have categorical features such as "age", "gender", 22 types of medication results that indicate the level of DOSAGE whether it went up or down or remain stable. To deal with all that stuff, we mapped some of the columns. However, "Age" column in our dataset has values in the form of Range such as (10-20] rather that discreet values such as 15. We used label encoder to map them. Moreover, the categorical values "Male" and "Female" into, "Gender" column are simple mapped to values 0 and 1 respectively. In "Races" Column, we normally have ["Caucasian", "African American" "Asian" "Hispanic"] as races which are encoded to values [2, 1, 0, 3] respectively.

*2) FEATURE SELECTION METHODS:* Feature Selection is the process of selecting only relevant features from the dataset so that our dataset gets free from all the irrelevant attributes. Methods applied for feature selection includes: -

1) Filter Approach
2) Wrapper Approach

In Filter Approach, we used Information Gain method to get the best features.

**INFORMATION GAIN**

Information Gain (IG) is an entropy-based feature evaluation method, widely used in the field of machine learning. It is defined as the amount of information provided by the feature items i.e. "Goodness of a Feature". • Features that perfectly partition should give maximal information. • Unrelated features should give no information. • It measures the reduction in entropy. – Entropy: (im)purity in an arbitrary collection of examples.

Inside Wrapper Approach, we used two methods namely "Chi Square Test" and "Extra Tree Classifier" for Feature Selection.

**RECURSIVE FEATURE ELIMINATION**

Recursive Feature Elimination (RFE) is a backward selection method that fits a model and removes the weakest feature (or features) until the specified number of features is reached. This technique begins by building a model on the entire set of predictors and computing an importance score on each predictor. The subset size that optimizes the performance criteria is used to select the predictors based on the importance rankings. RFE requires a specified number of features to keep, however it is often not known in advance how many features are valid. To find the optimal number of features cross-validation is used with RFE to score different feature subsets and select the best scoring collection of features. RFE can be an effective and relatively efficient technique for reducing the model complexity be removing irrelevant predictors

**EXTRA TREE CLASSIFIER**

Extra Tree classifier is a type of ensemble learning technique which aggregates the results of multiple de-correlated decision trees collected in a "forest" to output its classification result. In concept, it is very similar to a Random Forest Classifier and only differs from it in the manner of construction of the decision trees in the forest. Each Decision Tree in the Extra Trees Forest is constructed from the original training sample. Then, at each test node, Each tree is provided with a random sample of k features from the feature-set from which each decision tree must select the best feature to split the data based on some mathematical criteria (typically the Gini Index). This random sample of features leads to the creation of multiple de-

correlated decision trees. Again, Top 25 features are selected on the basis if Feature Score in Extra Trees Classifier.

**SELECTING NUMBER OF FEATURES**

We used PCA (Principal of Component Analysis) for selecting a total of 35 features from the original dataset. PCA simplifies the complrxity nature in high-dimensional data while holding patterns and examples. It does this by changing the data into less dimensions, which acts as summaries of Features. High-dimensional information are exceptionally basic in medical and emerge when different features. This kind of information introduces a few difficulties that PCA mitigates: computational cost and an expanded mistake rate because of different multiple test correction when testing each feature for relationship with a result. PCA is an unaided learning technique and is like clustering1—it discovers designs without reference to prior information about whether the examples originate from various treatment gatherings or have phenotypic contrasts.

## IV. LEARNING METHODS

### *A. KNN Classifier*

In this supervised machine learning algorithm, the output is a class membership. An object is classified by a plurality vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors If k = 1, then the object is simply assigned to the class of that single nearest neighbor. Distance measures used to calculate distance between features are Euclidean Distance for continuous features and Hamming Distance for Categorical ones.

$$\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$$

Fig. 1. Euclidean Distance Formula

$$D_H = \sum_{i=1}^{k}|x_i - y_i|$$

Fig. 2. Hamming Distance Formula

In this algorithm, we did Hyper Parameter tuning on k by taking different values of k between 1 and 100 and calculating accuracy for each value using stratified cross validation. The value of k giving highest accuracy among the tested values is selected.

### *B. Random Forest*

Random Forest Classifier consists of a large number of Decision Trees that works as an ensemble. Here every individual tree predicts its own class variable and the class with the most votes becomes our model's final prediction When training the model, each tree in a Random Forest gains from an arbitrary sample of the data points. The samples are drawn with substitution, known as bootstrapping, which implies that a few samples will be utilized on numerous occasions in a single tree. The thought is that via training each tree on various samples, although each tree may have high variance as for a specific set of the training data, generally, the whole forest will have lower variance yet not at the expense of increasing the bias.

For Hyper parameter tuning, we used the following parameters: -

1. n-estimators

The n-estimators parameter specifies the number of trees in the forest of the model. The default value for this parameter is 10, which means that 10 different decision trees will be constructed in the random forest. Usually the higher the number of trees the better to learn the data. However, adding a lot of trees can slow down the training process considerably, therefore we do a parameter search to find the sweet spot.

2. min-samples-split

The min-samples-split parameter specifies the minimum number of samples required to split an internal leaf node. The default value for this parameter is 2, which means that an internal node must have at least two samples before it can be split to have a more specific classification. When we increase this parameter, each tree in the forest becomes more constrained as it has to consider more samples at each node.

3. min-samples-leaf

The min-samples-leaf parameter specifies the minimum number of samples required to be at a leaf node. The default value for this parameter is 1, which means that every leaf must have at least 1 sample that it classifies.

4. max-features

The number of features to consider when considering the best split.

5. max-depth

The max-depth parameter specifies the maximum depth of each tree. he deeper the tree, the more splits it has and it captures more information about the data. We fit each decision tree with depths ranging from 10 to 107 and plot the training and test errors.

After tuning the parameters the final values are

- n-estimators = 733
- min-samples-split=10
- min-samples-leaf=2
- max-features=auto
- max-depth=87

The values used for Tuning Hyper Parameters in Random Forest Model using Grid Search Cross validation have been selected randomly. It might be possible that by selecting different subset of values for Hyper Parameters, we are further able to enhance the efficiency of model.

### C. Support Vector Machine

The goal of the Support vector machine Algorithm is to discover a hyperplane in a N-dimensional space (where N is the number of features) that particularly characterizes the data points. To separate the two classes of data points, there are numerous conceivable hyperplanes that could be picked. Our goal is to locate a plane that has the most extreme edge, i.e the greatest separation between data points of the two classes. Maximizing the edge separation gives some reinforcement so that the future data points can be classified with more certainty. data points falling on either side of the hyperplane can be classified to various classes. Likewise, the component of the hyperplane relies on the quantity of features inside dataset. Moreover, if the number of input features is 2, at that point the hyperplane is only a line. In the event that the number of features is 3, at that point the hyperplane turns into a two-dimensional plane. It gets hard to imagine when the quantity of highlights surpasses 3. In this algorithm we try to maximize the margin between data points and hyperplane. Hinge loss is the function used to maximize the margin.

$$c(x, y, f(x)) = \begin{cases} 0, & \text{if } y * f(x) \geq 1 \\ 1 - y * f(x), & \text{else} \end{cases}$$

Fig. 3.

$$c(x, y, f(x)) = (1 - y * f(x))_+$$

Fig. 4.

The cost is 0 if the predicted value and the actual value are of the same sign. If they are not, we then calculate the loss value. For tuning the c in the linear SVM (Support Vector Machine), here we have used the Grid Search Cross validation where the values are used for tuning where taken exponentially i.e C= [0.1, 1, 10, 100, 100] out which the best value of c comes out to be 0.1 for which we got the maximum accuracy.

## V. VALIDATION METHOD

### A. K-FOLD CROSS VALIDATION

This method is basically used to estimate the skill of the trained model on the test dataset. This method has a single parameter considered k that refers to the number of grops that a given dataset is to be separated into. All things considered, the strategy is frequently called k-fold cross validation. At the point when a particular value for k is picked, it might be utilized instead of k in the reference to the model, for example, k=5 turning out to be 5-fold cross-validation. An ineffectively picked value for k may bring about a mis-representative thought of the ability of the model, for example, a score with a high variance (over-fitting), or a high bias, (under-fitting).

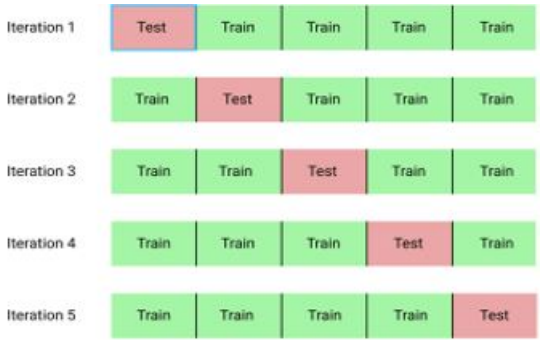


Fig. 5. k-fold cross validation

## VI. TOOLS/LIBRARIES

**1) Sklearn.model_selection.train_test_split**

This is used to split the dataset into training data and test data

**2)KNeighborsClassifier**

This library is used to run KNN Classifier

**3)Pandas, numpy, math**

We used these libraries to load data into dataframe and performing some mathematical operations on dataframe

**4)sklearn.base.TransformerMixin**

This is used to impute values into dataframe

**5)sklearn.metrics.accuracy_score**

This is used to get the accuracy of the model

**6)sklearn.metrics.confusion_metrics**

This library is used to get the confusion metrics for classification model

**7)sklearn.metrics.roc_curve**

this is used to get the True Positive Rate and False Positive rate values

**8)sklearn.metrics.roc_auc_score**

this is used to get the area under Receiver Under Characteristic (ROC)

**9)matplotlib.pyplot.plt**

This library is used to plot the result

**10)Sklearn.feature_selection.mutual_info_class_if**

This is used to calculate Information Gain

**11)Sklearn.preprocessing.LabelEncoder**

This is used to encode the categorical variable of the dataset

**12)Sklearn.preprocessing.MinMaxScaler**

This is used to normalize the data

**13)Sklearn.decomposition.PCA**

This is used to select top features which are required using feature selection technique

**14)Sklearn.model_selection.GridSearchCV**

This is used for Exhaustive search over specific parameter values over a model

**15)Sklearn.svm.SVC**

This is used to implement SVM Classifier

| | ACCURACY | FALSE NEGATIVE | SPECIFICITY | SENSITIVITY |
|---|---|---|---|---|
| K NEAREST NEIGHBOURS | 0.54 | 9045 | 0.51 | 0.56 |
| SUPPORT VECTOR MACHINE | 0.58 | 8565 | 0.55 | 0.48 |
| RANDOM FOREST CLASSIFIER | 0.59 | 7521 | 0.56 | 0.60 |

Fig. 6. Information gain

| | ACCURACY | FALSE NEGATIVE | SPECIFICITY | SENSITIVITY |
|---|---|---|---|---|
| K NEAREST NEIGHBOURS | 0.54 | 9045 | 0.51 | 0.56 |
| SUPPORT VECTOR MACHINE | 0.58 | 8621 | 0.54 | 0.55 |
| RANDOM FOREST CLASSIFIER | 0.59 | 7532 | 0.56 | 0.60 |

Fig. 7. Recursive Feature Elimination

## VII. TABLE AND PLOTS

## VIII. DISCUSSION AND CONCLUSION

### A. PRINCIPLE WORK DONE

We have executed the attributes that can best identify whether a patient will be readmitted to the hospital or not. We have used both filter and wrapper approach for feature selection. In Filter based approach, we have used "Information Gain" to calculate the importance of each feature whereas "Extra Tree Classifier" and "Recursive Feature Elimination" have been used for wrapper approach. Using PCA (Principle Component Analysis), we found that 35 Features can capture almost 98% of variance of entire dataset. Hence, we selected top 35 features amongst those obtained in these methods.

### B. RESULTS

As shown in the above table (Fig: - 6,7,8) Random Forest Classifier always outperform KNN in terms of accuracy. There always exists a difference of 5% in terms of accuracy. Also, the number of patients who actually have diabetes but are classified as healthy, is also less in case of Random Forest and Ensemble learning methods and thus, have better ability to reduce bias and variance. If we compare between all three feature selection methods, we found that we are getting better results using Extra Tree Classifier for both KNN and Random Forest. For Example, the "Accuracy" of KNN using the features selected by Extra Tree Classifier is 0.58 whereas it is 0.54 when we select features using Recursive Feature Elimination and Information Gain. Moreover, the number of false negative is also less. Similar results can be seen in case of Random Forests where we achieve an accuracy of 0.64 using

| | ACCURACY | FALSE NEGATIVE | SPECIFICITY | SENSITIVITY |
|---|---|---|---|---|
| K NEAREST NEIGHBOURS | 0.57 | 8838 | 0.55 | 0.58 |
| SUPPORT VECTOR MACHINE | 0.48 | 7342 | 0.59 | 0.46 |
| RANDOM FOREST CLASSIFIER | 0.64 | 6620 | 0.63 | 0.64 |

Fig. 8. Extra Tree Classifier

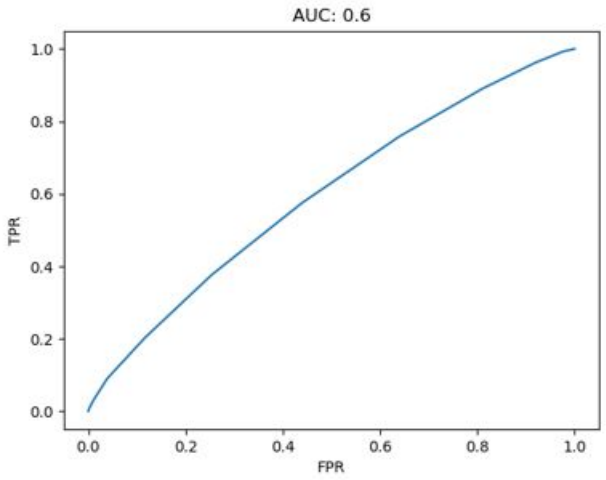


Fig. 9. Extra Tree Classifier

features selected by Extra Tree Classifier whereas it is 0.59 using other two methods. Furthermore, we have almost 900 less False Negatives using Extra Tree Classifier.

### C. MAJOR CONTRIBUTION

On the basis of above results, we came up with a list of 35 features selected by Extra Tree Classifier that have greater importance in determining whether a treated diabetic patient would be readmitted to the hospital or not. In earlier work done, Logistic Regression was used for the classification that achieves an accuracy of just .47 whereas we have been able to achieve an accuracy for 0.64. With improved accuracy in prediction, the Hospitals can provide better treatments to patients who have higher chances of re-admission. This would not only help the patients in health improvement but would also help them financially (as the chances of re-admission would be reduced by providing better treatment).

### D. CONCLUSION

We have found the following 35 features using Extra Trees Classifier that have a significant pairing on readmission of patients: -

"num_lab_procedures", "num_mrdications", "time_in_hospital","number_inpatient", "number_diagnosis",

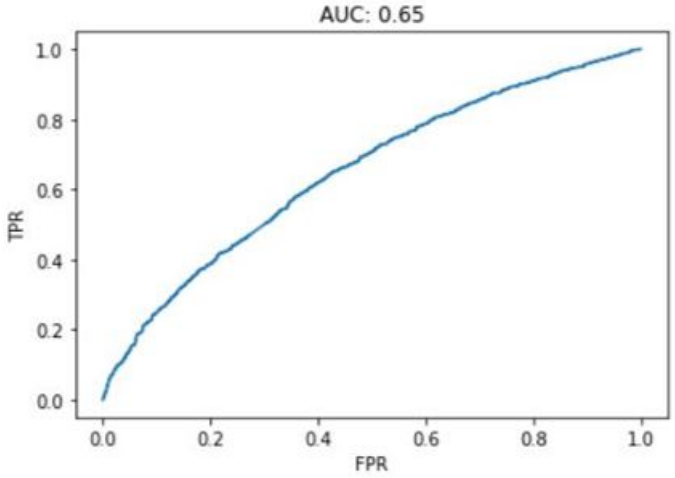


Fig. 10. Extra Tree Classifier

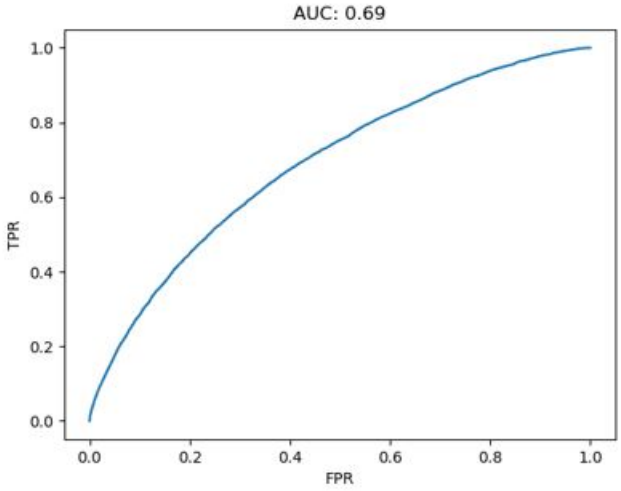


Fig. 11. Extra Tree Classifier

“discharge_disposition_id”, “num_procedures”, “number_outpatient”, “admission_type_id”, “admission_source_id”, “number_emergency”, “repaglinide”, “glimepiride-pioglitazone”, “acetohexamide”, “payer_code”, “chlorpropamide”, “glipizide_metformin”, “acarbose”, “medical_specialty”, “race”, “insulin”, “tolbutamide”, “metformin”, “diag_2”, ‘’troglitazone”, “A1Cresult”, “glimepride”, “pioglitazone”, “metformin_rosiglitazone”, “examide”, “glipizide”, “change”, “diag_1”, “nateglinide”, “max_glu_serum”

Since the dataset is quite old and there have been significant change in the lifestyle (Eating and Living Habits) of people, there may be some other factors as well that could directly affect the re-admission. So, the model can be further improved by collecting more data. The prediction accuracy might further improve by using neural networks. Moreover, the data pertains to only US population. So, it may not work with for patients from other countries who have different physiology and eating habits. Hence, we cannot generalize the model for the global population.

## IX. FUTURE WORK

In future, we shall focus on further improving our features subset using regularization methods such as LA550 (This penalizes the model parameters to avoid overfitting) and boosting methods such as XGBOOST. We also plan to exploit the domain knowledge by consulting medical experts who could further help in selecting the prediction variables to develop a model that can fit on global population.

## X. ACKNOWLEDGMENT

We would like to express our special thanks of gratitude to Dr Quazi Rahman who gave us the golden opportunity to do this project, which also aided me in doing a lot of Research and we came to know about so many new things and explore the noval concepts . Secondly, we are thankful to our university Lakehead university.

## REFERENCES

[1] G. E. Umpierrez, S. D. Isaacs, N. Bazargan, X. You, L. M. Thaler, and A. E. Kitabchi, “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, 2002.

[2] C. S. Levetan, M. Passaro, K. Jablonski, M. Kass, and R. E. Ratner, “Unrecognized diabetes among hospitalized patients,” Diabetes Care, vol. 21, no. 2, pp. 246–249, 1998.

[3] S. E. Siegelaar, J. B. L. Hoekstra, and J. H. Devries, “Special considerations for the diabetic patient in the ICU; targets for treatment and risks of hypoglycaemia,” Best Practice and Research: Clinical Endocrinology and Metabolism, vol. 25, no. 5, pp. 825–834, 2011.

[4] A. G. Pittas, R. D. Siegel, and J. Lau, “Insulin therapy for critically ill hospitalized patients: a meta-analysis of randomized controlled trials,” Archives of Internal Medicine, vol. 164, no. 18, pp. 2005–2011, 2004. View at Publisher · View at Google Scholar ·

[5] A. C. Tricco, N. M. Ivers, J. M. Grimshaw et al., “Effectiveness of quality improvement strategies on the management of diabetes: a systematic review and meta-analysis,” The Lancet, vol. 379, no. 9833, pp. 2252–2261, 2012.

[6] M. C. Lansang and G. E. Umpierrez, “Management of inpatient hyperglycemia in noncritically ill patients,” Diabetes Spectrum, vol. 21, no. 4, pp. 248–255, 2008. View at Publisher · View at Google Scholar ·

[7] R. Vinik and J. Clements, “Management of the hyperglycemic inpatient: tips, tools, and protocols for the clinician,” Hospital Practice, vol. 39, no. 2, pp. 40–46, 2011.

[8] K. J. Cios and G. W. Moore, “Uniqueness of medical data mining,” Artificial Intelligence in Medicine, vol. 26, no. 1-2, pp. 1–24, 2002.

[9] A. Frank and A. Asuncion, UCI Machine Learning Repository, University of California, School of Information and Computer Science, 2010.

[10] Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records Beata Strack,1 Jonathan P. DeShazo,2 Chris Gennings,3 Juan L. Olmo,4 Sebastian Ventura,4 Krzysztof J. Cios,1,5 and John N. Clore6