Type 1 Diabetes Prediction Using Data Mining Techniques

1st Jake E. Leslie  
*Department of Computer Science*  
*University of Detroit Mercy*Detroit, USA,  
leslieje@udmercy.edu  
  
2nd Shermana Akter  
*Department of Computer Science*  
*University of Detroit Mercy*Detroit, USA,  
aktersh2@udmercy.edu  
  
3rd Nathan A. Boyar  
*Department of Computer Science*  
*University of Detroit Mercy*Detroit, USA  
boyarna@udmercy.edu

***Abstract—******Type 1 diabetes is a chronic condition where one’s pancreas produces little to no insulin, which results in individuals using insulin therapy or other forms to keep their blood sugar levels at a healthy standard. Although type 1 diabetes impacts nearly 1.6 million people in the United States alone, some patients can go untreated which can create long-term negative effects on their health. Using data mining, the chances of patients who have type 1 diabetes going untreated can be cut down drastically. This research paper will focus on two datasets and different algorithms to determine the accuracy of identifying type 1 diabetes patients. Some of the sets will focus on looking at a patient’s glucose levels, blood pressure, insulin, diabetes pedigree function, and other traits that can determine if a patient may or may not have the disease. The results determine that the Decision Tree Classifier is the most accurate out of all classification algorithms for both datasets.***

Keywords—Type 1 diabetes, insulin, blood sugar levels, patients

# Introduction

The objective of this paper is to determine which classification algorithm is the best at detecting type 1 diabetes among patients according to their health data. Nearly 1.6 million Americans have Type 1 diabetes [1]. Although Type 1 develops slowly, the longer the patient goes undiagnosed, the more health risks the patient may experience. Also, being misdiagnosed by an inaccurate algorithm can be risky for an individual's health**.** Type 1 diabetes symptoms can be brushed off as other illnesses, which can result in a delayed diagnosis that could lead to diabetic ketoacidosis (DKA) [2] which can be life-threatening for the patient. The lack of insulin can make the body break down fat as fuel. This results in elevated ketones in urine, severe weight loss, blurry vision, extreme thirst, and disorientation. If DKA is left untreated because of usual misconceptions that these could be flu symptoms, the patient can go into a coma or die [2].

By using data mining to lower the chances of patients going undiagnosed to the point of DKA, data mining can instead help early diagnosis and early care. Type 1 is only shown to be projected to grow and affect an even larger population in the future [1]. Being able to diagnose and treat Type 1 in the early stages will result in a better way of life for the population that is affected by it.

The existing techniques to diagnose Type 1 diabetes are a random blood sugar test, a glycated hemoglobin test, and a fasting blood sugar test [3]. Instead of having these tests, if someone is undiagnosed but has frequent visits with the doctors for flu-like symptoms, these algorithms can be used to see if the patient has Type 1 diabetes according to their health. There is a list of datasets that will be used for this paper that have patient data, some of whom are diagnosed with T1D. Both datasets are from Kaggle’s website, the data science company The first set is the *Pima Indians Diabetes Database*, which is originally a dataset from the National Institute of Diabetes and Digestive and Kidney Diseases [4]. The second dataset from Kaggle is the *Early Stage Diabetes Risk Prediction Dataset*, the set comes from a direct questionnaire from patients at the Sylhet Diabetes Hospital in Bangladesh [5]. The objective is the same between these two datasets, which is to predict whether a patient has type 1 diabetes. The only main different for results is that the second data set, Early *Stage Diabetes Risk Prediction Dataset,* looks at type 1 and type 2 diabetic patients.

For the project, classification method will be used to determine which algorithm is the most accurate among all the datasets for the project. The three types of classification that will be used are support vector machines, logistic regression, and decision trees. These three classification methods will help determine which dataset is the most accurate at identifying if a patient has type 1 diabetes positive.

# Related Work

Woldaregay and colleagues [6] evaluated 47 different articles to determine what data mining techniques would be the most viable to analyze blood glucose anomalies in data. This would be helpful to those who are self-managing because it would alert them when blood glucose levels increase, not being knowledgeable on blood glucose levels can lead to organ damage or other bodily complications. They found that artificial neural networks, Gaussian process regression, and Bayesian neural networks among others are the most useful in terms of analyzing the data. Woldaregay and colleagues noted that interpatient and interpatient variation is the biggest issue when it comes to analysis. However, as technologies advance and techniques improve, they believe widespread use is imminent. This relates to our work as well because patient variation is a very big issue for our type 1 diabetes analysis. Features like BMI’s being different from person to person, varying insulin levels, and various illnesses that people can contract, make it hard to properly determine if a patient has diabetes.

Khan and colleagues [7] evaluated 80 reports to evaluate the data mining techniques used to predict and diagnose type 1 diabetes. They analyzed 6 techniques that proved successful towards prediction: classification-based, regression-based, association-based, clustering-based, sequential pattern mining, and a hybrid technique involving multiple different models. Their work which discusses their classification-based techniques was helpful for us so that we can understand how professionals would implement these techniques. Comparing their work to ours helped us understand what steps we can take next, and what steps not to make.

Kavakiotis and colleagues [8] acknowledge that data is only increasing in size and complexity as time moves forward so data mining techniques are crucial to traversing data efficiently and effectively. The authors note that obesity is one of the main risks of type 2 diabetes, which is the inverse to type 1 diabetes as lower BMI’s correlate directly with this variation of DM. They also include that early diagnosis and treatment have positive effects on disease progression, medication selection, life expectancy, and future complications. Kavakiotis and colleagues grouped 103 studies based on the outcomes of the data: diabetes prediction and diagnosis, complications, remedies, genetic background, and health care expectancy.

Marinov and colleagues [9] believe that diabetes is a serious disease and proper diagnosis, and treatment are pivotal for survival of the patients. 17 studies were analyzed and grouped based on interpretation and prediction of blood glucose levels, features selection, genomic data analysis, and other methods. They affirm that data mining will help future diabetes patients. This relates to our goal because our goal is to help future diabetes patients by finding the most consistent and accurate classification model so that these patients can find out their status sooner, and with more precision than before. If a patient is misdiagnosed with type 1 diabetes it can have severe complications if they start taking medication. The opposite is also true if a type 1 patient goes misdiagnosed because it can lead to severe issues and even death.

Duhamel and colleagues [10] examine the European Diabcare project. Founded in 1990, its goal was to assess the quality of care the diabetes patients were receiving. The researchers used a French database dating back to 1994 containing 32,551 records filled with information about patients with diabetes. Duhamel and colleagues then preprocessed and analyzed the data. They initially found 360 errors within the data which were then corrected. Then coherence rules were developed to correct any other errors that may occur; they found 756 additional errors. Cluster analysis was then performed which resulted in a surprisingly large amount of 2,367 incorrect files, which were then disposed of. Logistic regression and decision trees strengthened their conclusions on the errors, and predictions were somewhat scrambled. They fixed this by creating another category named “missing” to prevent statistical bias. This relates to our work because we applied the Logistic Regression and Decision Tree Classifier models to predict type 1 diabetes so seeing how the researchers applied the model and their results helped us in doing our own research.

# Methodology

## **Figure 1** displays the overall approach for the analysis of the two datasets with multiple classification algorithms

**Diagram

Description automatically generated**

## Datasets

The first dataset is the *Pima Indians Diabetes Database.* The dataset has nine columns and 768 rows. The columns are Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, Age, and the class column Outcome [4]. The dataset was pre-selected by the creator to only include women 21-years and older with the average age being 33. Finally, there were more non-diabetic women than those positive for diabetes.

**Table 1** displays all the attributes of the *Pima Indians Diabetes Database* dataset description of them

|  |  |
| --- | --- |
| Attribute | Description |
| Pregnancies | How many times the patient has been pregnant |
| Glucose | Simple sugar that is important for energy |
| BloodPressure | Pressure of circulating blood |
| SkinThickness | Collagen content |
| Insulin | Hormone that regulates the amount of glucose in the blood |
| BMI | Body Mass Index |
| DiabetesPedigreeFunction | Diabetic history in the family |
| Age | Age of the patient |
| Outcome | The results of whether the patient was positive or negative for type 1 diabetes |

The second dataset is the *Early Stage Diabetes Risk Prediction Dataset*. The dataset has 17 columns and 520 rows. The columns are Age, Gender, Polyuria, Polydipsia, SuddenWeightLoss, Weakness, Polyphagia, GenitalThrush, VisualBlurring, Itching, Irritability, DelayedHealing, PartialParesis, MuscleStiffness, Alopecia, Obesity, and the class column Class [5]. This dataset had more positive cases for diabetes than negative. Compared to the *Pima Indians* *dataset*, this dataset did not have an age or gender restriction.

**Table 2** displays all the attributes of the *Early Stage Diabetes Risk Prediction Dataset* and description of them

|  |  |
| --- | --- |
| Attribute | Description |
| Age | The age of the patient |
| Gender | The gender of the patient |
| Polyuria | Abnormal amount of diluted urine |
| Polydipsia | Abnormal amount of thirst |
| SuddenWeightLoss | Concerning weight loss |
| Weakness | The patient is weaker than others |
| Polyphagia | Excessive hunger |
| GenitalThrush | A yeast infection |
| VisualBlurring | Visual impairment |
| Itching | Skin being itchy |
| Irritability | Being irritable |
| DelayedHealing | Longer healing time for wounds |
| PartialParesis | Weakening of muscle/s |
| MuscleStiffness | Muscle/s feeling tight |
| Alopecia | Autoimmune disorder that causes hair loss |
| Obesity | Overweight |
| Class | Whether the patient is positive or negative for diabetes |

## Data Preprocessing

There were many steps taken to process this data. We started with the *Pima Indians* dataset by checking if there were any null values, after getting the result there turned out to zero null values. Moving on to filling any zero values inside of the features, our group filled in the columns: Glucose, BMI, SkinThickness, and Insulin with the median value calculated from those columns. To fill in the zero values for BloodPresure, the mean of the column was computed and then that mean value replaced the zeroes. Our next step was to make a heatmap of the modified data to see the significant features of the dataset. Looking at the results of the heatmap, it showed that all the features were relevant with no signs of negative correlation, leading to no features being dropped. We checked for any outliers inside of the dataset which resulted in a few being found, we ran our custom function get outliers, and calculated the interquartile range (IQR) as well as adding all outliers to a list and returning multiple outliers. Running this function located any outliers in the dataset and we then removed them from the dataset by dropping the values at the index of the outliers in place. After the outliers were dropped, the data preprocessing portion of the experiment was completed.

For the *Early Stage Diabetes Risk Prediction Dataset,* we took a different approach because the dataset was initially in nominal format. All the columns were yes or no, except for age which was numerical, and Class which was either positive or negative. First, we ran a preprocessing function called LabelEncoder(), and then created another function called nomTonumLabel which took in a column and let us fit and transform the columns data into 0’s or 1’s. This allowed our data to be easier to work with because our models needed the data to be in a numerical format. We checked for any null values in the columns and there were none. The heatmap was then made to analyze our feature importance and we noticed that there were a few negatively correlated features. The most important features were Polyuria and Polydipsia, they had a value of 0.67 and 0.65. The negatively correlated features were Gender, Alopecia, and itching. Gender had a value of -0.45, Alopecia had a value of -0.27, and Itching had a value of -0.1. As a result, these three features were removed from the dataset due to the negative correlation. After that, outliers were checked in the age column, and then the custom function was implemented to remove any possible outliers. That was how the data was preprocessed which made sure that it was a clean and relevant dataset to apply models on.

## Model Selection

Stratified Cross-Validation was used to evaluate both of our models. It was given ten folds so that an even more accurate result would be produced since more testing would be happening. Due to our data in both of our sets having a higher percentage of one class outcome to the other, it would not be sufficient to use standard Cross-Validation. Cross-Validation does not account for the fact that the classes are not balanced, whereas the Stratified form does. Using this method leads to both class outcomes being equally represented, eliminating any of the outcomes from being overrepresented. What this can do is skew results and make the classes biased which is not the fairest way of testing the accuracy and consistency of these models. Other than Stratified Cross-Validation, we also ran a classification report, so that we can see the weighted averages of the precision, recall, and f1-scores. These scores are important because it effectively tells us if the model did a good job at classifying the dataset. The precision score is intuitively the ability of the classifier not to label as positive a sample that is negative [11]. The recall score is intuitively the ability of the classifier to find all the positive samples [11]. Finally, the f1-score is a weighted harmonic mean of the precision and recall, where an F-beta score reaches its best value at 1 and worse score at 0 [11]. When evaluating a model, it may return a high accuracy score which looks very good, but if it has a low f1-score you cannot trust that the model deserves that score. This can be tragic as it may misdiagnose people with type 1 diabetes or any other disease.

## Evaluation

The analysis approach for this project on diabetes prediction was classification. The reason why we decided on this approach compared to clustering and regression is that we needed to identify patient by patient. Our goal was not to identify a possibility of a group of patients being T1D positive. The goal was to check if a full table that showed which patient was or negative was accurate. By doing classification to determine the accuracy of the datasets we chose, we were able to solidly portray the results. Clustering would have been difficult to use in this project because it would only allow the groups to come together on their possibility of being positive or negative for T1D. Regression would only discuss the quantity of the data that we were using. Classification is perfect because it puts a label on our results depending on what attributes we are looking at. Through classification, we used Stratified Cross-Validation, and a classification report. The benefits are highlighted prior, but it allows us to filter our data accurately and display the results according to our findings.

# Results and Discussion

From the conduction of our experiments to determine which machine learning model is the most accurate and consistent for predicting type-1 diabetes, we obtained key results in determining this. We received the heatmaps of the datasets showing us the feature importance, the accuracy score of the model’s classification efforts, and finally the precision, recall, and f1-score to tell us the accuracy of the model’s predictions.

**Figure 2** displays the heatmap for the *Pima Indians Diabetes dataset*

**Chart, treemap chart

Description automatically generated**

**Figure 3** displays the heatmap for the *Early Stage Diabetes Risk Prediction Dataset*. A picture containing table

Description automatically generated

**Figure 4** displays the accuracy score of the models computed with

10-fold Stratified Cross-Validation on the two datasets

Table

Description automatically generated

**Figure 5** displays the precision and recall scores of the models from the *Pima Indian Diabetes Dataset*

Graphical user interface, text, application

Description automatically generated

**Figure 6** displays the precision and recall scores of the models from the *Early Stage Diabetes Risk Prediction Dataset*

Graphical user interface, application

Description automatically generated

**Figure 7** displays the computed F1-Scores of all three of the models over both of the datasets.Chart, bar chart

Description automatically generated

Before we discuss the models, we must look back at the feature importance and their scores so that we can get a better sense of the dataset and why we removed what we did. Looking at the heatmap in **Figure 2**, we see that there are all positive values. What this means is that all the features in this dataset are relevant to the outcome and should not be removed. The most relevant features in the *Pima Indian Diabetes Dataset are* glucose at 0.49, and BMI at 0.31. The reason that glucose is so relevant to type 1 diabetes classification is because without enough insulin, glucose builds up in the bloodstream instead of going into the cells. This buildup of glucose in the blood is called hyperglycemia. The body is then unable to use glucose for energy. This leads to the symptoms of type 1 diabetes [12]. Since type 1 diabetes affects insulin production, it then causes glucose to be built up in the bloodstream, leading to that being a feature of type 1 diabetes. BMI is also relevant to type 1 diabetes classification because if a patient’s blood sugar levels are consistently high, patient may find it more difficult to put on weight. It is important for patients to not allow their blood sugar levels to run at high rates for an extended period. If they allow this to happen, it can lead to serious diabetic complications to occur [13]. This shows the relevance of BMI to type 1 diabetes. Lower BMI can occur due to high levels of sugar in the blood which is a common occurrence for patients with this illness.

 Looking at the heatmap in **Figure 3**, we can see that there are several negative values which stand out to us because those values are negatively correlated to the dataset which should be removed. The features with negative correlation are Gender at -0.45, Alopecia at -0.27, and Itching at -0.01. These are negative because they have a negative correlation with type 1 diabetes, so they are removed. Gender is not known to be related to type 1 diabetes. Alopecia is a separate disease that is the loss of hair, and Itching has not been proven to relevant to this dataset according to the heatmap. Further research has shown us that Alopecia and itchiness are related to type 1 diabetes although our heatmap said otherwise. The most relevant features in this dataset are Polyuria at 0.67 and Polydipsia at 0.65. Polyuria is related to diabetes because normally when someone’s kidneys create urine, the sugar is reabsorbed and is directed back into the bloodstream. With type 1 diabetes, patients have excess glucose that ends up in their urine, which results in producing more water and more urine [14]. Polydipsia is being excessively thirsty; it is caused by the results of high blood sugar levels [15]. Most health problems correlated with type 1 diabetes relates to glucose and blood sugar levels. This is why in the *Pima Indians* dataset; glucose is such a relevant feature to the outcome of the dataset.

In **Figure 4**, we have the accuracy percentages of all three of the models on both datasets. In our findings, we found that Support Vector Machines (SVM) had the highest accuracy score on both datasets, with 97.37% on the *Pima Indians Diabetes Dataset,* and 99.80% on the *Early Stage Diabetes Risk Prediction Dataset,* which leaves us with an overall accuracy percentage of 98.59%. While Decision Tree Classifier had an average accuracy of: 83.26% and Logistic Regression with 84.80%. The average percentage for Support Vector Machine is very high therefore at first it would be safe to think that this is the best score, but to decide our answer we must look at the f1-scores. As discussed earlier, having a higher f1-score means that we can trust the accuracy of a model more. The f1-score for Support Vector Machine on dataset 1 is 0.78 and on the second data set, it is 0.60. That is an average of 0.69. That means that there is a likely chance that the results are not true, which means that there was a significant amount of people that were misclassified. Compared to the other algorithms, Decision Tree Classifier has an average f1-score of 0.99, and Logistic Regression at 0.85. Decision Tree Classifier has a very high f1-score meaning that the predictions that it made were almost fully accurate amongst both datasets. With its average accuracy score being 83.26%, we believe that it is the most consistently accurate machine learning model, even over the Support Vector Machine. SVM may be 15.33% higher regarding accuracy score, but with its f1-score being 0.30 lower than Decision Tree Classifier, it is certain that the Decision Tree Classifier is the most consistent and accurate machine learning model for predicting type 1 diabetes according to our experiments.

# Conclusion

Data mining is an incredibly useful tool by compiling incredible amounts of data to solve problems that would be impossible to do by hand. In our case, understanding correlations in the data can help susceptible patients gain not only knowledge about type 1 diabetes but can help manage the lifelong illness. Type 1 diabetes affects a significant portion of the population; therefore, it is important that there should be efforts made to find the most accurate classification model to predict this disease so people can get the care they need faster. We learned from other similar reports that discuss type 1 diabetes and data mining, how important the issue is. These multiple reports also discuss the diagnosis of type 1 diabetes growing around the world. We can assume that this is because the way to diagnose type 1 diabetes is becoming a quicker process as technology advances. Additionally, but more importantly to our studies, we learned what data mining techniques would be the most useful to analyze patient data to determine if they are T1D positive.

Our work focuses on two datasets that have patient information to help determine if someone is T1D positive. We used different algorithms to conclude on which algorithm is most accurate when it came to the datasets. Through our methods of analysis, we found that the Decision Tree Classifier was the most accurate data mining model in terms of consistency and outcomes. We hope that scientists can use data mining techniques in order to predict those who may have type 1 diabetes before it gets to a point where they have long-lasting issues. As we learned from Kavakiotis and colleagues [8], early diagnosis can slow disease progression, improve medication selection, prolong life expectancy, and reduce future complications. Thus, data mining can be used to improve the lives of those who must type 1 diabetes. This type of algorithm-based data mining on datasets related to patient health can lead to early diagnosis for other illnesses as well. As data mining continues to progress and be more accurate, the hope is to help all patients in need, not just T1D positive patients.

##### References

1. “Facts,” Juvenile Diabetes Research Foundation. [Online]. Available:

https://www.jdrf.org/t1d-resources/about/facts/. [Accessed: 01-Dec-2021].

1. “All About DKA,” *Beyond Type 1*. [Online]. Available: https://www.diabetesresearch.org/diabetes-statistics. [Accessed: 01-Dec-2021].
2. “Type 1 Diabetes,” *Mayo Clinic.* [Online]. Available: https://www.mayoclinic.org/diseases-conditions/type-1-diabetes/diagnosis-treatment/drc-20353017. [Accessed: 01-Dec-2021].
3. “Pima Indians Diabetes Database,” *Kaggle*. [Online]. Available: https://www.kaggle.com/uciml/pima-indians-diabetes-database. [Accessed: 01-Dec-2021].
4. “Early Stage Diabetes Risk Prediction Dataset,” *Kaggle.* [Online]. Available: https://www.kaggle.com/ishandutta/early-stage-diabetes-risk-prediction-dataset. [Accessed: 01-Dec-2021].
5. A. Z. Woldaregay, E. Årsand, T. Botsis, D. Albers, L. Mamykina, G. Hartvigsen, D. of C. Science, and C. A. A. Z. Woldaregay, “Data-driven blood glucose pattern classification and anomalies detection: Machine-learning applications in type 1 diabetes,” *Journal of Medical Internet Research*, 01-May-2019. [Online]. Available: https://www.jmir.org/2019/5/e11030/. [Accessed: 29-Nov-2021].
6. F. A. Khan, K. Zeb, M. Al-Rakham, A. Derhab, and S. A. C. Bukhari, “Detection and Prediction of Diabetes Using Data Mining: A Comprehensive Review,” *IEEE Xplore Full-text PDF:* 12-Feb-2021. [Online]. Available:
7. I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda, “Machine Learning and data mining methods in diabetes research,” *Computational and structural biotechnology journal*, 08-Jan-2017. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5257026/. [Accessed: 29-Nov-2021].
8. M. Marinov, A. S. M. Mosa, I. Yoo, and S. A. Boren, “Data-mining technologies for diabetes: A systematic review,” *Journal of diabetes science and technology*, 01-Nov-2011. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3262726/. [Accessed: 29-Nov-2021].
9. A. Duhamel, M. Nuttens, P. Devos, M. Picavet, and R. Beuscart, “A preprocessing method for improving data mining techniques. application to a large medical diabetes database,” *IOS Press Ebooks*, 10-Jan-2012. [Online]. Available:
10. “sklearn.metrics.precision\_recall\_fscore\_support,” *Scikit Learn.* [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.precision\_recall\_fscore\_support.html. [Accessed: 1-Dec-2021].
11. “Type 1 Diabetes,” *MedlinePlus.* [Online]. Available: https://medlineplus.gov/ency/article/000305.htm. [Accessed: 1-Dec-2021].
12. “Insulin and Body Issues,” *Diabetes.co.uk.* [Online]. Available: https://www.diabetes.co.uk/teenagers/insulin-and-body-issues.html. [Accessed: 1-Dec-2021].
13. “Polyuria and Type 1 Diabetes,” *JDRF.* [Online]. Available: https://www.jdrf.org/t1d-resources/about/symptoms/frequent-urination/.  [Accessed: 1-Dec-2021].
14. “Polydipsia,” *Diabetes.co.uk.* [Online]. Available: https://www.diabetes.co.uk/symptoms/polydipsia.html. [Accessed: 1-Dec-2021].