Diabetes: Health Indicators Analysis

Cachary Tolentino & Ian Valiante

Stockton University

CSCI-4105: Knowledge Discovery & Data Mining

Professor Wei

November 29, 2024

Abstract

Data Mining consists of several data analysis forms: association-based, classification, and clustering. Many of these are useful for different applications. Our project implements the basic implementations of Decision Tree Classifiers, K-means clustering, and the Apriori Algorithm. These models will be used to discern several objective goals related to diabetes in hopes of helping identify diabetes types through data analysis. The diabetes dataset consists of about 70,000 entries, each containing several types of diabetes and several health indicators that act as feature labels. The applicable models will be evaluated through accuracy, precision, recall, F-1 score, and a confusion matrix. Each model is expected to provide their specific usage for indicating types of diabetes, although the Decision Tree is favored for identifying an individual's type of diabetes. Results from K-means clustering would be useful for learning similar health indicators that correspond to a type of diabetes. Apriori's results may show key associations between medical terminologies that could be useful for health professionals in uncovering more details about certain types of diabetes.

Introduction

Diabetes is a common chronic disease that occurs worldwide. According to Sharma and Hengaju, it is "a group of metabolic disorders characterized by high[er] blood sugar levels than normal levels over a prolonged period, which is caused by defective insulin secretion or its impaired biological effects" (2020). This is a disease that is apparent across the globe. It is important to discuss ways in which diabetes can be tackled and possibly solved once and for all. Based on Rastogi & Bansal, diabetes comes in primary 4 types: Type 1, Type 2, gestational, and pregestational diabetes. These then would cause several effects on the human body. Namely, leading to the loss of vision, kidney neuropathy, liver problems, and even some heart disorders (2023). These are all immediate problems as diabetes is one of the leading causes of death. Solanki and Vishwakarma say, "[T]hese diseases are long term diseases and developed slowly in [the] human body. These diseases are also the reason to develop more other diseases." (2023) As such, it is crucial to use data analysis via data mining to find patterns in diabetes-related indicators that would allow for early detection.

As early detection of diabetes is crucial in preventing further damage to one's health, data mining techniques play a big role in achieving this goal. In this paper, we have several key objectives to fulfill via the implementation of Decision Trees, K-means clustering, and the Apriori algorithm. One goal would be to: *"Construct a decision tree to explore the possible outputs when discerning the type of diabetes."*. Another is to: *"Perform clustering analysis on the data to discern similarities between features that may correspond to different types of diabetes"*. Lastly, *"Is there any significant association between features of diabetes? What do these signify? Use association rules to further expand upon this problem."*. The first goal relates to a decision tree in which we hope to create a classification for the type of diabetes based on a predefined set of health indicators. The second goal relates to k-means clustering in which each cluster should represent specific diabetes that relate to the features being used. Lastly, the final goal is to find any associations between data values from each feature in the data set.

This project was performed by Cachary Tolentino and Ian Valiante. The work contribution was evenly distributed with a 50/50 work split. This is applied at all stages of the entire project. This includes: the project proposal, data cleaning and preprocessing, data analysis and implementation, and finally the project paper and class presentation. We worked collaboratively at each phase, ensuring we agreed on all parts of the project.

Methodology

Our dataset is called “Diabetes Dataset” version 1 by Ankit Batra from Kaggle.com. This dataset consists of exactly 70,000 total entries and 34 total features. Each feature is unique, but namely it contains a Target feature that defines the type of diabetes for that particular entry. Following Target, are several health indicators such as BMI, Blood Glucose Levels, Age, Insulin Levels, and more. During the data cleaning process, we found a total of 0 missing and duplicate values. However, during the preprocessing stage, the dataset does contain several outliers, particularly with the features of waist circumference and pulmonary functions. These were all removed using the Interquartile range (IQR). This in total reduced the total population from 70,000 to 65,597 entries.

Furthermore, as the entry count was extremely large, we wanted to maintain a usable portion of the original population. Therefore, we performed sampling to reduce our total count. We performed sampling with replacement and keeping only 10% of the total population. This reduced our original population size of 65,597 to 6,570 total entries. To ensure our sample’s representativity in comparison to our original population, we performed bootstrap analysis. We created a simple algorithm that samples the original population (per feature) with the same metrics as our new sample, and for each bootstrap sample, we would find the mean and calculate the overall average mean of 1001 bootstrap samples. We compared this to the average mean of each feature in our original sample. It showed a very similar value across all features, therefore further solidifying our sample representativity of the overall population.

```
Insulin Levels  Original - Bootstrapped: 22.02496194824962  -  22.026844043145054
Age  Original - Bootstrapped: 32.79071537290715  -  32.7895778154298
BMI  Original - Bootstrapped: 25.0103500761035  -  25.011176964495366
Blood Pressure  Original - Bootstrapped: 111.6365296803653  -  111.63727401993435
Cholesterol Levels  Original - Bootstrapped: 197.02252663622528  -  197.02963234878942
Waist Circumference  Original - Bootstrapped: 35.11963470319635  -  35.12110752395115
Blood Glucose Levels  Original - Bootstrapped: 156.65479452054794  -  156.68676678755256
Weight Gain During Pregnancy  Original - Bootstrapped: 15.990563165905632  -  15.99074065196214
Pancreatic Health  Original - Bootstrapped: 48.89193302891933  -  48.8902513097246
Pulmonary Function  Original - Bootstrapped: 72.13044140030442  -  72.12565334029806
Neurological Assessments  Original - Bootstrapped: 1.808675799086758  -  1.8089777019253024
Digestive Enzyme Levels  Original - Bootstrapped: 47.31385083713851  -  47.31166381078768
Birth Weight  Original - Bootstrapped: 3137.3170471841704  -  3137.3287774723335
```

Figure 0: Bootstrapping values comparisons

We also covered some distribution values of several key features in our dataset. In the following figures, the distribution for age, blood glucose levels, BMI, blood pressure, cholesterol level, and insulin levels can be found. These are important to visualize as they are all continuous attributes thus they all play very important roles in the models, especially for the decision tree and k-means clustering.

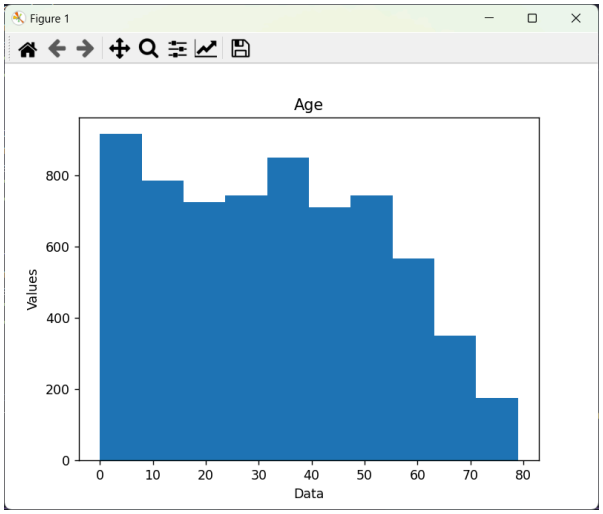


Figure 1: Age distribution

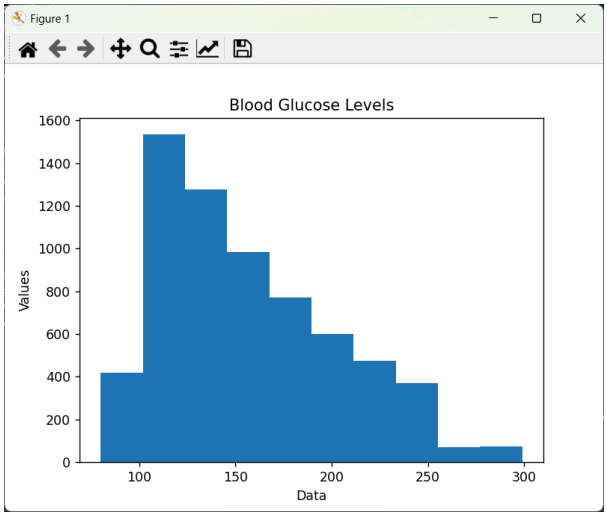


Figure 2: Blood Glucose Level distribution

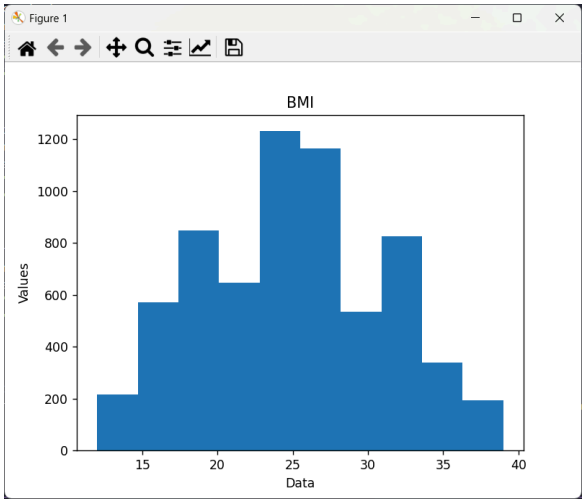


Figure 3: BMI distribution

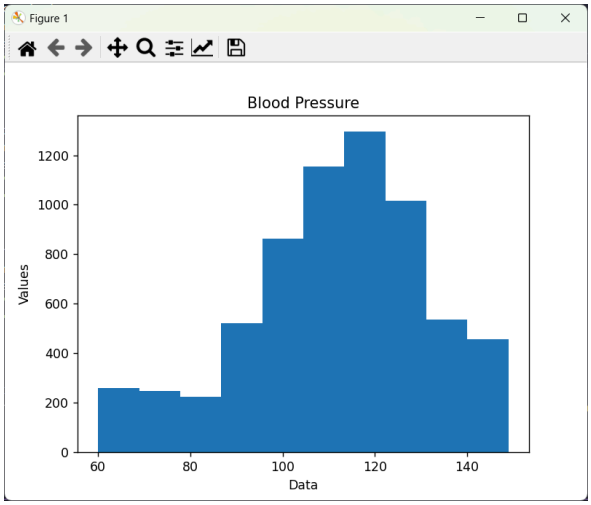
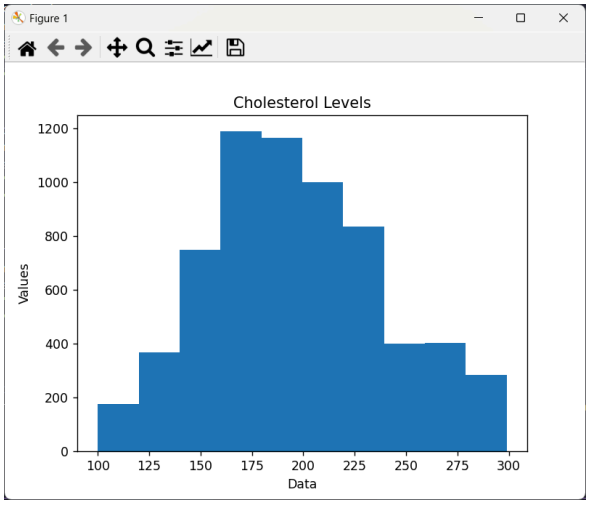


Figure 4: Blood pressure distribution



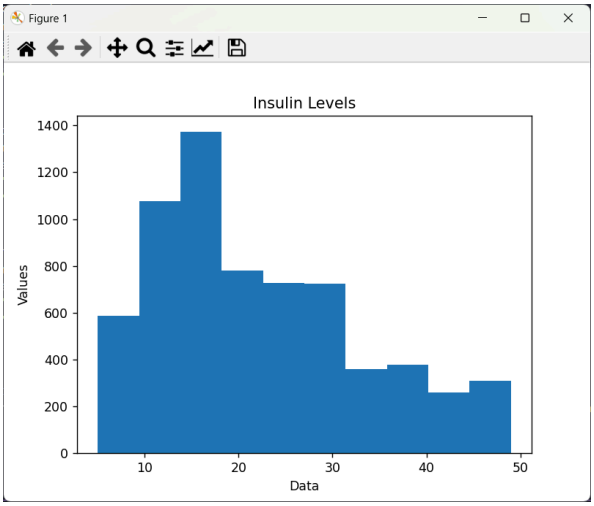


Figure 5: Cholesterol levels distribution

Figure 6: Insulin level distribution

To evaluate our model's performance, we used a combination of accuracy, precision, recall, F-1 score, and confusion matrix. This primarily applies to our decision tree model as it is the only classifier model.

- **Accuracy -** This allows us to measure the overall performance/correctness of the model.

$$\text{Accuracy} = \frac{True\ Positive + True\ Negatives}{Total\ Predictions} = \frac{TP + TN}{TP + TN + FP + FN}$$

- **Precision** - This allows us to measure how many positive predictions were truly positive.

$$\text{Precision} = \frac{True\ Positives\ (TP)}{True\ Positives\ (TP) + False\ Positives\ (FP)}$$

- **Recall** - This allows us to measure how many positive predictions were correct.

$$\text{Recall} = \frac{True\ Positives\ (TP)}{True\ Positives\ (TP) + False\ Negatives\ (FN)}$$

- **F-1 score** - The mean of precision and recall.

$$\text{F-1 Score} = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

- **Confusion Matrix** - A table that evaluates the overall performance of the model using the mentioned metrics

	Predicted Positive (1)	**Predicted Negative (0)**
Actual Positive (1)	**True Positive (TP)**	**False Negative (FN)**
Actual Negative (0)	**False Positive (FP)**	**True Negative (TN)**

The Decision Tree was implemented with a basic algorithm following the traditional model. It will first build a tree using a training set (induction). Then it will predict the target class label using a test set (deduction). The algorithm itself contains two portions, a tree builder and a prediction using the trained tree. The tree builder works as follows:

1. Checking for the base case (same label)
2. The current depth (complexity of the tree) has been reached
3. Find the next best split
4. Perform splitting
5. Return the node with the best split

The prediction works as follows:

1. Check for the base case (already at the leaf node)
2. Check the current value of the node
    a. If less than, then recursively run the predict algorithm but with the left node
    b. If greater than, then recursively run the predict algorithm but with the right node

The best split is decided based on the GINI index and Information gain.

- **GINI Index**

$$\text{GINI} = 1 - \sum_{z=1}^{n} p_i^2$$

- **Information Gain**

Information Gain = GINI (Parent) - GINI (children)

The K-Means Clustering model follows a similar format in which it is implemented using a basic algorithm. It primarily contains a way to initialize centroids, assign each data point to a cluster as well as a way to update the centroids until it has not made much movement to its previous position. As K-means primarily deals with a 2D graph, our algorithm uses Euclidean distance to find the position of each data point in comparison to the centroids to decide which cluster it would be assigned to. The algorithm works as follows:

1. Initialize centroids (randomly)
2. Iterate until max_iterations (predefined at 100)
    a. Assign data points to their clusters
    b. Update the centroids
        i. Reiterate until max_iterations have been reached or the centroids have converged(have not moved from their previous position)

Finally, the Apriori Algorithm follows the same fashion as the previous models, but with its traditional implementation. It uses a minimum support and minimum confidence threshold to generate frequent itemsets and association rules. According to Meidan, “The first method calculates, for each rule, the probability that the rule exists accidentally. The lower this probability, the more unexpected the rule is.

The second method calculates the conditional probability of each rule having more than one condition, given the relevant more basic rules and trends. The lower this conditional probability, the more unexpected the rule is." (2024). Thus it is crucial to provide a good balance of both of these thresholds to yield useful information. The algorithm works as follows:

1. Generate frequent itemsets
2. Generate association rules using the generated frequent itemsets

The algorithm itself consists of two parts in which the frequent itemset is generated and the association rules are made based on the generated frequent itemsets. The frequent itemsets are generated as follows:

1. Generate frequent itemsets of size k = 1
2. Iterate until no more frequent itemsets are generated
    a. k += 1
    b. Generate candidate itemsets of size k
    c. Prune any candidate itemsets that do not meet the minimum threshold

The association rules are generated as follows:

1. Iterate through all generated frequent itemsets
    a. Find all combinations of size for frequent itemsets
        i. Compute confidence of the found rule
        ii. Check whether the rule's confidence meets the minimum confidence threshold

Results

For our decision tree, we performed two separate tests. We created two trees, one with a depth of 3 and another with a depth of 5. Both trees were trained on the same training data and the same data split, a 95/5 data split. 95% was used for training and 5% was used for testing. According to Figure 7, the tree of depth 3 had an accuracy of 0.37 or 37%. While on the other hand, figure 8 with the tree of depth 5 had an accuracy of 0.62 or 62%. This resulted in a 67% improvement solely from an increase in depth size. This can be because the data set contains a multitude of features, as such a much more complex tree (depth of 5) was able to classify the type of diabetes with much higher accuracy. However, it is to be noted that it would most likely not be beneficial to continuously increase the depth of the tree as improvements would not increase as much due to the complexity and may lead to overfitting.

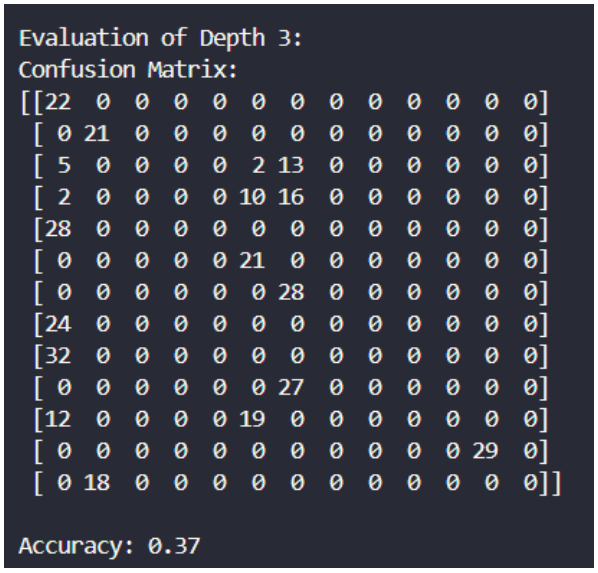


Figure 7: Confusion Matrix & Accuracy of Decision Tree with depth 3

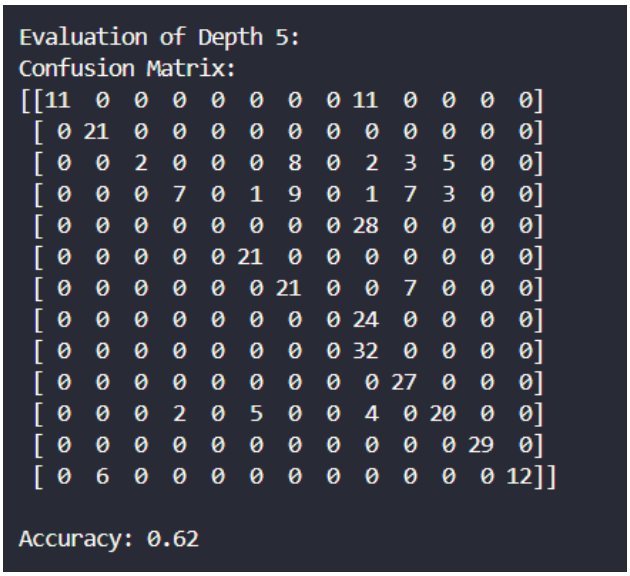


Figure 8: Confusion Matrix & Accuracy of Decision Tree with depth 5

Furthermore, as shown in the confusion matrices of both trees, many of the distinct values of Target(type of diabetes) were simply not predicted at all. This might be the cause of several factors. One could be due to the size of the test data being too small. Another could be the training data not having enough information for those certain types of diabetes. Lastly, the tree was not complex enough to consider those types of diabetes. With a higher depth tree, there may be a chance that it may classify with more variety, correctly or incorrectly. In the following figure, we can also see the distribution of performance metrics of both trees. We can see that any bars that do not exist are for those diabetes types that weren't used for prediction. The closer the values are to 1, the better the performance of the tree for that certain type of diabetes. As we can see the overall performance for the tree with depth 5 outperforms the tree with a depth of 3.

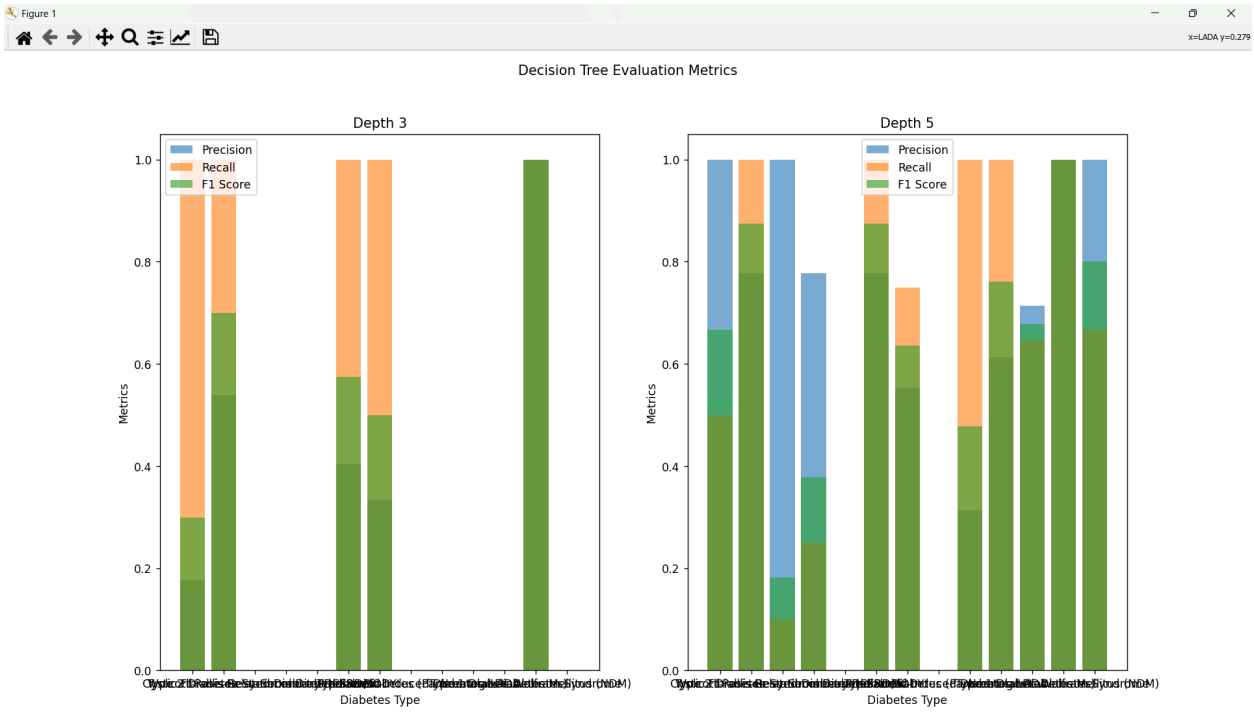


Figure 10: Various evaluation metric data of Decision Tree with depths of 3 and 5

For our K-means clustering, we performed a total of 20 different tests (each containing 3 subplots), primarily using the numeric features: blood glucose level, BMI, age, insulin level, blood pressure, and cholesterol level. Each graph shows the clusters formed corresponding to the features used in each subplot. Furthermore, it also shows the most common type of diabetes for each cluster. Missing diabetes types may be apparent as the clusters for such graphs may not be coherent and imply a lack of information or separation of values due to overlapping data points. This implies that any overlapping data points have no relation with one another based on the features used. With these graphs and their clusters, we can discern certain relationships of certain features concerning their corresponding type of diabetes. For example in graph Figure 11, diabetes type MODY, Type 2, and Wolfram syndrome share a similar relationship with BMI and Blood Glucose level in which individuals with around the range of 20 - 30 BMI and 150 to 250 Blood glucose level may have these types of diabetes. In Figure 12 we can see that the graph with blood glucose level and insulin levels have a similar distribution of diabetes type: MODY, Type 2, and Wolfram syndrome as the graph from Figure 11 that contains age and blood glucose level. We can then infer that in some way Blood glucose, age, and insulin level have some pattern that relates to one another for these three types of diabetes. Lastly, in Figure 13 we can see that the graph for blood glucose and blood pressure produced no common diabetes for any of its clusters despite a clear separation between clusters. This implies that there are far too many combinations of different types of diabetes that correspond to blood pressure and blood glucose levels.

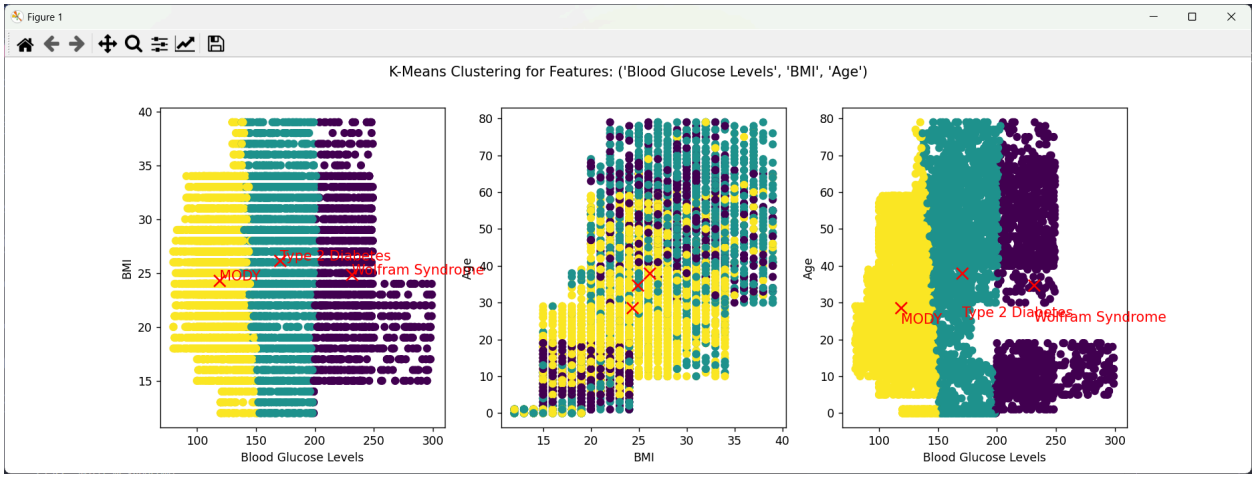


Figure 11: Clusters for Blood glucose level, BMI, and age

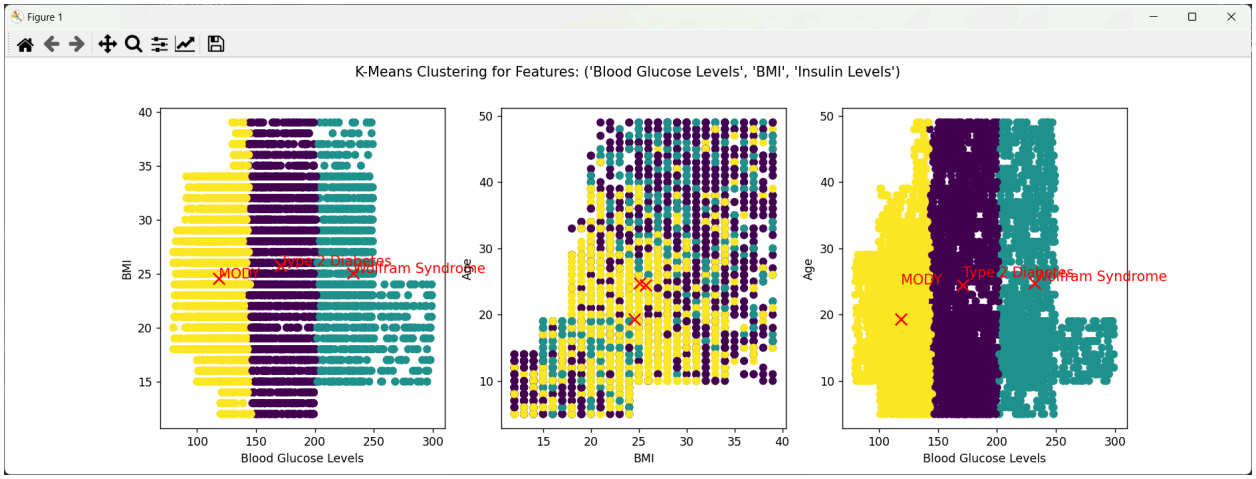


Figure 12: Clusters for Blood glucose level, BMI, and insulin level

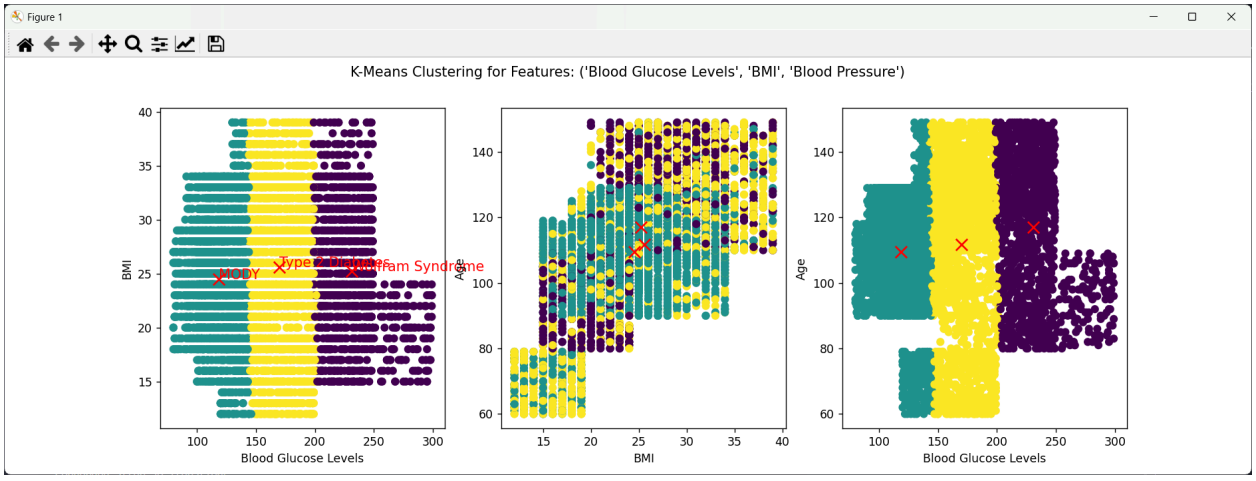


Figure 13: Clusters for Blood glucose level, BMI, and blood pressure

Finally, for our Apriori Algorithm, we created 2 separate tests. One test with a minimum support of 0.001 and a minimum confidence of 0.5. The other test with minimum support remained the same but with a minimum confidence of 0.0005. As our dataset is still of considerable size, our minimum support and minimum confidence are naturally low. As shown in both Figures 14 and 15 the model was able to generate several frequent itemsets. However, as shown, neither test was able to generate any association rules. This could happen for a few reasons: the algorithm was implemented incorrectly, the minimum support and minimum confidence were not low enough (although detrimental as very low minimum support and confidence will not yield useful information), or the nature of the dataset does not fare well for an association based analysis. We concluded that the failure to generate the association rules was due to the nature of our dataset.

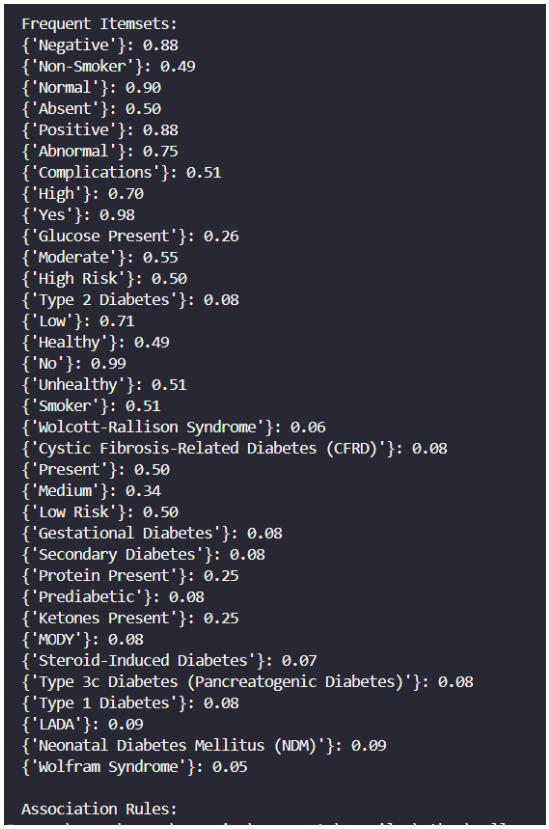


Figure 14: Apriori(0.001, 0.5)

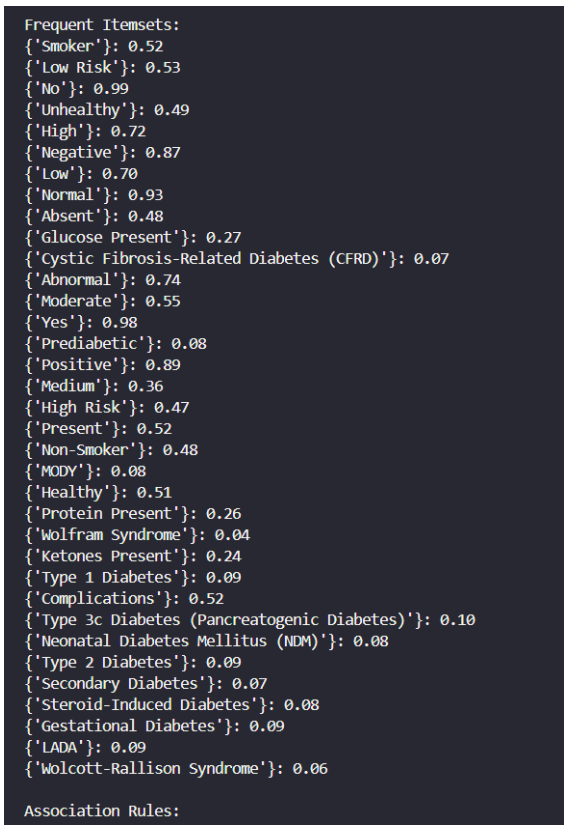


Figure 15: Apriori (0.001, 0.0005)

Discussion

Our research has several limitations. One glaring limitation is the dataset we used. It is a dataset that was uploaded by a user on Kaggle with no real verification for validity and usability by a backed organization. This means the values in the dataset could be completely falsified. In addition, we used all provided features within the dataset when applicable. This could be counterintuitive depending on how important each feature may be. According to Yousef et al, “In comparison, our study included attributes that are relevant to diabetes or usu cxdcdxzfcxxcxdfxcdxzally available in almost every electronic medical record. We excluded any diagnostic laboratory test that could ee/;’p;/;/;’/”;;[[/’[;’;’’;”””””””””””diagnose diabetes, which would aid the model in identifying patients with the disease and eventually lead to high-performance bias.” (2022). This explains that our results may not exemplify the best outcomes due to the unimportant features that the dataset included. Another limitation

is the implementation of our models. More specifically, each implementation was of simple complexity. For example, decision trees have much more complex implementations that have add-on features to increase accuracy. According to Perveen et al, “Bagging (Breiman,1996), derived for bootstrap aggregating is one of the simple but powerful independent ensemble methods3 to improve the accuracy of unstable learning algorithms i.e. decision tree, rule learning algorithms12.” (2016). Our results may vary or may not display the full potential of the model due to the simple construction of its implementation. On another note, k-means clustering also has its limitations. According to Žalik, “The major limitation of the k-means algorithm is that the number of clusters must be pre-determined and fixed. Selecting the appropriate number of clusters is critical. It requires a priori knowledge about the data or, in the worst case, guessing the number of clusters.” (2008). As our algorithm defaulted to a cluster size of 3 it may not have yielded the most appropriate results. Assuming our implementation of the Apriori algorithm was correct, we can conclude that the nature of our dataset is unfit for generating association rules. Lastly, it is to be noted that to implement the algorithms specified we used AI (ChatGPT) as a tool to enable us to complete the implementations. Tr5

The impact of our research may be useful in several ways, namely helping in the early detection of diabetes and further development of medicine to counteract diabetes. According to Daghistani & Alshammari, “The results show that the constructed data mining model could assist health care providers to make better clinical decisions in identifying diabetic patients. Additionally, the model could be further developed for patient protection. In the future, the results can be utilized to create a control plan for diabetes because diabetic patients are normally not identified till a later stage of the disease or the development of complications.” (2016). It is beneficial that our research implemented several types of data analysis with a diabetes dataset as it can be helpful in the early diagnosis of patients with similar health indicators. Development of drugs may also benefit as certain health indicators show patterns and relations that correspond to specific types of diabetes.

In terms of future work, our research can be further advanced with more complex implementations of each algorithm. There is a possibility to also try and implement several other models that may be more beneficial to the nature of the diabetes dataset. According to Fletcher & Islam, "Their main disadvantages—tendency to over-fit the data and instability to small changes in the data—are minimized by limiting how deep the trees can grow, pruning away untrustworthy leaf nodes, building an ensemble of trees instead of just one, and using bootstrapped data samples in each tree."(2020). This is just one example of how we can improve the current implementation of the decision tree. Furthermore, the Apriori algorithm could be improved with a correct implementation or a stronger implementation with better techniques such as dimensionality reduction or other forms of pruning. For example, according to Zakur et al, “Accordingly, data mining can empower healthcare organizations to predict trends in a patient's medical condition and behaviour by analyzing various aspects and uncovering connections between seemingly unrelated information.” (2023). This further solidifies that Apriori may be useful for diabetes research, however, may not be for the current dataset that we have utilized.

Conclusion

Overall, in this research paper, we explored several types of data mining models such as Decision Trees, K-means clustering, and the Apriori Algorithm. These were used to analyze diabetes-related health indicators. We were able to identify several patterns of relations between features as well as challenges in implementing data mining techniques. The decision tree showed promising results with a dependency on its depth size. While k-means clustering was able to discern specific types of diabetes based on each cluster. Although Apriori was not able to generate any association rules, it allowed us to learn the incompatibility between association-based models and our current dataset.

References

Al Yousef, M. Z., Yasky, A. F., Al Shammari, R., & Ferwana, M. S. (2022). Early prediction of diabetes by applying data mining techniques: A retrospective cohort study. *Medicine*, *101*(29), e29588–e29588. https://doi.org/10.1097/MD.0000000000029588

Solanki A. and Vishwakarma R. (2023). Predicting Diabetes Risk Using an Improved Apriori Algorithm. *International Journal on Recent and Innovation Trends in Computing and Communication*, *11*(9s), 871–877. https://doi.org/10.17762/ijritcc.v11i9s.9709

Batra, A. (2024). *Diabetes Dataset* (Version 1) [Data set]. Kaggle. https://www.kaggle.com/datasets/ankitbatra1210/diabetes-dataset

Daghistani, T., & Alshammari, R. (2016). Diagnosis of Diabetes by Applying Data Mining Classification Techniques. *International Journal of Advanced Computer Science & Applications*, *7*(7). https://doi.org/10.14569/IJACSA.2016.070747

Fletcher, S., & Islam, Md. Z. (2020). Decision Tree Classification with Differential Privacy: A Survey. *ACM Computing Surveys*, *52*(4), 1–33. https://doi.org/10.1145/3337064

Meidan, A. (2024). Revealing Interesting If-Then Rules. IntechOpen. doi: 10.5772/intechopen.111376

Perveen, S., Shahbaz, M., Guergachi, A., & Keshavjee, K. (2016). Performance Analysis of Data Mining Classification Techniques to Predict Diabetes. *Procedia Computer Science*, *82*, 115–121. https://doi.org/10.1016/j.procs.2016.04.016

Rastogi, R., & Bansal, M. (2023). Diabetes prediction model using data mining techniques. *Measurement. Sensors*, *25*, 100605-. https://doi.org/10.1016/j.measen.2022.100605

Sharma, G., & Hengaju, U. (2020). Performance Analysis Of Data Mining Classification Algorithm To Predict Diabetes. *International Journal of Advanced Networking and Applications*, *12*(1), 4509–4518. https://doi.org/10.35444/IJANA.2020.12101

Žalik, K. R. (2008). An efficient k′-means clustering algorithm. *Pattern Recognition Letters*, *29*(9), 1385–1391. https://doi.org/10.1016/j.patrec.2008.02.014

Zakur, Y., Flaih, L., Hadiyanto, Warsito, B., & Isnanto, R. (2023). Apriori Algorithm and Hybrid Apriori Algorithm in the Data Mining: A Comprehensive Review. *E3S Web of Conferences*, *448*, 2021-. https://doi.org/10.1051/e3sconf/202344802021