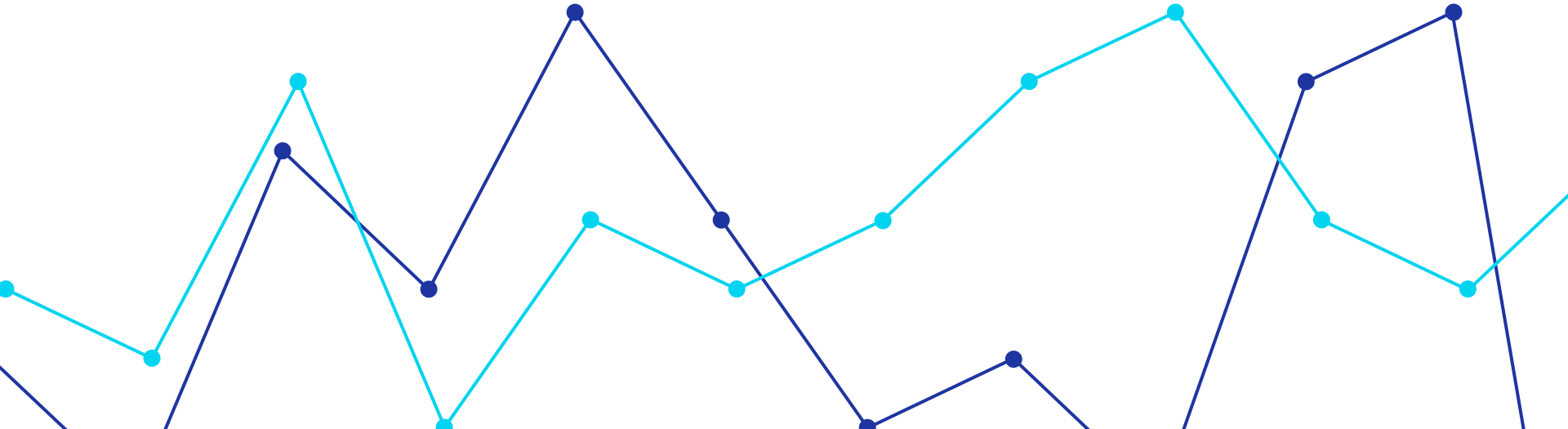


# Diabetes: Health Indicator Analysis

Group #6  
Cachary Tolentino and Ian Valiante





## Background



Diabetes is a chronic diseases that is rampant throughout the world, but especially in America. This illness constitutes the inability of one's body to produce enough insulin or use insulin properly, leading to high blood sugar levels. Many factors come into play when it comes to diabetes. Some are hereditary within the family, such as age, sex, and food consumption. As such, diabetes will be our primary topic, and we will discover information via data mining.

The dataset we will be using is “Diabetes Dataset” provided by Ankit Batra from Kaggle under a CC0: Public domain license (no copyright).

The dataset contains 65,697 entries and 34 total features.



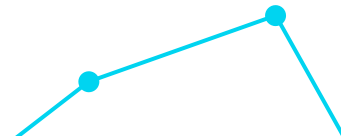
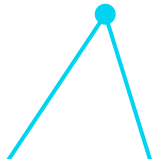


## Research Questions/Project Objectives



### Original → Revised

- **Construct a decision tree to explore the possible outputs when discerning diabetes positivity**
- Perform clustering analysis on the data to observe any primary patterns that imply similarity between forms of diabetes and their symptoms → **Perform clustering analysis on the data to discern similarities between features that may correspond to different types of diabetes**
- **Is there any significant association between features of diabetes? What do these signify? Use association rules to further expand upon this problem**
- ~~Based on a predefined set of symptoms, lifestyle choices, and hereditary traits, predict whether an individual may or may not have diabetes.~~





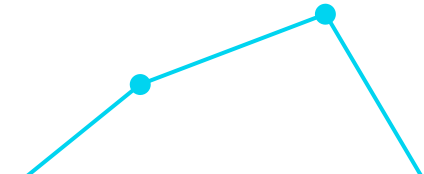
## Timeline/Division of Labor



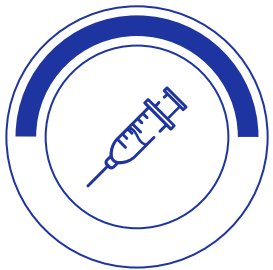
### Proposed Timeline:

- Phase 0 (Data & Proposal): 1 - 2 Day(s)
  - Phase 1 (Data Preprocessing): 1 - 2 week(s)
  - Phase 2 (Data Mining): 3 - 4 weeks
  - Phase 3 (Post Processing): 4 - 5 weeks
  - Phase 4 (Presentation & Research Paper): 1 week
- 

### Actual Timeline:

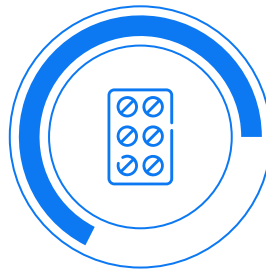
- Phase 1 (Data & Proposal): 1 day
  - Phase 2 (Data Preprocessing): 1 week
  - Phase 3 (Data Analysis & Implementation): 5 - 7 weeks
  - Phase 4 (Project Paper & Class Presentation): 1 week
- 

## Methodology



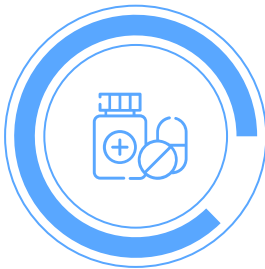
### Decision Tree

A decision tree was used to predict the type of diabetes of an entry (set of health biometrics).



### K-Means Clustering

K-Means clustering was used to analyze any similarities between features and what diabetes they may indicate.



### Apriori Algorithm

Apriori algorithm was used to find any association between several biometrics to see any similarities between diabetes types.

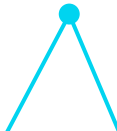
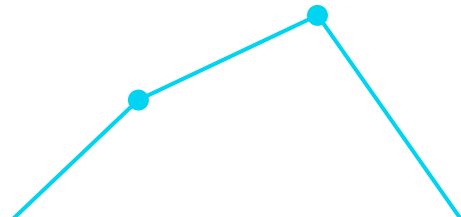


## Methodology



### Evaluation Metrics

To evaluate our algorithms, primarily our Decision Tree (as it is the only classifier algorithm), we used...

- Confusion Matrix
  - Accuracy
  - Precision
  - Recall
  - F-1 Score
- 
- 



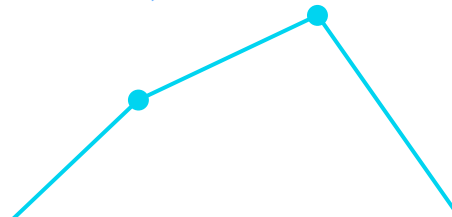
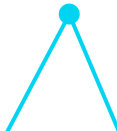
## Methodology



### Decision Tree

It's a basic implementation of a Decision Tree Classifier. It uses GINI Index as its splitting criteria. It also uses Information Gain to determine the effectiveness of each splitting criteria via GINI Index.

- How it works:
  1. Choose your training data and test data (ours was a 95/5 split)
  2. Create a target\_mapping(the set of diabetes types - the class label to be predicted)
  3. Train the Tree using the training data (induction)
  4. Make predictions with the trained tree (deduction)






## Methodology



### K-Means Clustering

Our algorithm is also a basic implementation of K-Means Clustering. It follows the general structure of how the algorithm would work.

- How it works:
    1. Initialize the centroids (done randomly using `np.random.choice` by making a random sample from the given data sample)
    2. Then clusters will be formed via `np.linalg.norm` which calculates the distance of the vector (in our case the distance of actual data points from each centroid). It will then assign the data points to the closest centroid.
    3. After each data point have been assigned to a cluster, the centroids will be updated via the mean value of the cluster it is within.
    4. Steps 2 and 3 will repeat until the centroids have converged (have not moved much from its previous position).
- 



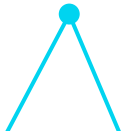
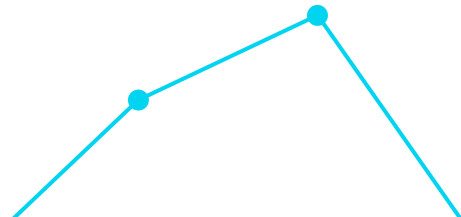


## Methodology



### Apriori Algorithm

Similarly to the other algorithms, our implementation of the Apriori Algorithm follows the traditional means in which it generates frequent itemsets and rules based on the discovered frequent itemsets.

- How it works:
    1. Generates frequent itemsets based on a given minimum support threshold.
    2. Rules will be generated using the frequent itemsets generates based on a given minimum confidence threshold.
    3. Each frequent itemset and rules will be printed out.
- 
- 



# Implementation Demonstration

(\*Open VSC)







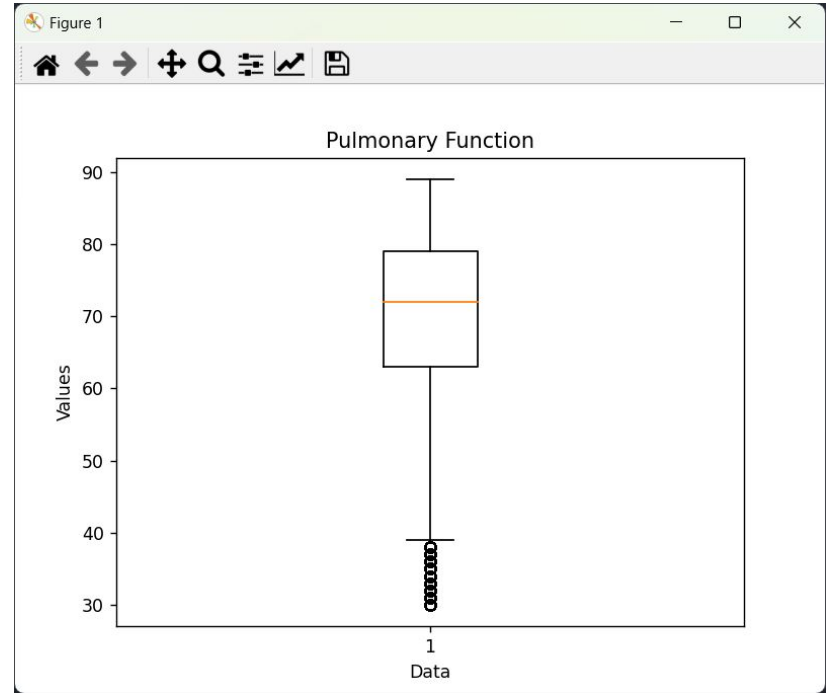
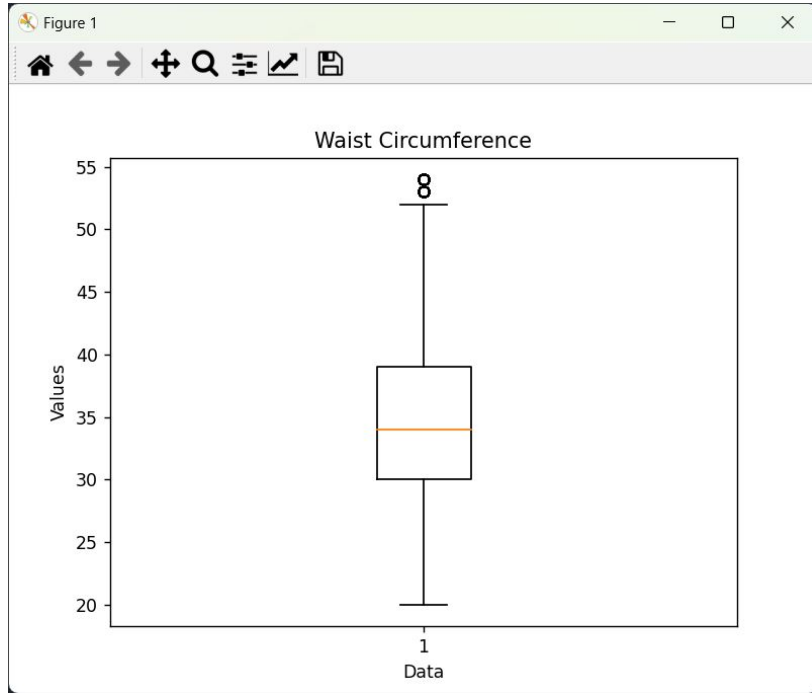
## Technical Challenges



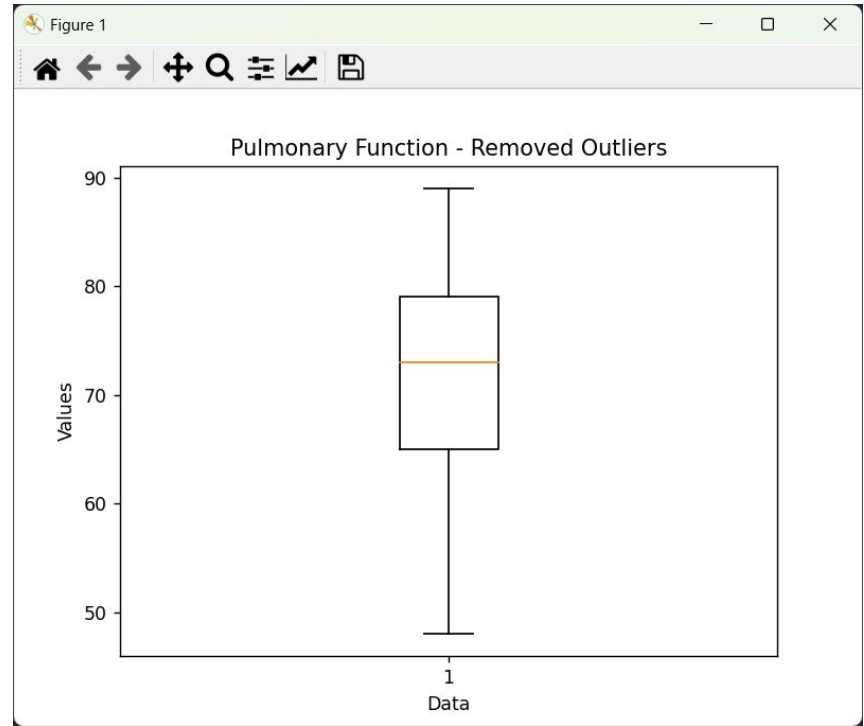
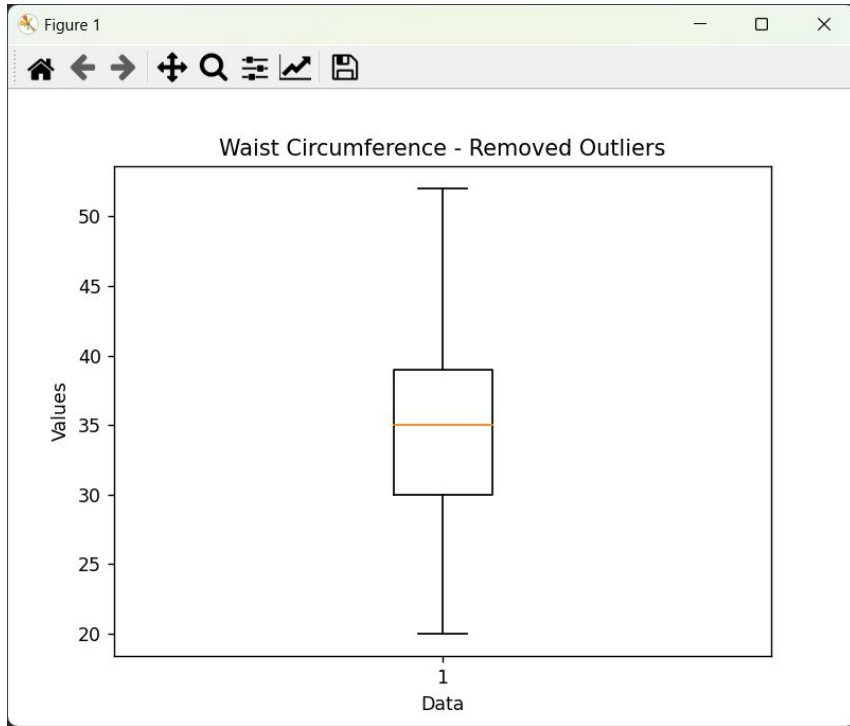
### Data Cleaning

- Outliers
    - Our dataset had a couple of outliers. Primarily for the feature waist circumference and pulmonary function. We removed these outliers using Interquartile Range (IQR) technique. Essentially, we set a bound of acceptable values for our dataset specifically for these two features.
- 
- 

## Technical Challenges



## Technical Challenges

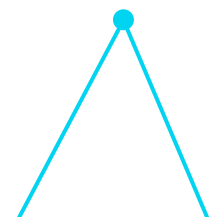
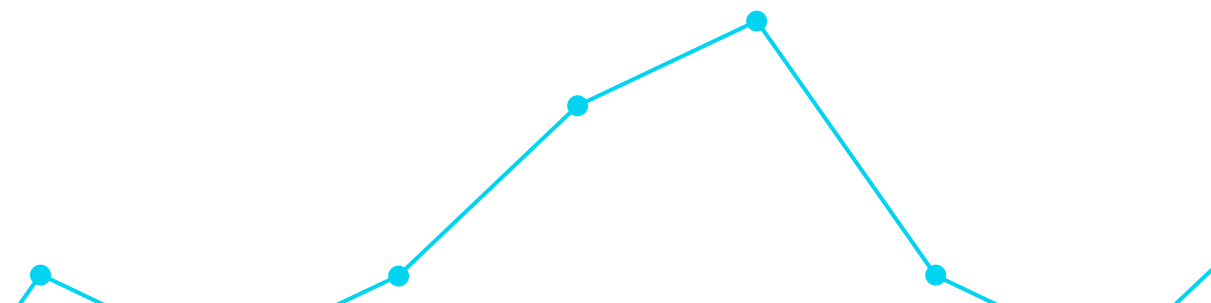




## Technical Challenges



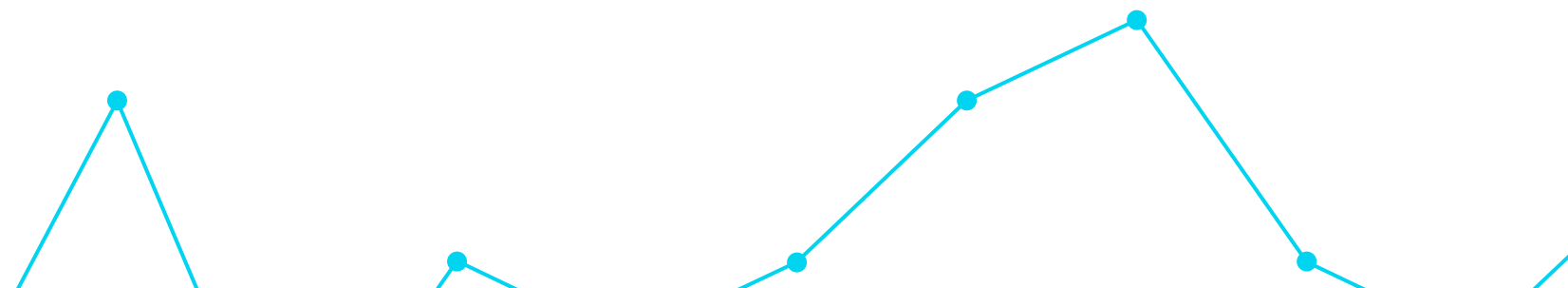
### Data Preprocessing

- Sampling
    - With a large dataset (+65,000 entries), we wanted to reduce the data sample. We used sampling with replacement.
    - Our reduced sample size was about 10% of the original size (~6,500 entries).
    - But we wanted to ensure the representativity (how evenly distributed each entry is) within our new found sample.
    - Therefore, we used Bootstrapping! (Only works for numeric values)
- 
- 



## Technical Challenges

### Data Preprocessing

- Bootstrapping Explanation
    - Essentially, we found the mean for each feature.
    - Then for the bootstrapping portion, we created an algorithm that finds the average of means for about 1001 values for each numeric feature.
    - This is useful for us to show that the means for our original sample is close to the bootstrapped means, therefore, the sample is representative of the entire population.
- 

## Technical Challenges

Number of duplicated values: 0  
Original population size: 65697  
New population size: 6570


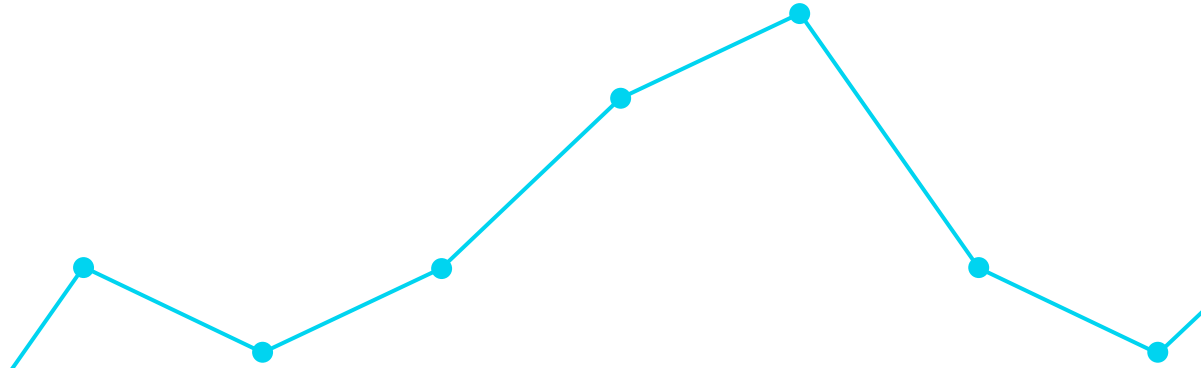
Insulin Levels Original - Bootstrapped: 22.02496194824962 - 22.026844043145054  
Age Original - Bootstrapped: 32.79071537290715 - 32.7895778154298  
BMI Original - Bootstrapped: 25.0103500761035 - 25.011176964495366  
Blood Pressure Original - Bootstrapped: 111.6365296803653 - 111.63727401993435  
Cholesterol 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





## Highlights

### Apriori Results:

- We made two tests with our apriori algorithm
    - One with a min\_sup of 0.001 and min\_conf of 0.5
    - Another with a min\_sup of 0.001 and min\_conf of 0.0005
  - Both yielded some frequent itemsets, but no association rules.
  - We concluded that these results are natural as our dataset, specifically each feature values, have no association with each other.
- 
- 

## Highlights

```
Frequent Itemsets:
{'Negative'}: 0.88
{'Non-Smoker'}: 0.49
{'Normal'}: 0.90
{'Absent'}: 0.50
{'Positive'}: 0.88
{'Abnormal'}: 0.75
{'Complications'}: 0.51
{'High'}: 0.70
{'Yes'}: 0.98
{'Glucose Present'}: 0.26
{'Moderate'}: 0.55
{'High Risk'}: 0.50
{'Type 2 Diabetes'}: 0.08
{'Low'}: 0.71
{'Healthy'}: 0.49
{'No'}: 0.99
{'Unhealthy'}: 0.51
{'Smoker'}: 0.51
{'Wolcott-Rallison Syndrome'}: 0.06
{'Cystic Fibrosis-Related Diabetes (CFRD)'}: 0.08
{'Present'}: 0.50
{'Medium'}: 0.34
{'Low Risk'}: 0.50
{'Gestational Diabetes'}: 0.08
{'Secondary Diabetes'}: 0.08
{'Protein Present'}: 0.25
{'Prediabetic'}: 0.08
{'Ketones Present'}: 0.25
{'MODY'}: 0.08
{'Steroid-Induced Diabetes'}: 0.07
{'Type 3c Diabetes (Pancreatogenic Diabetes)'}: 0.08
{'Type 1 Diabetes'}: 0.08
{'LADA'}: 0.09
{'Neonatal Diabetes Mellitus (NDM)'}: 0.09
{'Wolfram Syndrome'}: 0.05
```

Association Rules:

← First  
Call

Second  
Call →

```
Frequent Itemsets:
{'Smoker'}: 0.52
{'Low Risk'}: 0.53
{'No'}: 0.99
{'Unhealthy'}: 0.49
{'High'}: 0.72
{'Negative'}: 0.87
{'Low'}: 0.70
{'Normal'}: 0.93
{'Absent'}: 0.48
{'Glucose Present'}: 0.27
{'Cystic Fibrosis-Related Diabetes (CFRD)'}: 0.07
{'Abnormal'}: 0.74
{'Moderate'}: 0.55
{'Yes'}: 0.98
{'Prediabetic'}: 0.08
{'Positive'}: 0.89
{'Medium'}: 0.36
{'High Risk'}: 0.47
{'Present'}: 0.52
{'Non-Smoker'}: 0.48
{'MODY'}: 0.08
{'Healthy'}: 0.51
{'Protein Present'}: 0.26
{'Wolfram Syndrome'}: 0.04
{'Ketones Present'}: 0.24
{'Type 1 Diabetes'}: 0.09
{'Complications'}: 0.52
{'Type 3c Diabetes (Pancreatogenic Diabetes)'}: 0.10
{'Neonatal Diabetes Mellitus (NDM)'}: 0.08
{'Type 2 Diabetes'}: 0.09
{'Secondary Diabetes'}: 0.07
{'Steroid-Induced Diabetes'}: 0.08
{'Gestational Diabetes'}: 0.09
{'LADA'}: 0.09
{'Wolcott-Rallison Syndrome'}: 0.06
```


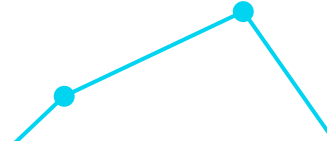
Association Rules:



## Highlights



### Decision Tree Results:

- We created two tests
    - One with a tree max depth of 3
    - Another with a tree max depth of 5
  - The tree with max depth of 3 resulted in an overall 0.37 or 37% accuracy.
  - While the tree with max depth of 5 resulted in an overall 0.62 or 62% accuracy.
  - This is a **67% improvement!**
  - We concluded that this improvement is due to the increase in depth size, primarily due to the large dimensionality of the dataset (13 features), a more complex tree is able to classify the type of diabetes more accurately.
  - We do understand that this is not a linear relationship and would eventually not see much performance gains while increasing tree complexity.
- 
- 

## Highlights

Evaluation of Depth 3:

Confusion Matrix:

```
[[22  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0 21  0  0  0  0  0  0  0  0  0  0  0]
 [ 5  0  0  0  0  2 13  0  0  0  0  0  0]
 [ 2  0  0  0  0 10 16  0  0  0  0  0  0]
 [28  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0 21  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0 28  0  0  0  0  0  0]
 [24  0  0  0  0  0  0  0  0  0  0  0  0]
 [32  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0 27  0  0  0  0  0  0]
 [12  0  0  0  0 19  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  0  0  0  0 29  0]
 [ 0 18  0  0  0  0  0  0  0  0  0  0  0]]
```

Accuracy: 0.37

Evaluation of Depth 5:

Confusion Matrix:

```
[[11  0  0  0  0  0  0  0 11  0  0  0  0]
 [ 0 21  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  2  0  0  0  8  0  2  3  5  0  0]
 [ 0  0  0  7  0  1  9  0  1  7  3  0  0]
 [ 0  0  0  0  0  0  0  0 28  0  0  0  0]
 [ 0  0  0  0  0 21  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0 21  0  0  7  0  0  0]
 [ 0  0  0  0  0  0  0  0 24  0  0  0  0]
 [ 0  0  0  0  0  0  0  0 32  0  0  0  0]
 [ 0  0  0  0  0  0  0  0  0 27  0  0  0]
 [ 0  0  0  2  0  5  0  0  4  0 20  0  0]
 [ 0  0  0  0  0  0  0  0  0  0 29  0  0]
 [ 0  6  0  0  0  0  0  0  0  0  0 12  0]]
```

Accuracy: 0.62

# Highlights

```
Class-wise Metrics:
Class: Type 2 Diabetes
  Precision: 0.18
  Recall: 1.00
  F1-score: 0.30
Class: Wolcott-Rallison Syndrome
  Precision: 0.54
  Recall: 1.00
  F1-score: 0.70
Class: Cystic Fibrosis-Related Diabetes (CFRD)
  Precision: 0.00
  Recall: 0.00
  F1-score: 0.00
Class: Gestational Diabetes
  Precision: 0.00
  Recall: 0.00
  F1-score: 0.00
Class: Secondary Diabetes
  Precision: 0.00
  Recall: 0.00
  F1-score: 0.00
Class: Prediabetic
  Precision: 0.40
  Recall: 1.00
  F1-score: 0.58
Class: MODY
  Precision: 0.33
  Recall: 1.00
  F1-score: 0.50
Class: Steroid-Induced Diabetes
  Precision: 0.00
  Recall: 0.00
  F1-score: 0.00
Class: Type 3c Diabetes (Pancreatogenic Diabetes)
  Precision: 0.00
  Recall: 0.00
  F1-score: 0.00
Class: Type 1 Diabetes
  Precision: 0.00
  Recall: 0.00
  F1-score: 0.00
Class: LADA
  Precision: 0.00
  Recall: 0.00
```

```
Precision: 0.00
Recall: 0.00
F1-score: 0.00
Class: LADA
  Precision: 0.00
  Recall: 0.00
  F1-score: 0.00
Class: Neonatal Diabetes Mellitus (NDM)
  Precision: 1.00
  Recall: 1.00
  F1-score: 1.00
Class: Wolfram Syndrome
  Precision: 0.00
  Recall: 0.00
  F1-score: 0.00
```

## Highlights

```
Class-wise Metrics:
Class: Type 2 Diabetes
  Precision: 1.00
  Recall: 0.50
  F1-score: 0.67
Class: Wolcott-Rallison Syndrome
  Precision: 0.78
  Recall: 1.00
  F1-score: 0.88
Class: Cystic Fibrosis-Related Diabetes (CFRD)
  Precision: 1.00
  Recall: 0.10
  F1-score: 0.18
Class: Gestational Diabetes
  Precision: 0.78
  Recall: 0.25
  F1-score: 0.38
Class: Secondary Diabetes
  Precision: 0.00
  Recall: 0.00
  F1-score: 0.00
Class: Prediabetic
  Precision: 0.78
  Recall: 1.00
  F1-score: 0.88
Class: MODY
  Precision: 0.55
  Recall: 0.75
  F1-score: 0.64
Class: Steroid-Induced Diabetes
  Precision: 0.00
  Recall: 0.00
  F1-score: 0.00
Class: Type 3c Diabetes (Pancreatogenic Diabetes)
  Precision: 0.31
  Recall: 1.00
  F1-score: 0.48
Class: Type 1 Diabetes
  Precision: 0.61
  Recall: 1.00
  F1-score: 0.76
Class: LADA
  Precision: 0.71
  Recall: 0.65
  F1-score: 0.68
```

```
F1-score: 0.76
Class: LADA
  Precision: 0.71
  Recall: 0.65
  F1-score: 0.68
Class: Neonatal Diabetes Mellitus (NDM)
  Precision: 1.00
  Recall: 1.00
  F1-score: 1.00
Class: Wolfram Syndrome
  Precision: 1.00
  Recall: 0.67
  F1-score: 0.80
```


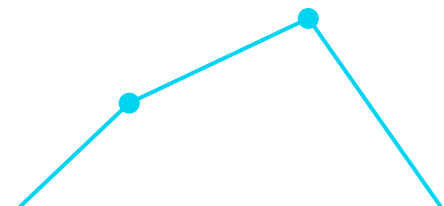




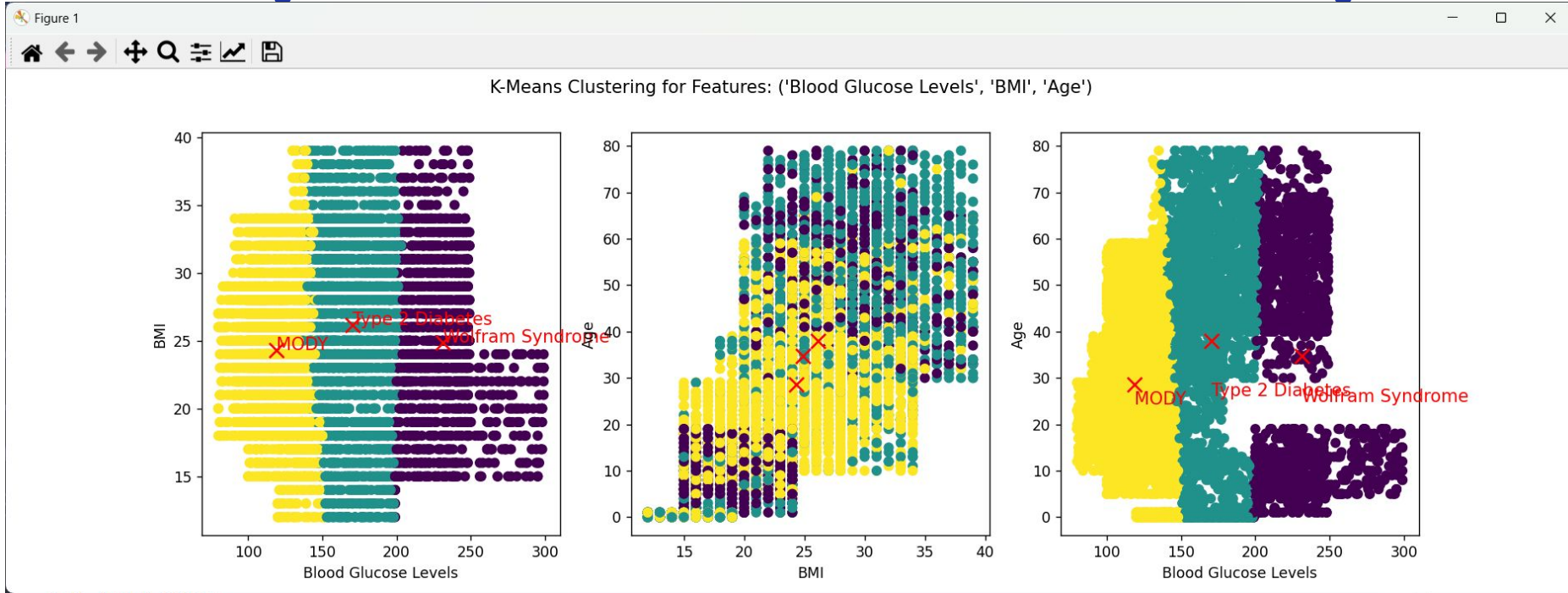
## Highlights



### K-Means Clustering Results:

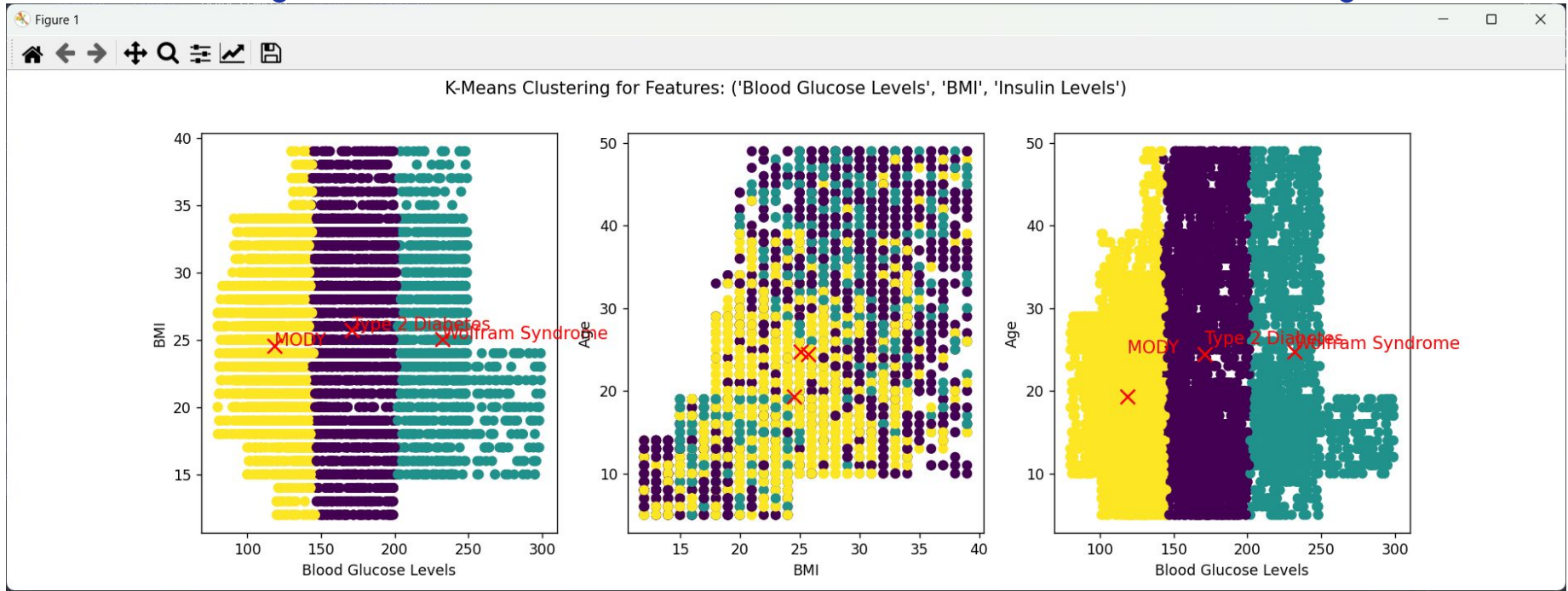
- We created several test, specifically about 20 total combinations for 6 features (Blood Glucose Levels, BMI, Age, Insulin Levels, Blood Pressure, and Cholesterol Levels).
  - Each data point can be treated as a patient with a combination of the indicated features at varying levels. The clusters then represents the similarities of each patient.
  - The graphs also indicates the most common diabetes that appear in each cluster.
  - As we can see in the following graphs there are some interesting relationships between some of these features.
  - This can be used as a guide to discern similar diabetes in the context of biometrics that they may affect or act as a health indicator.
  - \*Note\* The graphs were not able to fully label the correct features used for each axis and diabetes type.
- 
- 

# Highlights

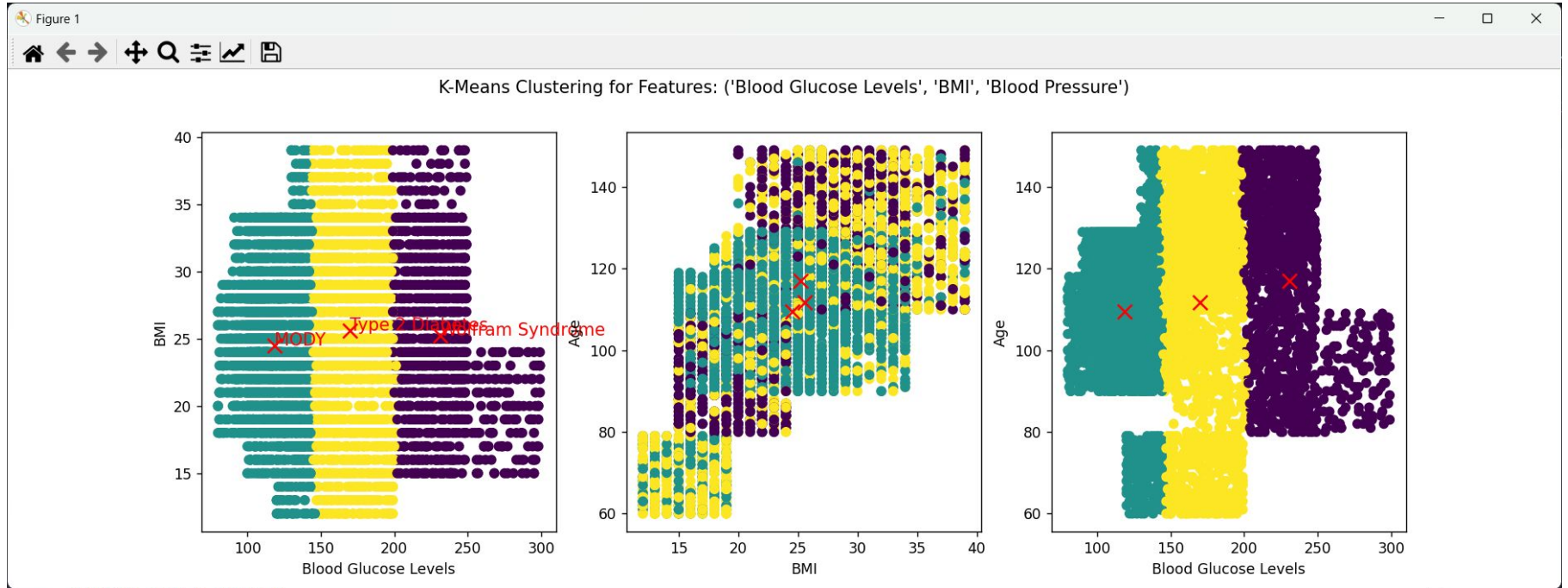




# Highlights



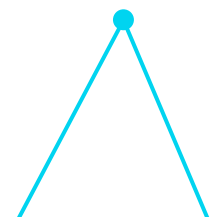
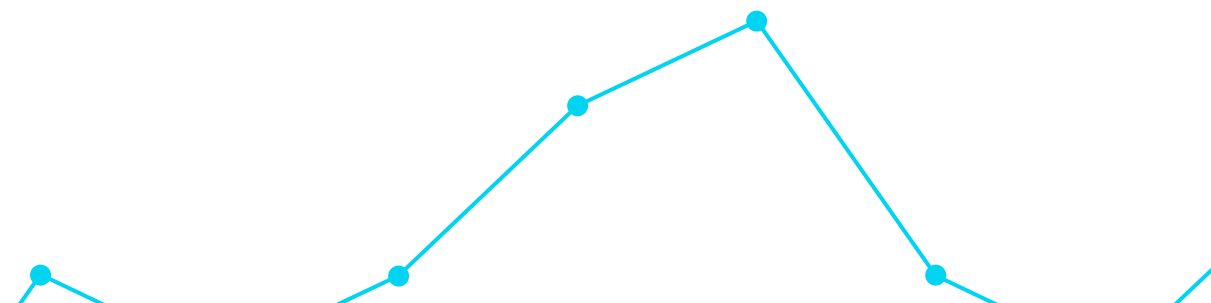
# Highlights





## Conclusion



- The Decision Tree Classifier was useful in classifying the type of diabetes based on defined biometric data. Furthermore, its accuracy can be increased with a more complex tree.
  - The K-Means Clustering was useful for finding similarities in biometric data that corresponds to certain types of diabetes.
  - The Apriori Algorithm was unfortunately not useful in finding any association rules in our given dataset.
  - Overall, each algorithm could possibly be improved with a more complex implementation and further consideration specifically for our dataset.
  - Each of these can be used to help solve issues relating to diabetes identification.
- 
- 



## 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.00000000000029588>

Arpit Solanki, E. al. (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>





## References

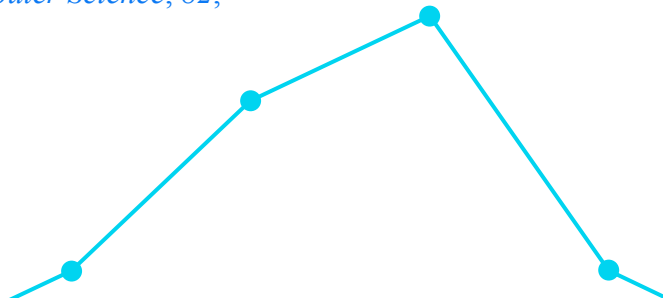

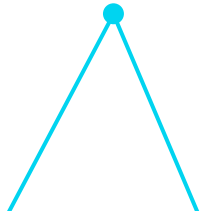


Daghistani, T., & Alshammari, R. (2016). Diagnosis of Diabetes by Applying Data Mining Classification Techniques. *International Journal of Advanced Computer Science & Applications*, 7(7). <https://doi.org/10.14569/IJACSA.2016.070747>

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>





## References



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>

