**MSIS 645 Data Mining and Predictive Analysis**

Project Report

Prediction, Association and Clustering on Diabetes and Its Factors

Swetha Adike

Table of Contents

[**1.** **Abstract** 3](#_Toc102718130)

[**2.** **Introduction** 3](#_Toc102718131)

[**3.** **Data** **Explanation** 3](#_Toc102718132)

[**4.** **Analysis**/**Methodology** 6](#_Toc102718133)

[**4.1.** **Prediction** 8](#_Toc102718134)

[**4.2.** **Association** 12](#_Toc102718135)

[**4.3.** **Clustering** 15](#_Toc102718136)

[**5.** **Conclusions** 17](#_Toc102718137)

[**6.** **References**: 18](#_Toc102718138)

# **Abstract**

Diabetes is a serios chronic disease which is increasing rapidly now a days. This disease is increasing its strength and moving upon every individual who is not disciplined to one’s health habits, no leaving even the children. In this project, an analysis of this disease is been made by imposing few data mining tasks on a chosen dataset. This dataset is collection of responses from the individuals on the status of diabetes based on few indices like BMI, cholesterol, age, etc. This project work is an attempt in which relevant data mining tasks such as prediction, association and clustering are imposed on the dataset to evaluate the presence of diabetes, association rules between the factors and clustering into few groups with similar data attributes. The resultant data is used for further analysis on evaluating and deciding which factors are more prevailing, the dependents and the common groups.

# **Introduction**

Diabetes is one of the serious chronic diseases, impacting number of humans every year. The individuals affected by this disease lose the ability to effectively regulate glucose in blood and will have reduced quality of life and life expectancy. Complications like heart disease, kidney disease, amputations are associated with this disease. It is not an exaggeration to say that there is no cure for this disease except following a balanced life to have a disciplined body. Early diagnosis of diabetes can lead to lifestyle change and help in controlling the disease.

This project work is done on this theme to predict the chances of diabetes for a person on 21 different health factors. Based on these health factors, a model is developed using Naïve Bayesian classification, to predict the chances of diabetes for a person based on 21 health factors. As an extension to this, the datapoints are analyzed for association rules and then clustering is formed by choosing particular data types. The data used in this model is a part of dataset taken from a health-related telephonic survey conducted by Behavioral Risk Factor Surveillance System (BRFSS), available on Kaggle for the year 2015. The original dataset has records from 441,455 individuals and has 330 attributes from which 21 were selected and used in this project work.

# **Data** **Explanation**

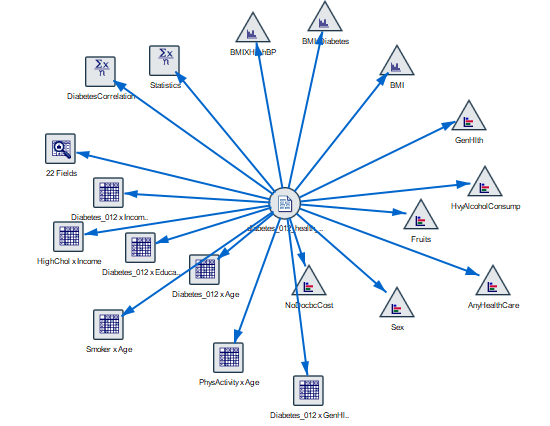
Before getting to deep analysis into the data, a detailed explanation of the data is as below

The reduced dataset used in this work consists of 70,692 responses on the diabetes condition in the individual based on 21 below factors

|  |  |  |
| --- | --- | --- |
| Feature Name | Description | Type |
| Diabetes\_012 | Stage of diabetes in a person – 0 for no diabetes, 1 for diabetes | Discrete |
| HighBP | 0 = no high BP, 1 = high BP | Discrete |
| HighChol | 0 = no high cholesterol, 1 = high cholesterol | Discrete |
| CholCheck | 0 = no cholesterol check in 5 years, 1 = yes cholesterol check in 5 years | Discrete |
| BMI | Body Mass Index | Continuous |
| Smoker | Smoked at least 100 cigarettes in your entire life: 0= no, 1 = yes | Discrete |
| Stroke | (Ever told) had a stroke: 0= no, 1 = yes | Discrete |
| HeartDiseaseorAttack | Coronary heart disease (CHD) or myocardial infarction (MI): 0= no, 1 = yes | Discrete |
| PhysActivity | Physical activity in past 30 days – not including job: 0= no, 1 = yes | Discrete |
| Fruits | Consume fruit one or more times per day: 0= no, 1 = yes | Discrete |
| Veggies | Consume vegetable one or more times per day: 0= no, 1 = yes | Discrete |
| HvyAlcoholConsump | Adult men >=14 drinks per week and adult women >=7 drinks per week: 0 = no, 1 = yes | Discrete |
| AnyHealthcare | Having any health care coverage including health insurance, prepaid plans etc. 0 = no, 1 = yes | Discrete |
| NoDocbcCost | Skipped from past 12 months when needed to see doctor but couldn’t because of cost: 0 =no, 1= yes | Discrete |
| GenHlth | General health on scale 1-5: 1 = excellent, 2 = very good, 3 = good, 4 = fair, 5 = poor | Categorical |
| MentHlth | Days of poor mental health from past 30 days, scaled from 1-30 | Continuous |
| PhyHlth | Physical illness or injury days in past 30 days, scaled from 1-30 | Continuous |
| DiffWalk | Serious difficulty in walking or climbing stairs 0= no, 1= yes | Discrete |
| Sex | 0 = Female, 1= Male | Categorical |
| Age | Age on scale 1- 13: Age 18-24 = 1, Age 25 to 29 = 2, Age 30-34 = 3, Age 35-39 = 4, Age 40-44 = 5, Age 45-49 = 6, Age 50-54 = 7, Age 55-59 = 8, Age 60-64 = 9, Age 65-69 = 10, Age 70-74 = 11, Age 75-79 = 12, Age 79-84 = 13 | Categorical |
| Education | Education on scale 1-6: Never attended school or only kindergarten = 1, Elementary education = 2, High school = 3, High school graduate = 4, College or technical study = 5, College graduate = 6 | Categorical |
| Income | Income interpreted on scale 1-8: less than $10,000 = 1; less than $35,000 = 5; $75,000 or more = 8 | Categorical |

# **Analysis**/**Methodology**

The dataset is analyzed in detail by creating a stream in SPSS modeler as below with data audit node, statistics and correlations and few other graph plots. This stream file is attached with the file name DMProject2Analysis.str

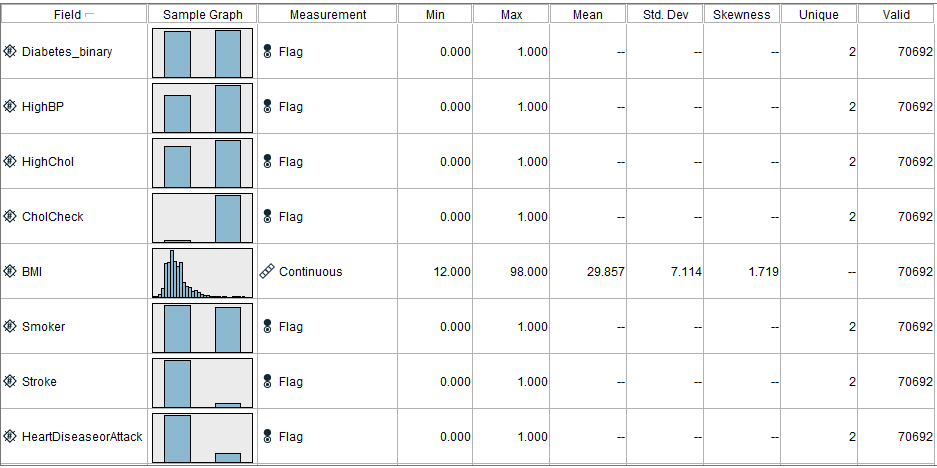


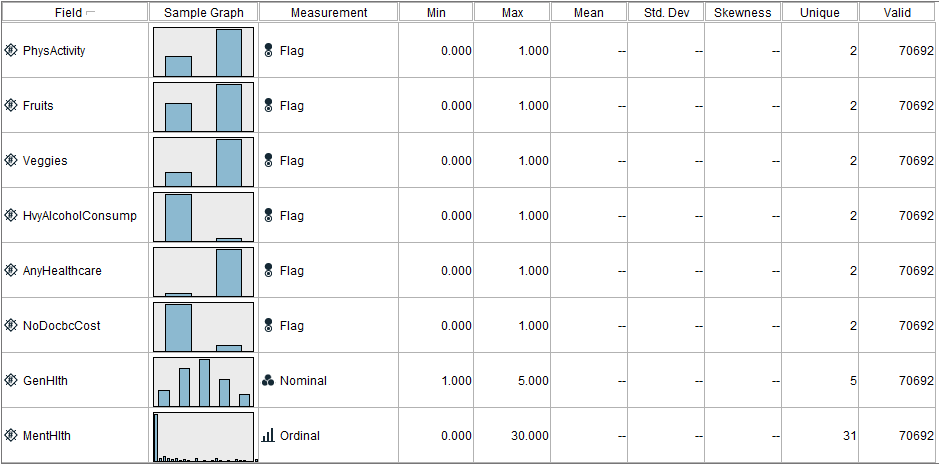
In the stream, while the details of data audit node are discussed, below are the inferences from other nodes

* The statistics node gives all the statistical details like mean, median, standard deviation, etc., and the correlations of 21 fields with diabetes.
* Since age, education, income and general health are categorical values, they are compared with diabetes by matrix node to get categorical diabetes results.
* The distribution graph gives the comparison of NoDocbcCost, Sex, AnyHealthCare, Fruits, HvyAlcoholConsump, GenHlt, normalized by colors.
* Other than comparisons with diabetes, general analysis is also done creating a matrix node from Age and physical activity, Smoking with age, cholesterol variation with income etc.,

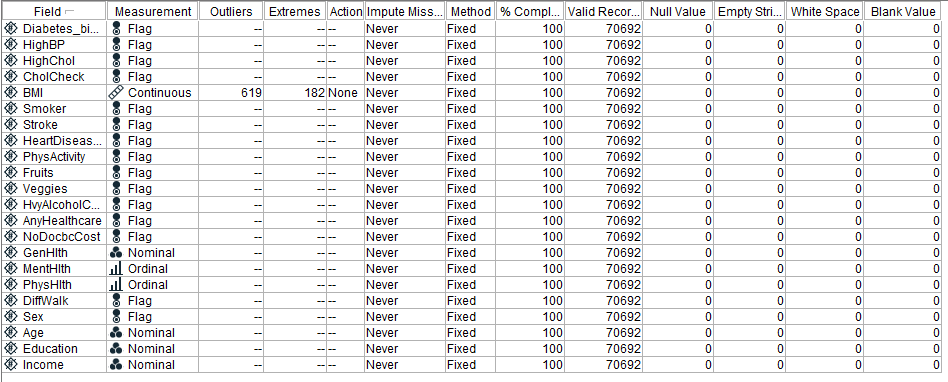
The details of audit node, comparing to above data table with datatypes can be verified with the below figures

* From the above table, the fields are divided into categorical, discrete and continuous values where there are 6 categorical, 15 discrete and 1 continuous attributes, accounting to all 22 fields.
* Since there is no field with all zeros, all are good to be considered for developing the model.
* From the above quality table, there are no missing values and no null or blank values and the considered data set is 100% complete with valid details.
* There are extreme values few outliers in BMI field, other than this all the fields are good for the model consideration





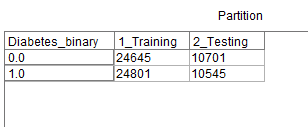




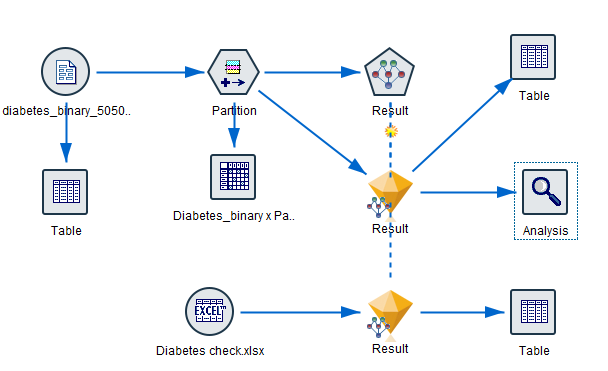
## **Prediction**

Using the above data with discussed data types, the following model is created with 70% of the data partitioned for training and 30% for testing so that there will be an assessment on how well the model is functioning by analyzing the confusion matrix. The model is also attached as stream file named DMProject2Prediction

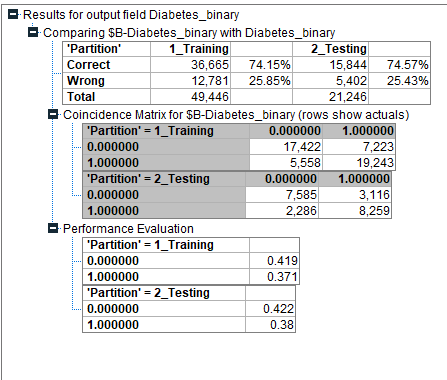
The following matrix is generated for correlating the count between the diabetes data with training and testing data, shows respective records partitioned for training and testing corresponding to diabetes with the responses no diabetes, and diabetes (0, and 1)



In the modeling tab, Bayesian network node is chosen for classification and prediction of diabetes from the given data.



The analysis node attached to the generated diabetes prediction super node gives the following table



From the above figure, the accuracy for testing data is 74.57% which is nearly equal to training data which shows that there is no overfitting in the model.

The accuracy can be considered as an adequate matric as the dataset is balanced.

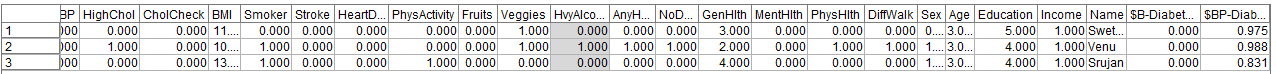
Recall = 8259/(8259+2286) = 73%,

precision = 8259/(8259+3116) = 72% and

specificity = 7585/(7585+3116) = 70%

Hence from the above, 72% of the individuals are predicted with diabetes are actually with diabetes and 1-specificity = approximately 30% are with not with diabetes.

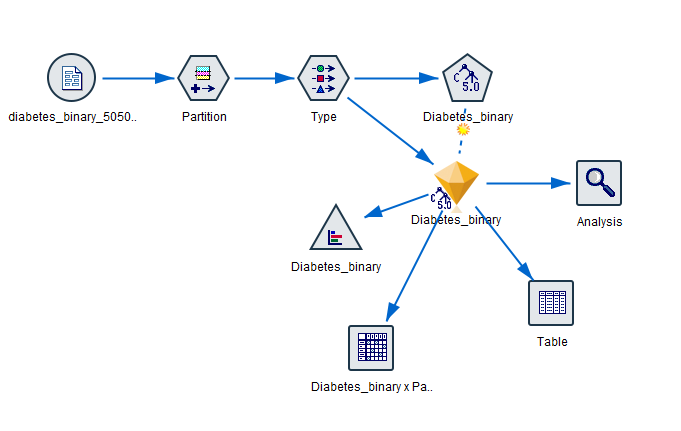
The model has been checked by giving our daily details in an excel sheet, (named as Diabetes check.xlsx) (the details are however exaggerated from the real) to predict whether we are in the diabetic zone and the results are



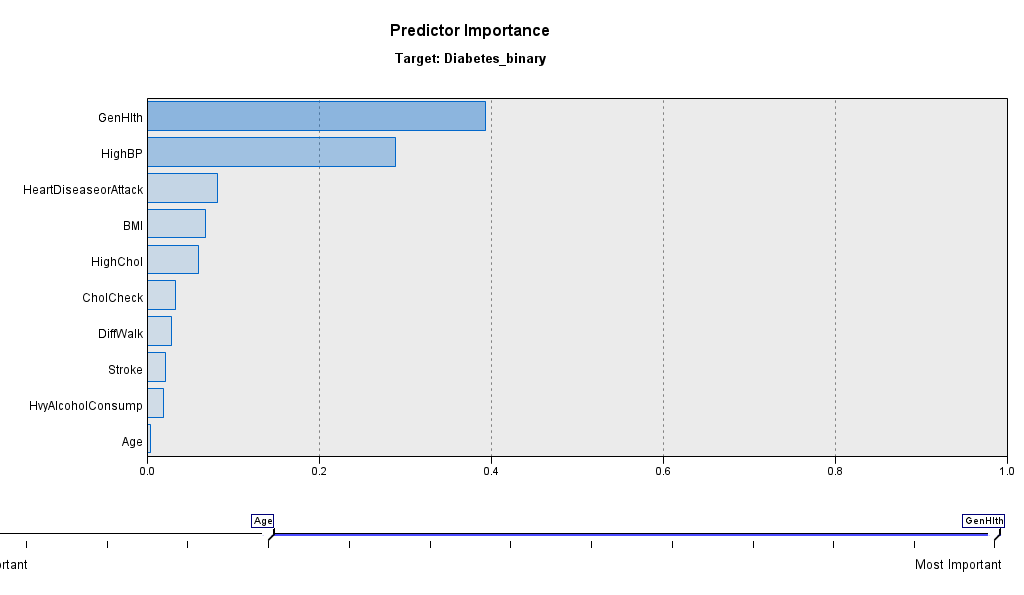
Hence from the above, there is no diabetes predicted for all the three individuals with respective probabilities (which correlates in the real time)

Also, the model is run under C 5 classification, where similar results compared to Naïve Bayesian model are obtained

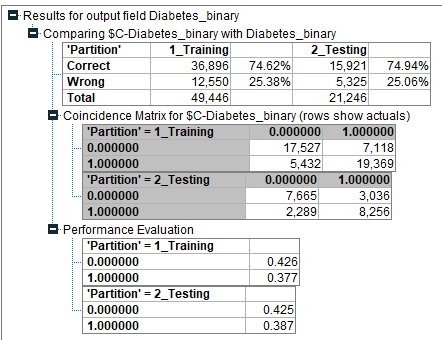
The stream is figured as below, attached as DMProject2C5



The predictor importance is as below where GenHlth ans HighBP has high importance and then comes the heart disease attack, BMI and other fields.



The analysis node gives results similar to Bayesian classification in above sections



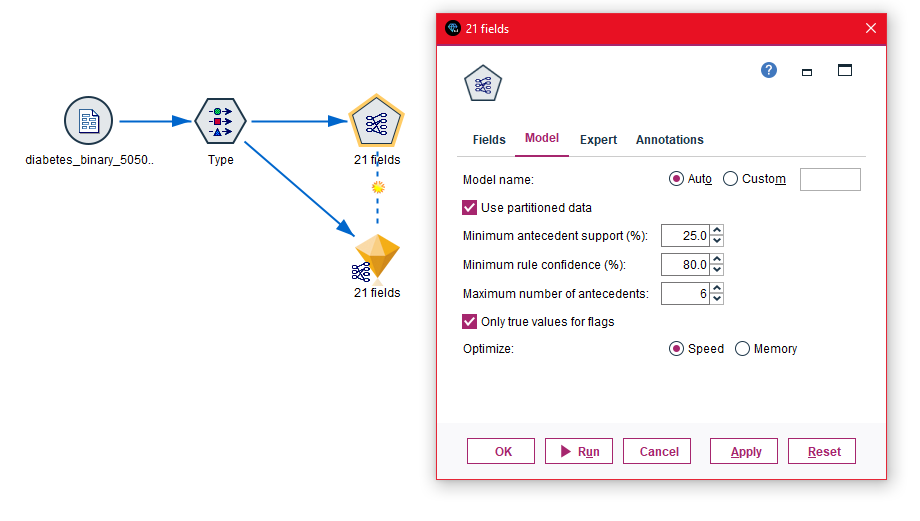
It is clear from the above figure that the model is accurate and similar results compared to Bayesian classification are obtained for recall, accuracy and specificity. The graphical analysis of the results are attached in the file.

## **Association**

Association analysis is done to find out what factors go in correlation with other factors. An attempt is made to associate the data in the given data set.

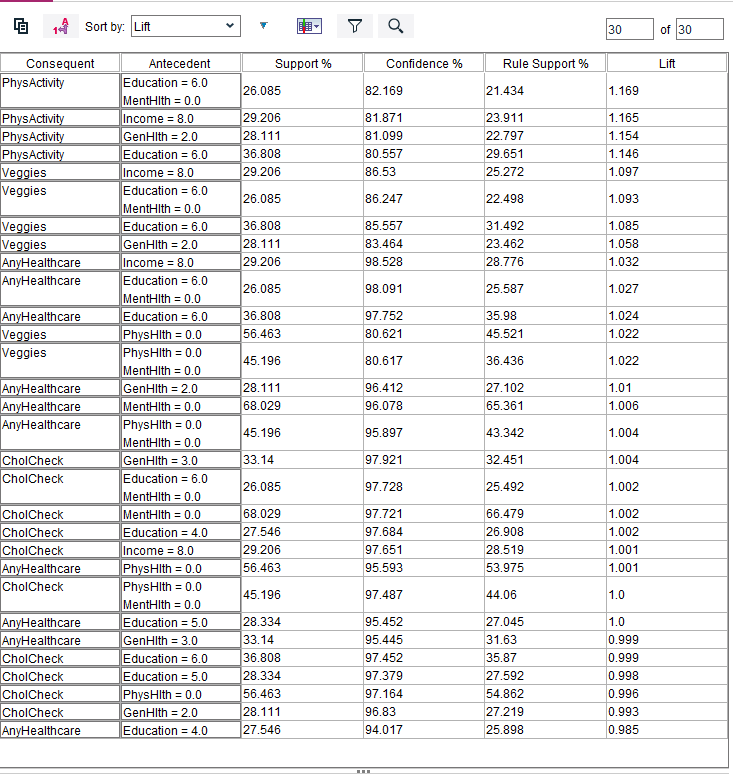
An association-based analysis is done on this dataset by creating the following model. Though the model is trivial and not a good use of this rule, an attempt is made to show the extension to all the data types in this data set except continuous data type attributes. The stream is attached with file name DMProject2Association

Here Apriori algorithm is imposed by the Apriori model in modeling tab od SPSS modeler. The algorithm is applied to all the variables except data with continuous data type using a minimum antecedent support of 25% and minimum confidence of 80%. In the field tab, all the 22 fields are given as the consequents while Antecedents considered are GenHlth, MentHlth, PhysHlth, Age, Education, Income (all the other data types except continuous and flag)

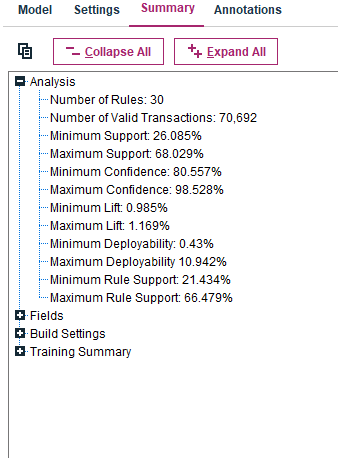


The result is the figure below

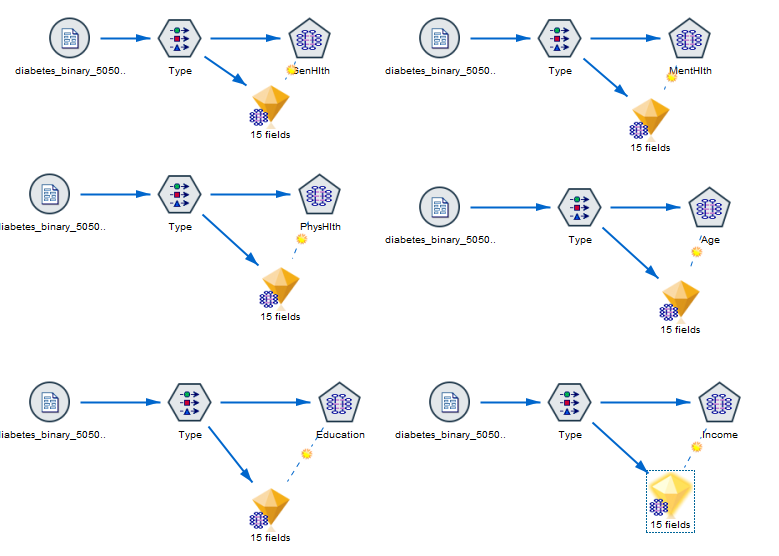
In the above output table, all the lift values greater than 1 can be considered as the best rules. However, the improvement value for consequent as PhyActivity and antecedents Education = 6 and MentHlth = 0 is higher with 1.169 lift with 82% confidence, 26% support, and 21.434 rule support. This can be formulated to rank 1 as { PhyActivity} = > { Education = 6, MentHlth = 0}.



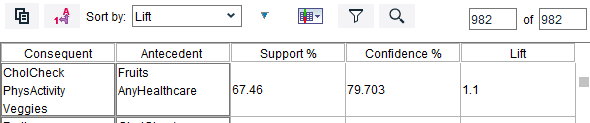
The outputs are summarized as below



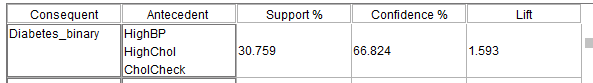
The data is imposed to similar association analysis using Carma node as in below figure. In this analysis all the data fields are consider corresponding to individual nominal fields.



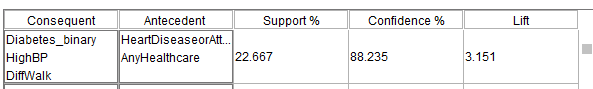
GenHlth resulted in following lift



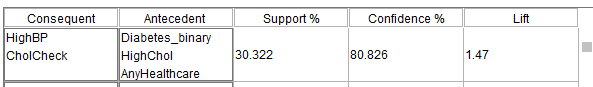
PhysHlth



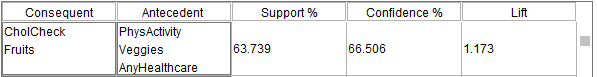
Education



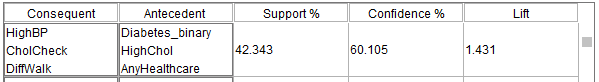
MentHlth



Age



Income

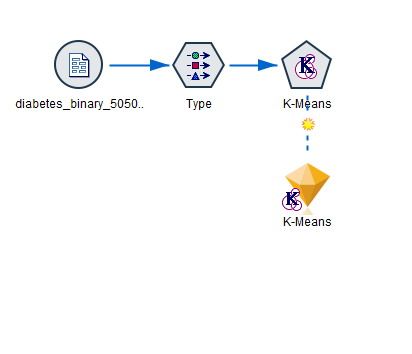


From the above figures, it can be inferred that the lift value is high for the association of all the data fields with education field lift of 3.51 with a confidence of 88% and support of 22%.

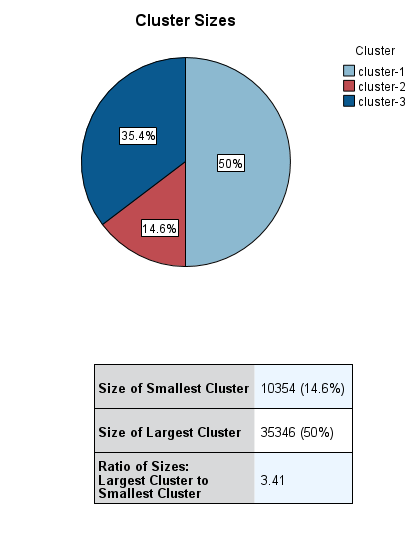
## **Clustering**

Clustering is a collection of data objects which are similar to one another. The objects within the same cluster represent similarity and dissimilar objects belong to another cluster. It is nothing but grouping of the data. Here since we have a large data set, an attempt is made to cluster the whole data into 4 small clusters based on few similarities.

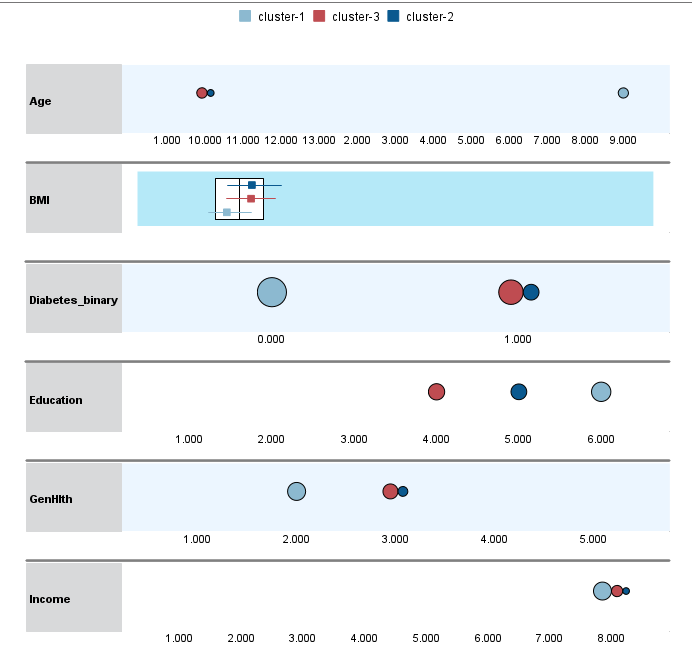
In this analysis, the data fields corresponding to BMI, GenHlth, Age, Education and income are chosen as primary factors on which the clustering depends.



The resulting output from K means super node is



Where there are three clusters with largest cluster size of 50% and smallest with 14.6% with 10,354 records.



The above table summarizes the clusters comparison. Two clusters are formed at age scale 10 and one at 9. The BMI index of the three clusters form at closer data points. Large cluster is formed with more no diabetes individuals while 2 clusters are formed covering the individuals with diabetes. While similar pattern is taken for the GenHlth and Income fields, education field has each of the clusters at scale 4,5,6. Since there are no defined target fields in this analysis, no graphical or matrix comparison can be made in between two fields

# **Conclusions**

As per the results and discussions in the above sections, the conclusions on he final outcome is as below

* The data analysis part helped in choosing the data types for different fields as the data types chosen in the fields play a vital role in changing the result in every model.
* Since there are not many extreme values and outliers, the dataset is used as it is with no change, assuming the data is accurate with no errors.
* Then the analysis is proceeded with the prediction, where a mode is developed to predict the occurrence of diabetes in an individual and the model is checked by giving 3 examples and evaluated for its accuracy.
* This model developed by Naïve Bayesian classification is them compared with C5 classification, where the results seems to be correlating. The performance matrix for both the classification methods give almost same results. Hence, we can conclude on the correctness of the model.
* After predicting the diabetes fields, the model is checked for any associations. This is also done in two methods by Apriori method and Carma method. In Apriori method, a higher lift value of 1.169 is obtained while in Carma method, by individually associating all the data fields to one nominal data type field, associations with Education attribute give a maximum lift of 3.51 with 88 confidence and ranked 1. This shows that the associations are more with education field than other fields.
* The model is then analyzed for clustering which does not give satisfactory result, though the silhouette value for this model is 0.3, fair. Since there are many scaled attributes, the clusters are more grouped in one scale rather than dispersed cluster formation.

Hence from the above conclusions, the data set is evaluated under different types of data mining analysis and predictions which can be helpful in future when someone needs to make use of analyzing different data types, predicting the diabetes, clustering a particular record in the data set and finally checking for associations.

# **References**:

1. <https://www.cdc.gov/brfss/annual_data/2015/pdf/2015_calculated_variables_version4.pdf>
2. <https://rstudio-pubs-tatic.s3.amazonaws.com/482619_cb8f1da13960497ebbcbe1d9e1efa7d5.html>
3. <https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset?select=diabetes_binary_health_indicators_BRFSS2015.csv>
4. <https://favtutor.com/blogs/data-mining-projects>
5. <https://www.researchgate.net/publication/338581650_Diabetic_Prediction_System_Using_Data_Mining>