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

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.

2. 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.

3. Data Explanation

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

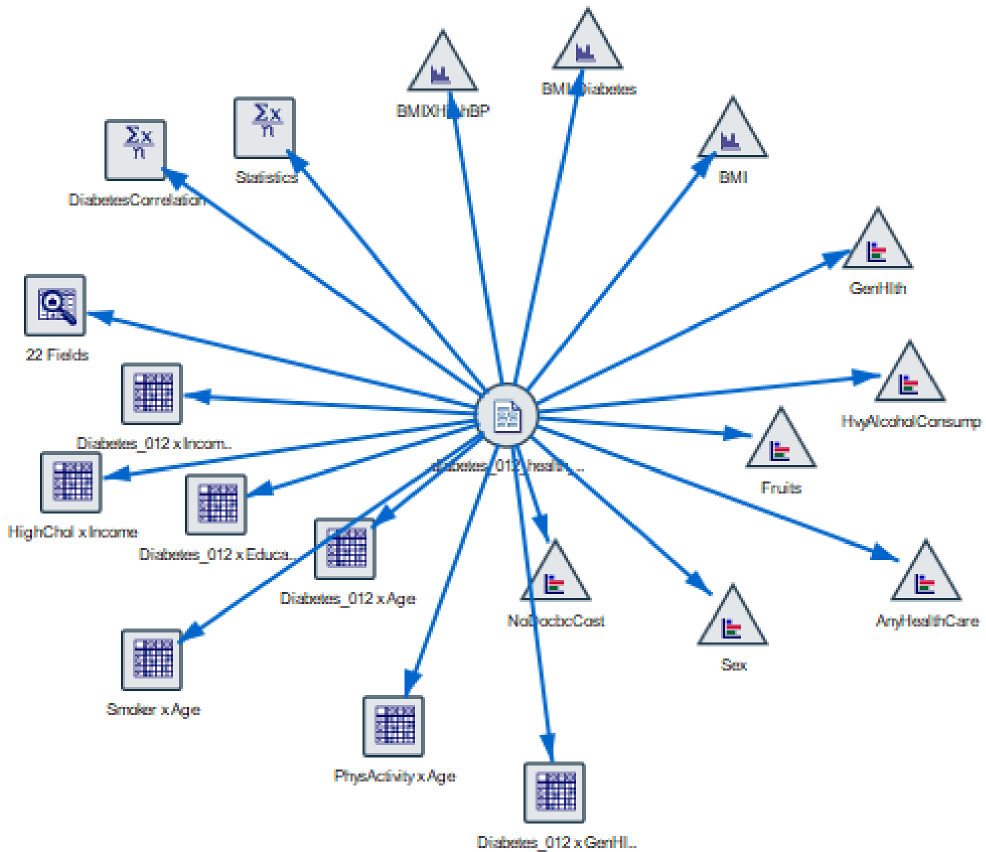
The reduced dataset used in this work consists of 70,692 responses on the diabetes condition in the individual based on 21 below factors

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	Discrete

	couldn’t because of cost: 0 =no, 1= yes	
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

4. 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

Field	Sample Graph	Measurement	Min	Max	Mean	Std. Dev	Skewness	Unique	Valid
Diabetes_binary		Flag	0.000	1.000	--	--	--	2	70692
HighBP		Flag	0.000	1.000	--	--	--	2	70692
HighChol		Flag	0.000	1.000	--	--	--	2	70692
CholCheck		Flag	0.000	1.000	--	--	--	2	70692
BMI		Continuous	12.000	98.000	29.857	7.114	1.719	--	70692
Smoker		Flag	0.000	1.000	--	--	--	2	70692
Stroke		Flag	0.000	1.000	--	--	--	2	70692
HeartDiseaseorAttack		Flag	0.000	1.000	--	--	--	2	70692

Field	Sample Graph	Measurement	Min	Max	Mean	Std. Dev	Skewness	Unique	Valid
PhysActivity		Flag	0.000	1.000	--	--	--	2	70692
Fruits		Flag	0.000	1.000	--	--	--	2	70692
Veggies		Flag	0.000	1.000	--	--	--	2	70692
HvyAlcoholConsump		Flag	0.000	1.000	--	--	--	2	70692
AnyHealthcare		Flag	0.000	1.000	--	--	--	2	70692
NoDocbcCost		Flag	0.000	1.000	--	--	--	2	70692
GenHlth		Nominal	1.000	5.000	--	--	--	5	70692
MentHlth		Ordinal	0.000	30.000	--	--	--	31	70692

PhysHlth		Ordinal	0.000	30.000	--	--	--	31	70692
DiffWalk		Flag	0.000	1.000	--	--	--	2	70692
Sex		Flag	0.000	1.000	--	--	--	2	70692
Age		Nominal	1.000	13.000	--	--	--	13	70692
Education		Nominal	1.000	6.000	--	--	--	6	70692
Income		Nominal	1.000	8.000	--	--	--	8	70692

Field	Measurement	Outliers	Extremes	Action	Impute Miss...	Method	% Compl...	Valid Recor...	Null Value	Empty Stri...	White Space	Blank Value
Diabetes_bi...	Flag	--	--	--	Never	Fixed	100	70692	0	0	0	0
HighBP	Flag	--	--	--	Never	Fixed	100	70692	0	0	0	0
HighChol	Flag	--	--	--	Never	Fixed	100	70692	0	0	0	0
CholCheck	Flag	--	--	--	Never	Fixed	100	70692	0	0	0	0
BMI	Continuous	619	182	None	Never	Fixed	100	70692	0	0	0	0
Smoker	Flag	--	--	--	Never	Fixed	100	70692	0	0	0	0
Stroke	Flag	--	--	--	Never	Fixed	100	70692	0	0	0	0
HeartDiseas...	Flag	--	--	--	Never	Fixed	100	70692	0	0	0	0
PhysActivity	Flag	--	--	--	Never	Fixed	100	70692	0	0	0	0
Fruits	Flag	--	--	--	Never	Fixed	100	70692	0	0	0	0
Veggies	Flag	--	--	--	Never	Fixed	100	70692	0	0	0	0
HvyAlcoholC...	Flag	--	--	--	Never	Fixed	100	70692	0	0	0	0
AnyHealthcare	Flag	--	--	--	Never	Fixed	100	70692	0	0	0	0
NoDocbcCost	Flag	--	--	--	Never	Fixed	100	70692	0	0	0	0
GenHlth	Nominal	--	--	--	Never	Fixed	100	70692	0	0	0	0
MentHlth	Ordinal	--	--	--	Never	Fixed	100	70692	0	0	0	0
PhysHlth	Ordinal	--	--	--	Never	Fixed	100	70692	0	0	0	0
DiffWalk	Flag	--	--	--	Never	Fixed	100	70692	0	0	0	0
Sex	Flag	--	--	--	Never	Fixed	100	70692	0	0	0	0
Age	Nominal	--	--	--	Never	Fixed	100	70692	0	0	0	0
Education	Nominal	--	--	--	Never	Fixed	100	70692	0	0	0	0
Income	Nominal	--	--	--	Never	Fixed	100	70692	0	0	0	0

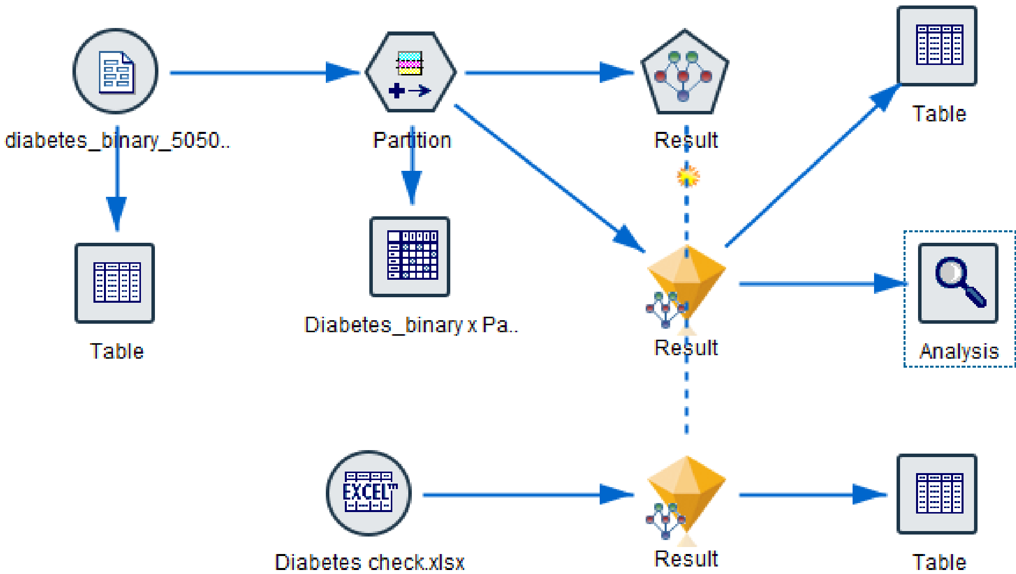
4.1.Prediction

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)

	Partition	
Diabetes_binary	1_Training	2_Testing
0.0	24645	10701
1.0	24801	10545

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

Results for output field Diabetes_binary

Comparing $B-Diabetes_binary with Diabetes_binary

'Partition'	1_Training		2_Testing	
Correct	36,665	74.15%	15,844	74.57%
Wrong	12,781	25.85%	5,402	25.43%
Total	49,446		21,246	

Coincidence Matrix for $B-Diabetes_binary (rows show actuals)

'Partition' = 1_Training	0.000000	1.000000
0.000000	17,422	7,223
1.000000	5,558	19,243

'Partition' = 2_Testing	0.000000	1.000000
0.000000	7,585	3,116
1.000000	2,286	8,259

Performance Evaluation

'Partition' = 1_Training	
0.000000	0.419
1.000000	0.371

'Partition' = 2_Testing	
0.000000	0.422
1.000000	0.38

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.

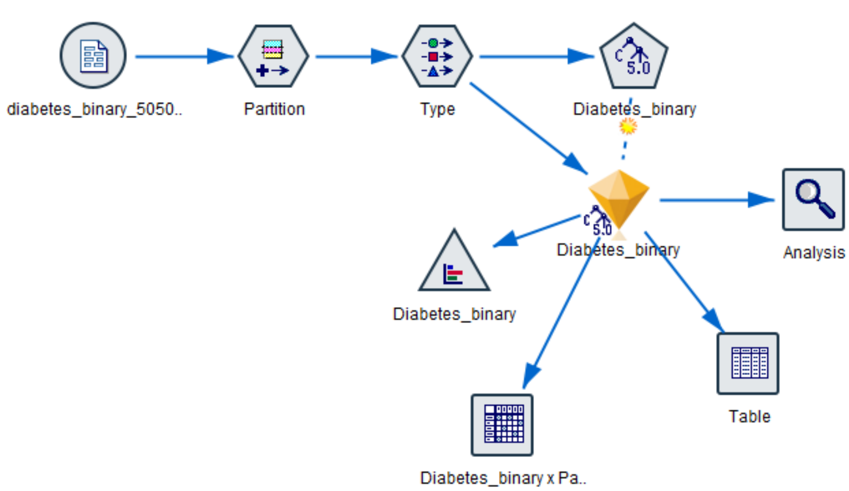
The model has been checked by giving our daily details in an excel sheet, (named as Diabetes check.xlsx) (the details are however exaggerated from the real) to predict whether we are in the diabetic zone and the results are

	BP	HighChol	CholCheck	BMI	Smoker	Stroke	HeartD...	PhysActivity	Fruits	Veggies	HvyAlco...	AnyH...	NoD...	GenHlth	MentHlth	PhysHlth	DiffWalk	Sex	Age	Education	Income	Name	$B-Diabet...	$BP-Diab...
1	00	0.000	0.000	11....	0.000	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000	3.000	0.000	0.000	0.000	0....	3.0...	5.000	1.000	Swet...	0.000	0.975
2	00	1.000	0.000	10....	1.000	0.000	0.000	0.000	0.000	1.000	1.000	1.000	1.000	2.000	0.000	1.000	1.000	1....	3.0...	4.000	1.000	Venu	0.000	0.988
3	00	0.000	0.000	13....	1.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	4.000	0.000	0.000	0.000	1....	3.0...	4.000	1.000	Srujan	0.000	0.831

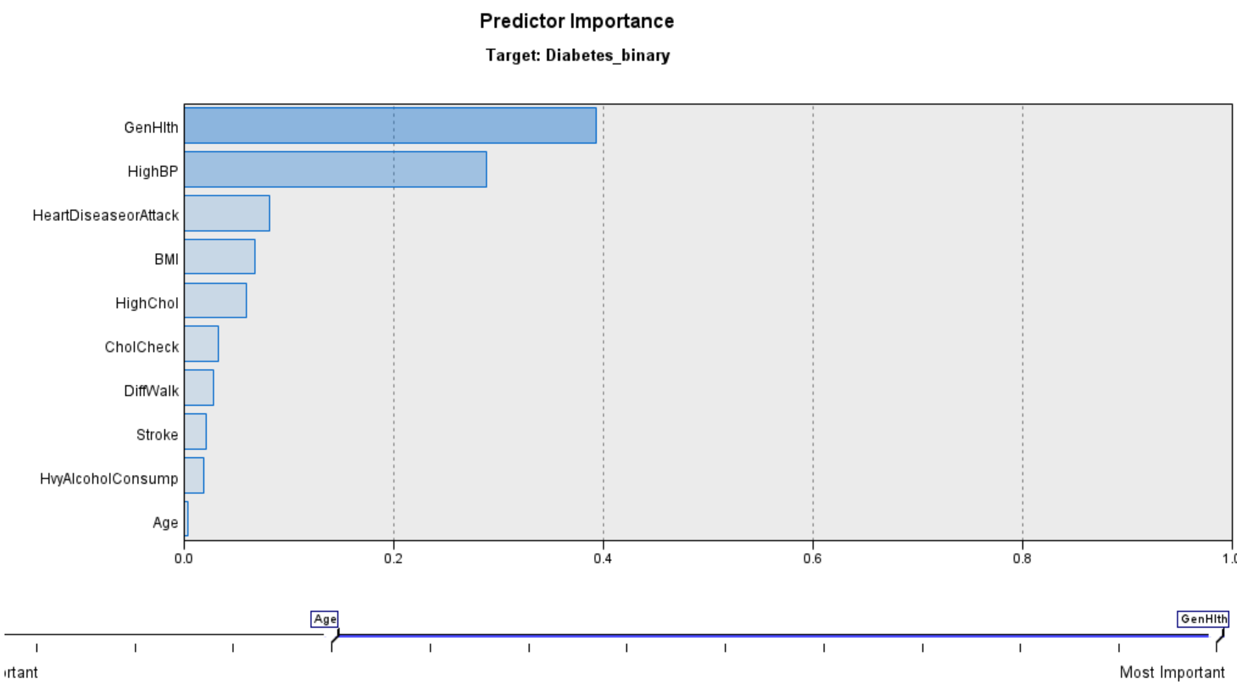
Hence from the above, there is no diabetes predicted for all the three individuals with respective probabilities (which correlates in the real time)

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

Results for output field Diabetes_binary

Comparing $C-Diabetes_binary with Diabetes_binary

'Partition'	1_Training		2_Testing	
Correct	36,896	74.62%	15,921	74.94%
Wrong	12,550	25.38%	5,325	25.06%
Total	49,446		21,246	

Coincidence Matrix for $C-Diabetes_binary (rows show actuals)

'Partition' = 1_Training	0.000000	1.000000
0.000000	17,527	7,118
1.000000	5,432	19,369

'Partition' = 2_Testing	0.000000	1.000000
0.000000	7,665	3,036
1.000000	2,289	8,256

Performance Evaluation

'Partition' = 1_Training	
0.000000	0.426
1.000000	0.377

'Partition' = 2_Testing	
0.000000	0.425
1.000000	0.387

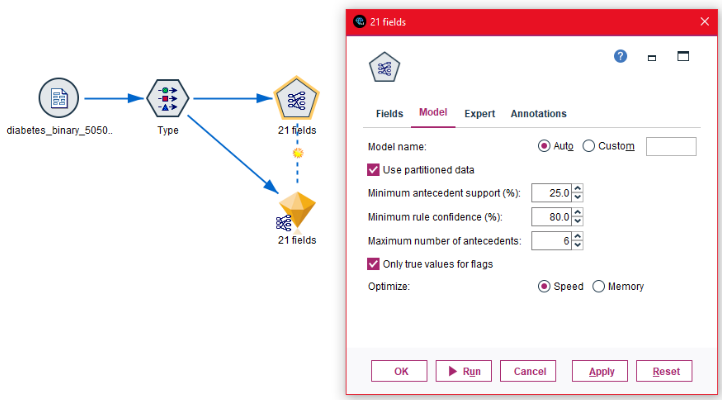
It is clear from the above figure that the model is accurate and similar results compared to Bayesian classification are obtained for recall, accuracy and specificity. The graphical analysis of the results are attached in the file.

4.2.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}.

Sort by: Lift — 30 of 30

Consequent	Antecedent	Support %	Confidence %	Rule Support %	Lift
PhysActivity	Education = 6.0 MentHlth = 0.0	26.085	82.169	21.434	1.169
PhysActivity	Income = 8.0	29.206	81.871	23.911	1.165
PhysActivity	GenHlth = 2.0	28.111	81.099	22.797	1.154
PhysActivity	Education = 6.0	36.808	80.557	29.651	1.146
Veggies	Income = 8.0	29.206	86.53	25.272	1.097
Veggies	Education = 6.0 MentHlth = 0.0	26.085	86.247	22.498	1.093
Veggies	Education = 6.0	36.808	85.557	31.492	1.085
Veggies	GenHlth = 2.0	28.111	83.464	23.462	1.058
AnyHealthcare	Income = 8.0	29.206	98.528	28.776	1.032
AnyHealthcare	Education = 6.0 MentHlth = 0.0	26.085	98.091	25.587	1.027
AnyHealthcare	Education = 6.0	36.808	97.752	35.98	1.024
Veggies	PhysHlth = 0.0	56.463	80.621	45.521	1.022
Veggies	PhysHlth = 0.0 MentHlth = 0.0	45.196	80.617	36.436	1.022
AnyHealthcare	GenHlth = 2.0	28.111	96.412	27.102	1.01
AnyHealthcare	MentHlth = 0.0	68.029	96.078	65.361	1.006
AnyHealthcare	PhysHlth = 0.0 MentHlth = 0.0	45.196	95.897	43.342	1.004
CholCheck	GenHlth = 3.0	33.14	97.921	32.451	1.004
CholCheck	Education = 6.0 MentHlth = 0.0	26.085	97.728	25.492	1.002
CholCheck	MentHlth = 0.0	68.029	97.721	66.479	1.002
CholCheck	Education = 4.0	27.546	97.684	26.908	1.002
CholCheck	Income = 8.0	29.206	97.651	28.519	1.001
AnyHealthcare	PhysHlth = 0.0	56.463	95.593	53.975	1.001
CholCheck	PhysHlth = 0.0 MentHlth = 0.0	45.196	97.487	44.06	1.0
AnyHealthcare	Education = 5.0	28.334	95.452	27.045	1.0
AnyHealthcare	GenHlth = 3.0	33.14	95.445	31.63	0.999
CholCheck	Education = 6.0	36.808	97.452	35.87	0.999
CholCheck	Education = 5.0	28.334	97.379	27.592	0.998
CholCheck	PhysHlth = 0.0	56.463	97.164	54.862	0.996
CholCheck	GenHlth = 2.0	28.111	96.83	27.219	0.993
AnyHealthcare	Education = 4.0	27.546	94.017	25.898	0.985

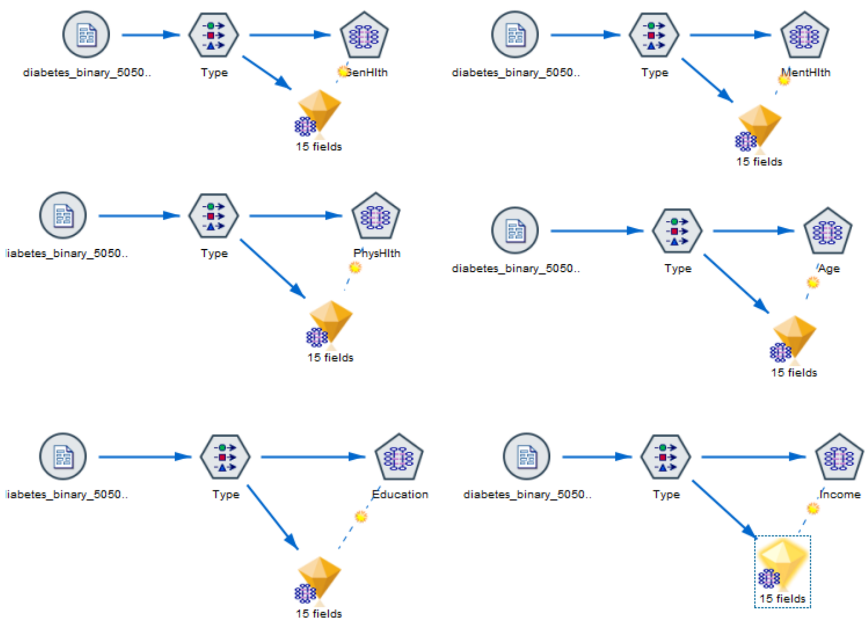
The outputs are summarized as below

Model Settings **Summary** Annotations

Collapse All Expand All

- Analysis
  - Number of Rules: 30
  - Number of Valid Transactions: 70,692
  - Minimum Support: 26.085%
  - Maximum Support: 68.029%
  - Minimum Confidence: 80.557%
  - Maximum Confidence: 98.528%
  - Minimum Lift: 0.985%
  - Maximum Lift: 1.169%
  - Minimum Deployability: 0.43%
  - Maximum Deployability 10.942%
  - Minimum Rule Support: 21.434%
  - Maximum Rule Support: 66.479%
- Fields
- Build Settings
- Training Summary

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

Sort by: Lift 982 of 982

Consequent	Antecedent	Support %	Confidence %	Lift
CholCheck PhysActivity Veggies	Fruits AnyHealthcare	67.46	79.703	1.1

PhysHlth

Consequent	Antecedent	Support %	Confidence %	Lift
Diabetes_binary	HighBP HighChol CholCheck	30.759	66.824	1.593

Education

Consequent	Antecedent	Support %	Confidence %	Lift
Diabetes_binary HighBP DiffWalk	HeartDiseaseorAtt... AnyHealthcare	22.667	88.235	3.151

MentHlth

Consequent	Antecedent	Support %	Confidence %	Lift
HighBP CholCheck	Diabetes_binary HighChol AnyHealthcare	30.322	80.826	1.47

Age

Consequent	Antecedent	Support %	Confidence %	Lift
CholCheck Fruits	PhysActivity Veggies AnyHealthcare	63.739	66.506	1.173

Income

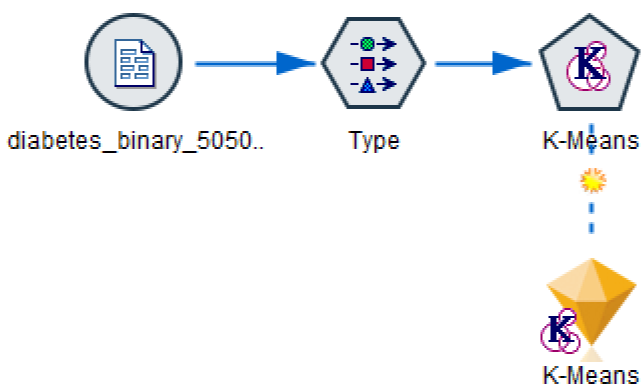
Consequent	Antecedent	Support %	Confidence %	Lift
HighBP CholCheck DiffWalk	Diabetes_binary HighChol AnyHealthcare	42.343	60.105	1.431

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%.

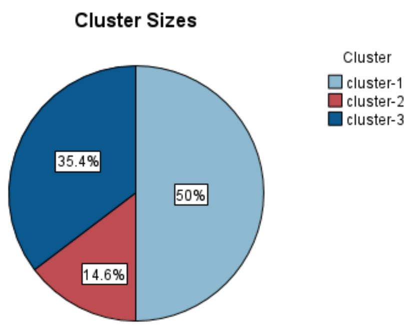
4.3.Clustering

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.

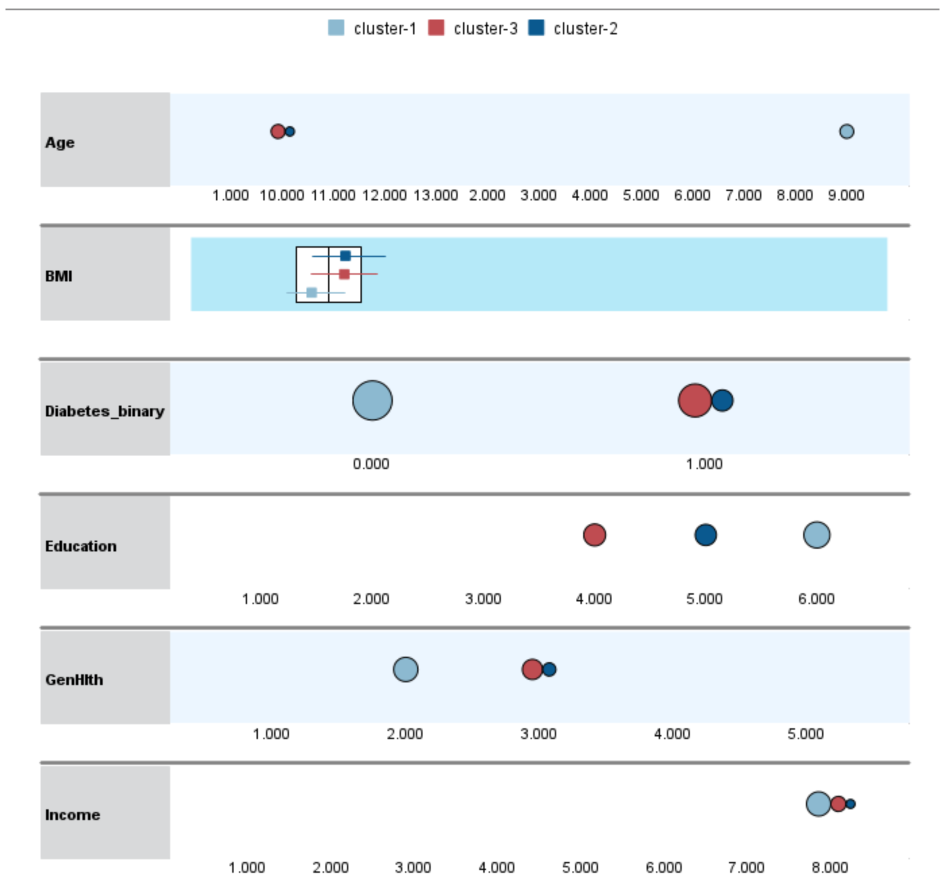


The resulting output from K means super node is



Size of Smallest Cluster	10354 (14.6%)
Size of Largest Cluster	35346 (50%)
Ratio of Sizes: Largest Cluster to Smallest Cluster	3.41

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

5. 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.

6. **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