e-ISBN : 978-629-98755-5-0

SIG : e-Learning@CS
https://appspenang.uitm.edu.my/sigcs/
Publication Date : 24 – Mar - 2025

ANALYSIS USING DATA MINING TECHNIQUES: THE EXPLORATION AND REVIEW DATA OF DIABETES PATIENTS

*Syarifah Adilah Mohamed Yusoff[1], Jamal Othman[2], Elly Johana Johan[3], Azlina Mohd Mydin[4] and Wan Anisha Wan Mohamad[5]
**syarifah.adilah@uitm.edu.my[1], jamalothman@uitm.edu.my[2], ellyjohana@uitm.edu.my[3], azlin143@usm.edu.my[4], wanan122@usm.edu.my[5]*

[1,2,3,4,5]Jabatan Sains Komputer & Matematik (JSKM),
Universiti Teknologi MARA Cawangan Pulau Pinang, Malaysia

**Corresponding author*

ABSTRACT

Data mining is undergoing a transformative phase driven by advancements in Artificial Intelligence, statistics, database technology, real-time processing and integration of diverse data sources. These trends are not only enhancing the efficiency and accuracy of data mining but also expanding its applications across different industries. The subsequent step involves a comprehensive study of the dataset, incorporating both data exploration and analysis of data variables to achieve a structural and statistical understanding of the data. In this statistical summary procedure, the distribution of attributes and their interactions are crucial for accurately processing the data in accordance with the selected classification or data mining techniques to be performed. In examining the distribution of diabetes data, there are intricate interactions among the attributes. Therefore, it is advisable for future studies to implement robust classification algorithms, such as ensemble methods, to effectively manage and extract potential insights.

Keywords: data mining, classification, data interaction, attribute representation, data exploration

Introduction to fundamental concepts of data mining

In today's data-driven world, data mining has become an essential tool for organizations seeking to extract valuable insights from vast amounts of data. Data mining involves the process of discovering patterns, correlations, and anomalies within large datasets to predict outcomes and make informed decisions. As technology continues to evolve, several emerging trends are reshaping the landscape of data mining, making it more efficient, accurate, and impactful.

A prominent trend in data mining is the incorporation of artificial intelligence (AI) and machine learning (ML) to enhance algorithms derived from statistical methods and metaheuristic approaches. These technologies have revolutionized data mining by automating complex analysis processes and providing deeper, more actionable insights. AI-driven analytics enable organizations to process and analyze data at unprecedented speeds, uncovering hidden patterns and predicting future trends with greater accuracy (Current Trends & Future Scope of Data Mining, 2021). This has opened new avenues for innovation across various sectors, including finance, healthcare, retail, and manufacturing.

Figure 1 illustrates the relationship between AI and many fields of knowledge in the development of data mining technology, enhancing analytical outcomes. The descriptive analytics helps organizations understand past events and trends by summarizing historical data. Next, predictive analytics uses this historical data to forecast future events and identify potential risks and opportunities. Afterwards, prescriptive analytics goes a step further by recommending actions to achieve desired outcomes based on these predictions (han et al., 2011). Together, these three types of analytics enable organizations to make informed, data-driven decisions that enhance their performance and strategic planning.

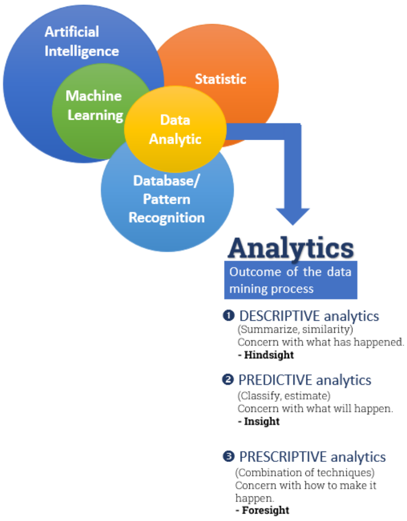


Figure 1: Data mining trends, disciplines and outcomes

Analytics starts upon the collection of data. This process has multiple steps or phases contingent upon the selected data mining techniques. Various techniques exist, including classification, clustering, association analysis, and text mining. The selected technique is reliant upon the expected outcomes of the research and the types of datasets employed. The most common technique is classification and Figure 2 illustrates the typical procedures involved in carrying out classification analysis on the utilized dataset. Initially, the acquisition of the dataset confronted certain issues that need a resolution. The subsequent step involves a comprehensive study of the dataset, incorporating both data exploration and analysis of data variables to achieve a structural and statistical understanding of the data. In this statistical summary procedure, the distribution of attributes and their interactions are crucial for accurately processing the data in accordance with the selected classification or data mining techniques to be performed. Afterwards, the data is prepared for pre-processing processes, which include addressing any missing data, outliers, and noise, as well as transforming or digitizing the data. Data is now prepared for classification analysis through the implementation of algorithms for training and testing prior to the generation of results (Shukri et al, 2024). Thus, this study intends to discuss thoroughly the phase of reviewing the data where the most fundamental statistical concept is established.

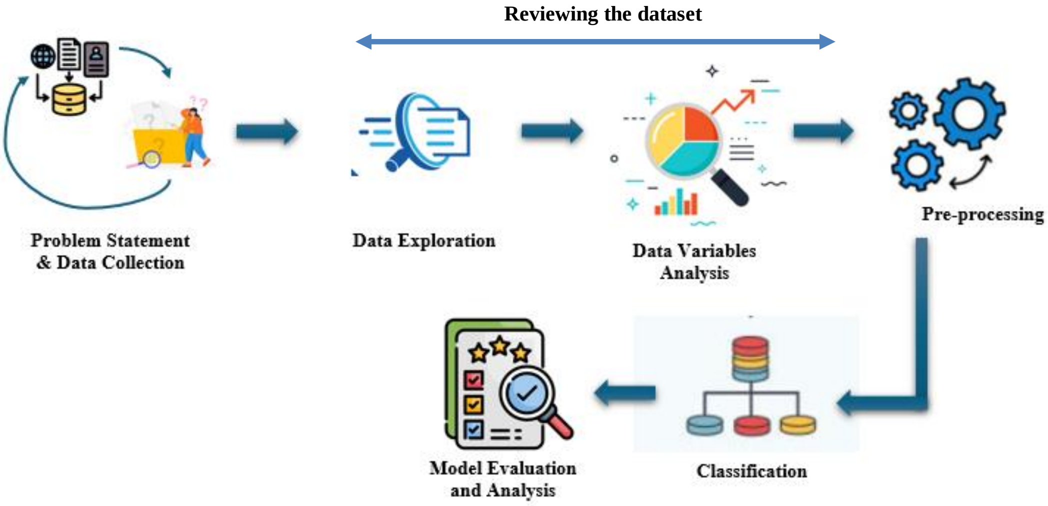


Figure 2: Six good reasons to take notes

Data Exploration

The diabetes patient dataset originally was taken through Kaggle website (TEBOUL, 2022) from Behavioral Risk Factor Surveillance System Overview 2015 of Centers for Disease Control and Prevention(.gov). The dataset consists of 13 attributes and a total of 253680 records or instances. Table 1 describes the detailed implementation of each attribute. Based on the description, the attributes were not limited to medical information only but consist of daily activities, diet and mental health record.

Table 1: Better life index 2024 description

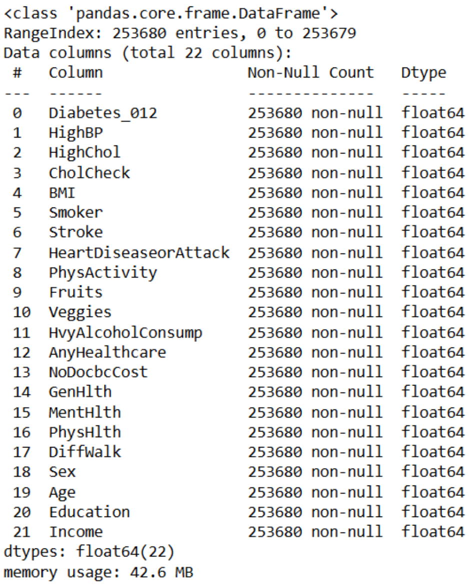
No	Attributes	Explanation
1	Diabetes_012	0 = no diabetes 1 = prediabetes 2 = diabetes
2	HighBP	0 = no high, BP 1 = high BP
3	HighChol	0 = no high cholesterol. 1 = high cholesterol
4	CholCheck	0 = no cholesterol check in 5 years. 1 = yes cholesterol check in 5 years
5	BMI	Body Mass Index
6	Smoker	Have you smoked at least 100 cigarettes in your entire life? [Note: 5 packs = 100 cigarettes]. (0 = no 1 = yes)
7	Stroke	you ever had a stroke. (0 = no, 1 = yes)
8	HeartDiseaseorAttack	coronary heart disease (CHD) or myocardial infarction (MI). (0 = no 1 = yes)
9	PhysActivity	physical activity in past 30 days - not including job. (0 = no 1 = yes)

10	Fruits	Consume Fruit 1 or more times per day. (0 = no 1 = yes)
11	Veggies	Consume Vegetables 1 or more times per day. (0 = no 1 = yes)
12	HvyAlcoholConsump	Adult men >=14 drinks per week and adult women>=7 drinks per week. (0 = no 1 = yes)
13	AnyHealthcare	Have any kind of health care coverage, including health insurance, prepaid plans such as HMO, etc. 0 = no 1 = yes
14	NoDocbcCost	Was there a time in the past 12 months when you needed to see a doctor but could not because of cost? 0 = no 1 = yes
15	GenHlth	Would you say that in general your health is: scale 1-5 (1 = excellent 2 = very good 3 = good 4 = fair 5 = poor)
16	MentHlth	Days of poor mental health scale 1-30 days
17	PhysHlth	Physical illness or injury days in past 30 days scale 1-30 days
18	DiffWalk	Do you have serious difficulty walking or climbing stairs? (0 = no 1 = yes)
19	Sex	Patient's gender (1: male; 0: female)
20	Age	13-level age category (1 = 18-24yrs / 2 = 25-29 yrs / 3 = 30-34 yrs / 4 = 35-39 yrs / 5 = 40-44 yrs / 6 = 45-49 yrs / 7 = 50-54 yrs / 8 = 55-59 yrs / 9 = 60-64 yrs / 10 = 65-69 yrs / 11 = 70-74 yrs / 12 = 75-79 yrs / 13 = 80 or older)
21	Education	Education level (EDUCA see codebook) scale 1-6 1 = Never attended school or only kindergarten 2 = Grades 1 - 8 (Elementary) 3 = Grades 9 - 11 (Some high school) 4 = Grade 12 or GED (High school graduate) 5 = College 1 year to 3 years (Some college or technical school) 6 = College 4 years or more (College graduate)
22	Income	Income scale (INCOME2 see codebook) scale 1-8 1 = less than $10,000, 2= $10,000 to less than $15,000, 3=$15,000 to less than $20,000, 4= $20,000 to less than $25,000, 5 = $25,000 to less than $35,000, 6= $35,000 to less than $50,000, 7= $50,000 to less than $75,000, 8 = $75,000 or more

Reviewing the content of the dataset

The summary of variables's data types, range index, columns, non-values and memory usage were illustrated as in Figure 3, where we can recognize that dimensions of the dataset was 253680 x 22 indicates the total of records of person was 253680 and each share 22 attributes information varies from medical, diets and general health. Further, the presence of non-null values indicated that all attributes were devoid of null values, signifying the absence of missing data. Finally, the data types 'float64' indicate that all attributes consist of decimal numeric values.

In comparison to the information presented in Table 1, some attribute data types designated as objects signify categorical values meanwhile the data frame as illustrated in Figure 2 saved the data as continuous numeric type. Since different data types may require different statistical analyses in machine learning approaches, it was essential to re-assign the appropriate data types according to the original attributes.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 253680 entries, 0 to 253679
Data columns (total 22 columns):
 #   Column                Non-Null Count   Dtype  
---  ------                --------------   -----  
 0   Diabetes_012          253680 non-null  float64
 1   HighBP                253680 non-null  float64
 2   HighChol              253680 non-null  float64
 3   CholCheck             253680 non-null  float64
 4   BMI                   253680 non-null  float64
 5   Smoker                253680 non-null  float64
 6   Stroke                253680 non-null  float64
 7   HeartDiseaseorAttack  253680 non-null  float64
 8   PhysActivity          253680 non-null  float64
 9   Fruits                253680 non-null  float64
 10  Veggies               253680 non-null  float64
 11  HvyAlcoholConsump     253680 non-null  float64
 12  AnyHealthcare         253680 non-null  float64
 13  NoDocbcCost           253680 non-null  float64
 14  GenHlth               253680 non-null  float64
 15  MentHlth              253680 non-null  float64
 16  PhysHlth              253680 non-null  float64
 17  DiffWalk              253680 non-null  float64
 18  Sex                   253680 non-null  float64
 19  Age                   253680 non-null  float64
 20  Education             253680 non-null  float64
 21  Income                253680 non-null  float64
dtypes: float64(22)
memory usage: 42.6 MB
```

Figure 3: Summary of the data in rows and columns of the diabetes dataset.

Univariate Analysis

Univariate analysis was employed to examine the distribution of data for each attribute and assess the significance of the data, determining whether it was suitable for further analysis or necessitated statistical correction. Figure 4 depicts the distribution of data for each attribute. As highlighted earlier the output attribute was Diabetes012 and other attributes were candidates of input attributes. Review the total of 21 attributes of input variables, some of the data poorly distributed and imbalance such as CholCheck, Stroke, HeartDiseaseorAttack, PhysActivity, veggies, HvyAlchoholConsump, AnyHealthcare, NoDocbcCost, DiffWalk. The distribution of each attribute in Figure 4 indicates they were categorized as float (represented automatically in the Python data frame in Figure 3) but nominal data types, as originally described in Table 1. Therefore, the data types of each attribute except BMI will be changed to nominal as originally stated in Table 1 and in line with the distribution of the original data.

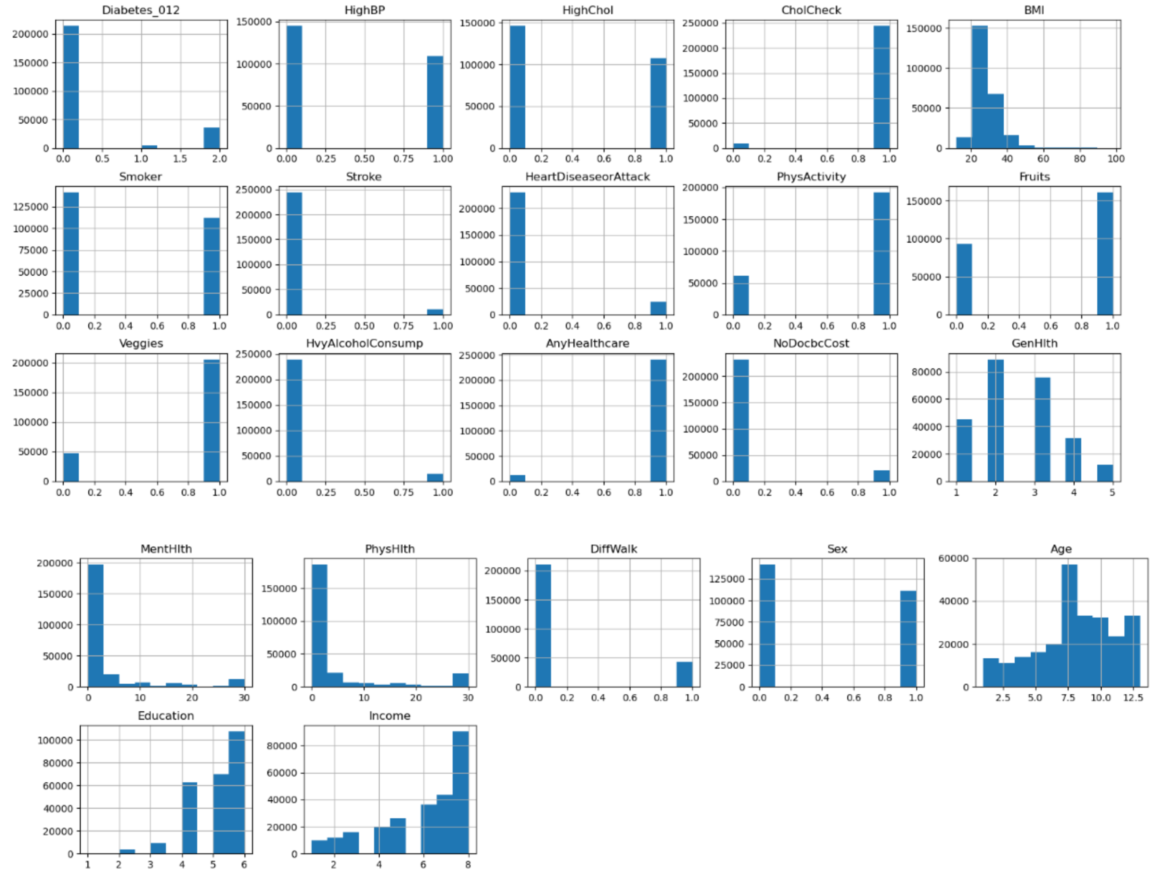
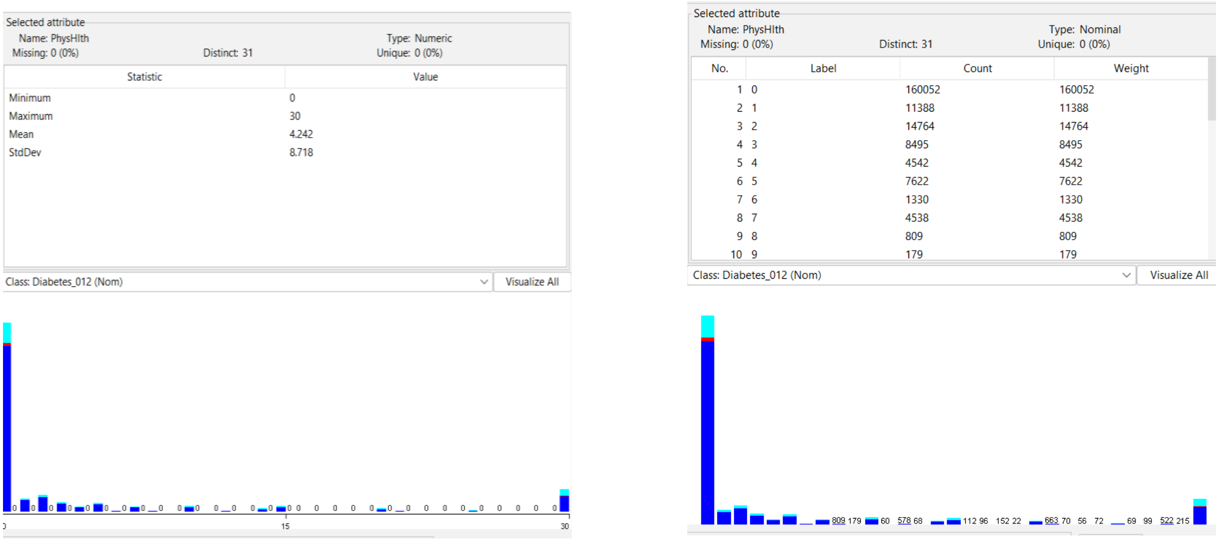


Figure 4: Frequency of each value from all attributes.

Next, the data transferred into WEKA application to make easier for further statistical exploration which at first all the suppose nominal attributes were converted using filter ‘*numerictonominal*’ and attribute Diabetes_012 was assigned as class attribute (output attribute). Figure 5 depict the changes of the statistical properties of the attribute named PhysHlth, when the attribute’s types was changed from numeric to nominal types. The figure shows that if the attribute’s type was numeric, the statistical properties observation based on min, max, mean and standard deviation. Meanwhile if nominal the distribution based on count of each given value for the attribute which were 0 to 30 that indicate the number or patient that had experienced PhysHlth problem for last 30 days.



Selected attribute
Name: PhysHlth
Missing: 0 (0%)
Distinct: 31
Type: Numeric
Unique: 0 (0%)

Statistic	Value
Minimum	0
Maximum	30
Mean	4.242
StdDev	8.718

Class: Diabetes_012 (Nom)

Selected attribute
Name: PhysHlth
Missing: 0 (0%)
Distinct: 31
Type: Nominal
Unique: 0 (0%)

No.	Label	Count	Weight
1	0	160052	160052
2	1	11388	11388
3	2	14764	14764
4	3	8495	8495
5	4	4542	4542
6	5	7622	7622
7	6	1330	1330
8	7	4538	4538
9	8	809	809
10	9	179	179

Class: Diabetes_012 (Nom)

Figure 5: Comparison of statistical data distribution between numeric and nominal data.

The next Figure 6 illustrates the latest distribution of data for each attribute in relation to the class attribute. The blue color shows no diabetes patients, red color shows pre-diabetes patients and turquoise shows diabetes patients. From each visualization of the attribute, distribution of the data based on the three class easily being observed. Each attribute was dominate by blue class color and very minimal from turquoise and red classes. This indicates further learning analysis of machine learning will dominate by blue class since the model will learn too much from blue class data and overshadow the minimal classes of turquoise and red class (Wongvorachan et al, 2023). To resolve this issue, the imbalance in class distribution must be rectified.

Figure 7(a) illustrates the current distribution of class output consist of 84.24 percent of the data is from class no diabetes, 1.83 percent of pre-diabetes and 13.93 percent of the data is diabetes. This imbalance class distribution can be rectified in three ways which are; 1) Oversampling; 2) Undersampling; and 3) SMOTE (Synthetic Minority Over-Sampling Technique) (Liu et al, 2022). In overall all classes contain data that are more than 1000 which is adequate to run data mining analysis. Further, in comparing these 3 classes, the pre-diabetes class is disproportionately small, including just 1.83 percent. Hence, excluding this data appears more advantageous for facilitating a seamless and successful analysis.

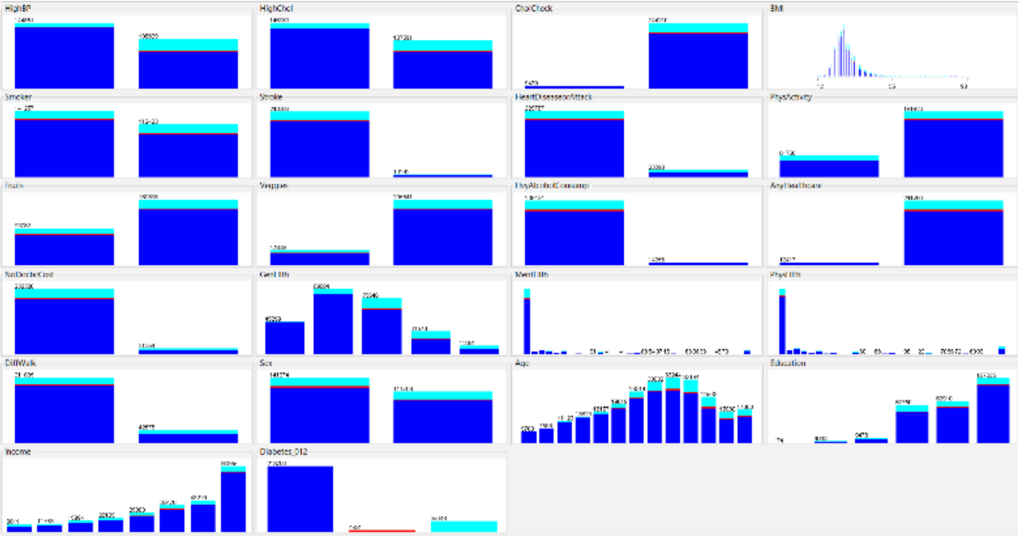
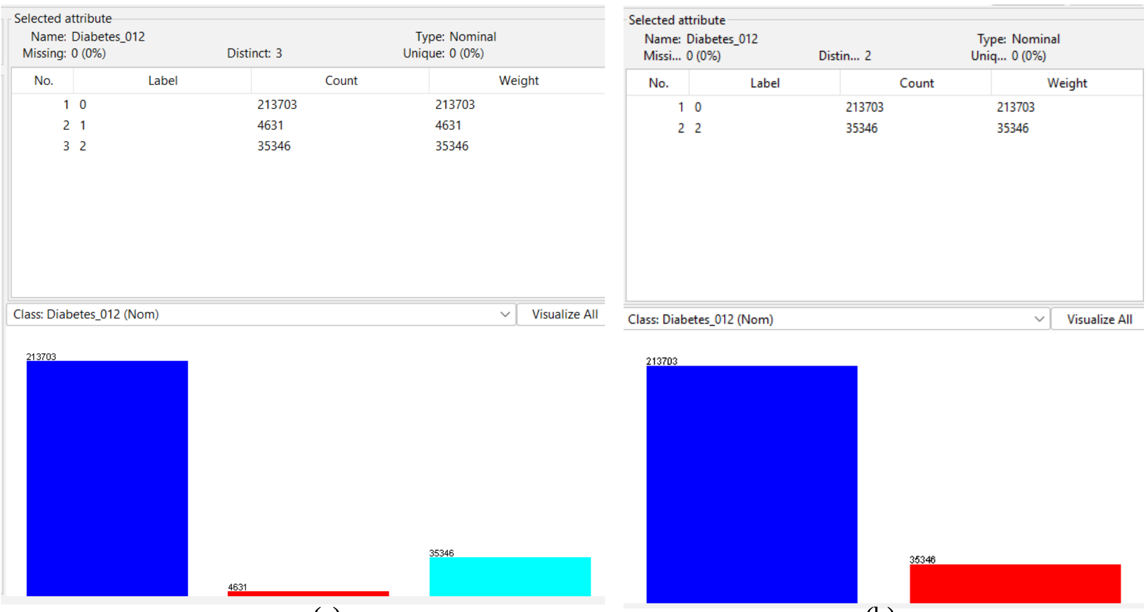


Figure 6: Distribution of the attributes with relation to class attribute Diabetes_012

Figure 7(b) shows the distribution of the class diabetes (red color) and no diabetes (blue color). To address the imbalance issue, the undersampling strategy was selected, which involves reducing the number of samples from the majority class to align with the minority class, after initially examining the whole data in both classes without generating new synthetic samples. Figure 8 shows both classes were balance in term of distribution of the data using undersampling approach. Nonetheless, there is no study indicating that any methods of undersampling, oversampling, or SMOTE are preferable to one another.



Selected attribute
Name: Diabetes_012 Missing: 0 (0%) Distinct: 3 Type: Nominal Unique: 0 (0%)

No.	Label	Count	Weight
1	0	213703	213703
2	1	4631	4631
3	2	35346	35346

Class: Diabetes_012 (Nom)

(a)

Selected attribute
Name: Diabetes_012 Missi... 0 (0%) Distin... 2 Type: Nominal Uniq... 0 (0%)

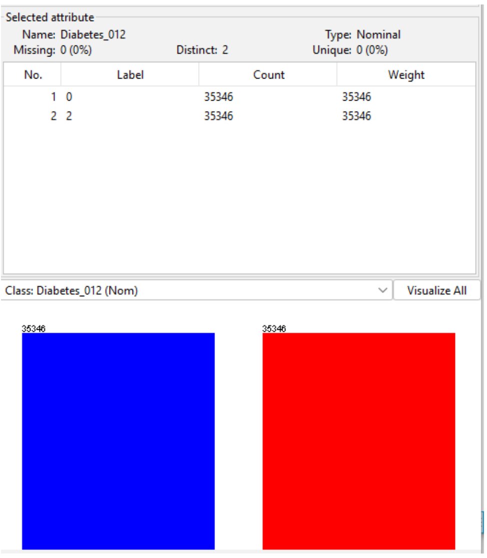
No.	Label	Count	Weight
1	0	213703	213703
2	2	35346	35346

Class: Diabetes_012 (Nom)

(b)

Figure 7: Distribution of the attributes with relation to class attribute Diabetes_012

Next, Figure 9 illustrates overall data distribution of 70692 instances or observations among 21 input attributes and one output attribute of Diabetes_012. Following the aforementioned enhancements to the classes, the distribution of the data indicated by red and blue colors have significantly improved across all attributes.



Selected attribute
Name: Diabetes_012 Type: Nominal
Missing: 0 (0%) Distinct: 2 Unique: 0 (0%)

No.	Label	Count	Weight
1	0	35346	35346
2	2	35346	35346

Class: Diabetes_012 (Nom)

Figure 8: Balance data between two class using undersampling approach

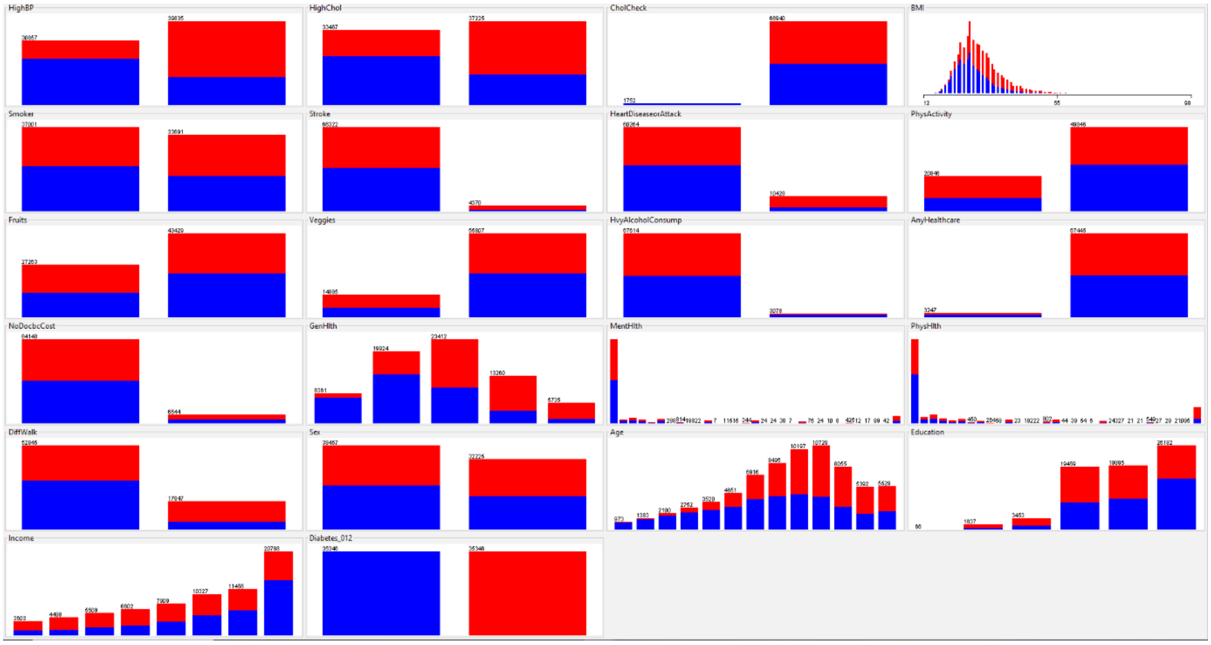


Figure 9: Distribution of all attributes in relation with class attribute Diabetes_012 after rectification.

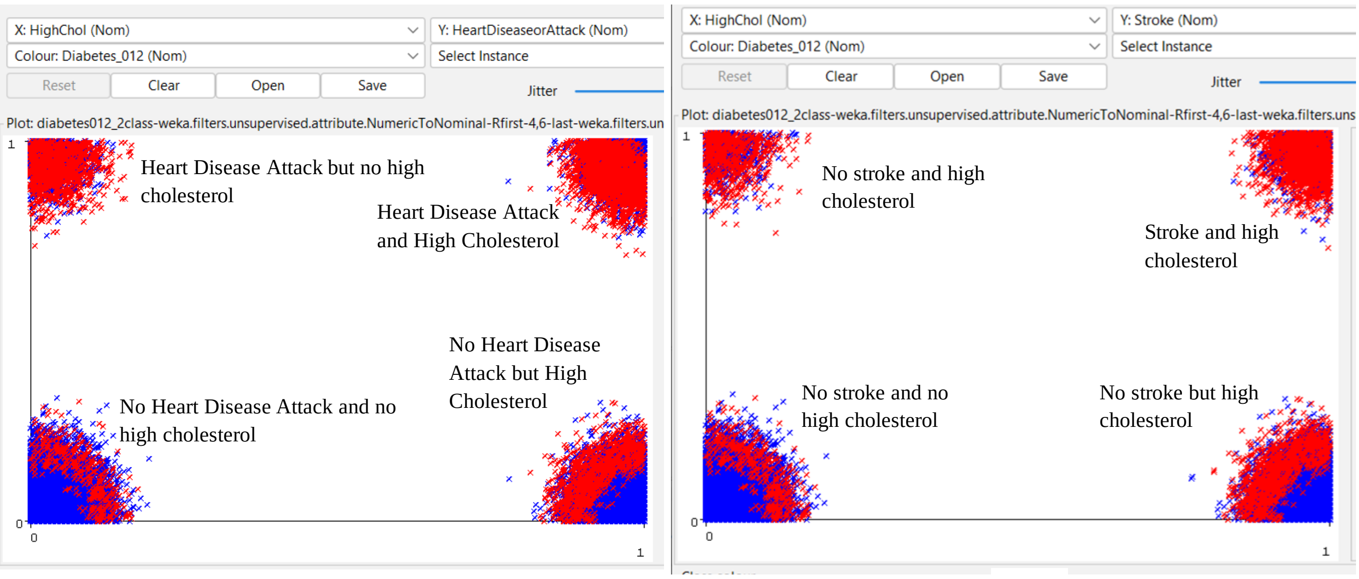
Attribute interaction

The interaction among attributes, referred to as multivariate analysis, investigates the interplay between variables to comprehend complex relationships and facilitate classification, with an emphasis on the distribution across many classes. The distributions of nominal and numeric data types differ, as nominal data is segmented into distinct intervals, resulting in data clustering around those intervals, but numeric data is continuous, leading to variability across the scales.

Figure 10(a) shows interaction among two nominal attributes of high cholesterol and heart disease attack. Meanwhile Figure 10(b) shows interaction among high cholesterol and another nominal attribute of stroke. Both figures show that diabetes patients indicated by red color are predominant on the upper level and less for no diabetes data. Consequently, this case highlights that high cholesterol, stroke incidents, and heart disease are significant contributors to the emergence of diabetes.

The interaction between 13 age categories and stroke incidence is depicted in Figure 11, emphasizing differences between diabetic and non-diabetic patients. This distribution indicates that no diabetes patients (shown in blue) dominate the no-stroke group; nevertheless, in the stroke group, the prevalence of diabetes patients increases from age level 6 to 13. According to Table 1, this level comprised a group of patients aged 54 and older than 80. The next Figure 12 illustrates the distribution of the data across numeric types of attribute of BMI in x-axis versus nominal data type of Diabetes_012 in y-axis which consist of no diabetes (label as 0) and diabetes (label as 2). The scatter distribution from the left to the right of the graph indicates a continuous data distribution for BMI, with a minimum of 12 and a maximum of 98. The majority of the sample data (patients) exhibit a BMI range of 12 to 55 for both diabetic and non-diabetic individuals. Therefore, there is no significant difference between these two attributes for further analysis.

Analyzing and comparing the interactions among attributes reveals that this is not a straightforward modeling problem, and there is no simple relationship among the attributes. Therefore, the tuning of the model requires careful consideration, employing advanced techniques like the ensemble method when classifying this dataset.



X: HighChol (Nom)
Y: HeartDiseaseorAttack (Nom)
Colour: Diabetes_012 (Nom)
Select Instance
Plot: diabetes012_2class-weka.filters.unsupervised.attribute.NumericToNominal-Rfirst-4,6-last-weka.filters.un

X: HighChol (Nom)
Y: Stroke (Nom)
Colour: Diabetes_012 (Nom)
Select Instance
Plot: diabetes012_2class-weka.filters.unsupervised.attribute.NumericToNominal-Rfirst-4,6-last-weka.filters.uns

Reset
Clear
Open
Save
Jitter
1
0
0
1
Heart Disease Attack but no high cholesterol
Heart Disease Attack and High Cholesterol
No Heart Disease Attack but High Cholesterol
No Heart Disease Attack and no high cholesterol
Reset
Clear
Open
Save
Jitter
1
0
0
1
No stroke and high cholesterol
Stroke and high cholesterol
No stroke and no high cholesterol
No stroke but high cholesterol

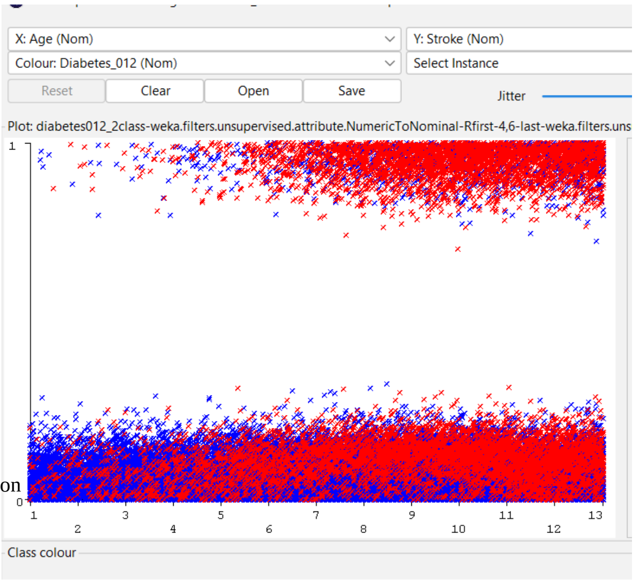
(a)

(b)

Figure 10: Distribution of data across three different attributes of high cholesterol, heart disease attack and

Has stroke and in variation of 13 interval of age

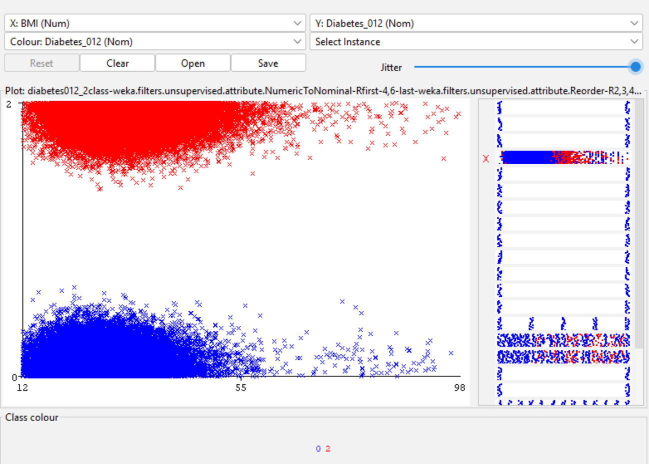
No stroke and in variation of 13 interval of age



X: Age (Nom)
Y: Stroke (Nom)
Colour: Diabetes_012 (Nom)
Select Instance
Plot: diabetes012_2class-weka.filters.unsupervised.attribute.NumericToNominal-Rfirst-4,6-last-weka.filters.uns

stroke

Figure 11: Distribution of data across attribute age vs stroke for diabetes and no diabetes patient



X: BMI (Num)

Y: Diabetes_012 (Nom)

Colour: Diabetes_012 (Nom)

Select Instance

Plot: diabetes012_2class-weka.filters.unsupervised.attribute.NumericToNominal-Rfirst-4,6-last-weka.filters.unsupervised.attribute.Reorder-R2,3,4...

Figure 12: Distribution of data across attribute BMI vs Diabetes_012

Conclusion

Originally this diabetes dataset consisted of 253680 samples of data known as observation and 22 attributes that might have a relation. The objective of the study which classification of diabetes patients with possible attributes that contribute to the disease. Thus, the target attribute selected was Diabetes_012 and the other 21 attributes remains as independent and predicted attributes. Statistical concepts play a crucial role in classification modeling, which is a fundamental technique in machine learning and data mining from the beginning of the data collected. Data distribution, which researchers require to review and understand the data, is essential for not only selecting appropriate classification algorithm but to make sure the data was accurately justified by the types and any possible constraint such as imbalance data. Diabetes list of predicted attributes originally dominance by nominal distribution except BMI. This justifies that there are several algorithms running well with nominal attributes for classification. The statistical analysis in this work elucidates the complex interplay among various attributes that necessitate advanced machine learning methods to differentiate insights between diabetic and non-diabetic patients. Additionally, there is a noteworthy connection among stroke, heart attack disease, and high cholesterol levels. Consequently, the future study will introduce a rigorous and robust machine learning algorithm to uncover predictive insights.

References:

Ahmad Adib Baihaqi, S., Syarifah Adilah, M. Y., Saiful Nizam, W., Mohd Saifulnizam, A. B., & Rozita, K. (2024). Machine Learning Approach of Predicting Airline Flight Delay using Naïve Bayes Algorithm. *Journal of Computing Research and Innovation Machine*, *9*(2), 140–155. https://doi.org/10.24191/jcrinn.v9i2.460

Current Trends & Future Scope of Data Mining. (2021, July 30). Datamation. https://www.datamation.com/big-data/data-mining-trends/

Liu, L., Wu, X., Li, S., Li, Y., Tan, S., & Bai, Y. (2022). Solving the class imbalance problem using ensemble algorithm: application of screening for aortic dissection. *BMC Medical Informatics and Decision Making*, *22*(1), 1–16. https://doi.org/10.1186/s12911-022-01821-w

Han, Jiawei, Kamber, Micheline, & Pei, Jian. (2011). *Data Mining Third Edition*. https://myweb.sabanciuniv.edu/rdehkharghani/files/2016/02/The-Morgan-Kaufmann-Series-in-Data-Management-Systems-Jiawei-Han-Micheline-Kamber-Jian-Pei-Data-Mining.-Concepts-and-Techniques-3rd-Edition-Morgan-Kaufmann-2011.pdf

Rushdi Shams. (2013, December 12). *Weka Tutorial 33: Random Undersampling (Class Imbalance Problem)*. YouTube. https://www.youtube.com/watch?v=ocOlm73HeNs

Teboul, A. (2022). *Diabetes Health Indicators Dataset*. Www.kaggle.com. https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset

Wongvorachan, T., He, S., & Bulut, O. (2023). A Comparison of Undersampling, Oversampling, and SMOTE Methods for Dealing with Imbalanced Classification in Educational Data Mining. *Information (Switzerland)*, *14*(1). https://doi.org/10.3390/info14010054