Numerical and relationship Analysis of Diabetic data set

B.Vishnu Prabha[1]
Artficial Intelligence and Data Science
Jeppiaar Institute of Technology
Tamilnadu, India
vishnuprabha.be@gmail.com

R.Elamathi[1]
Artficial Intelligence and Data Science
Jeppiaar Institute of Technology
Tamilnadu, India
elamathiramesh12@gmail.com

***Abstract*—Numerical analysis is very essential in analysing the data set in various fields. By performing numerical analysis the characteristics of data set is easily predicted. Now a days the diabetic is the most common disease in all over the world and its makes huge impact on everyone's lives. In this paper the analysis is made by taking the various characteristics of diabetic cases. The analysis of single, two and multiple characteristics are also known as univariate, Bivariate and multi variate analysis. In this analysis we predict the mean, median, standard deviation and various numerical summaries of diabetic data set and also the comparison between the various numerical summaries. The relationship analysis between features and target variables also performed. The results is very useful for the researchers to do their research on diabetic dataset.**

Keywords—Diabetic data set, Univariate, Bivariate, and Multivariate.

I. DATA ANALYSIS

Data analysis is a critical process that involves examining, transforming, and modeling data to extract meaningful insights that can inform decision-making. The process typically begins with data collection, followed by data cleaning and preparation to ensure the data is accurate and complete. Next, data is analyzed using a range of statistical and machine learning techniques, including descriptive, exploratory, inferential, predictive, and prescriptive analysis. Numerical and relationship analysis are important components of data analysis, as they help identify patterns and relationships in the data that can be used to make predictions or inform decisions. The results of data analysis are often presented using visualizations or reports, which can be used to communicate insights to stakeholders. Effective data analysis requires a combination of technical skills,

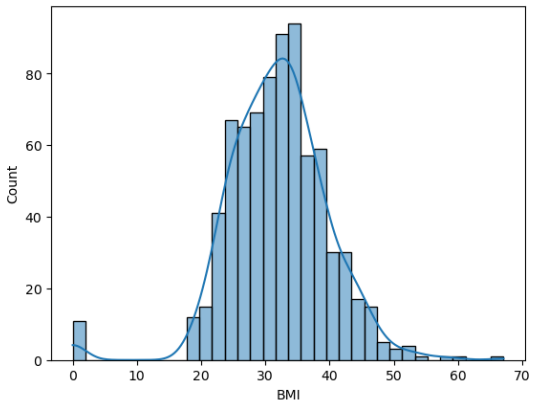

Fig. 1. Data analysis on diabetic dataset.

domain knowledge, and critical thinking, as well as an understanding of the business problem or research question being addressed [2][3].

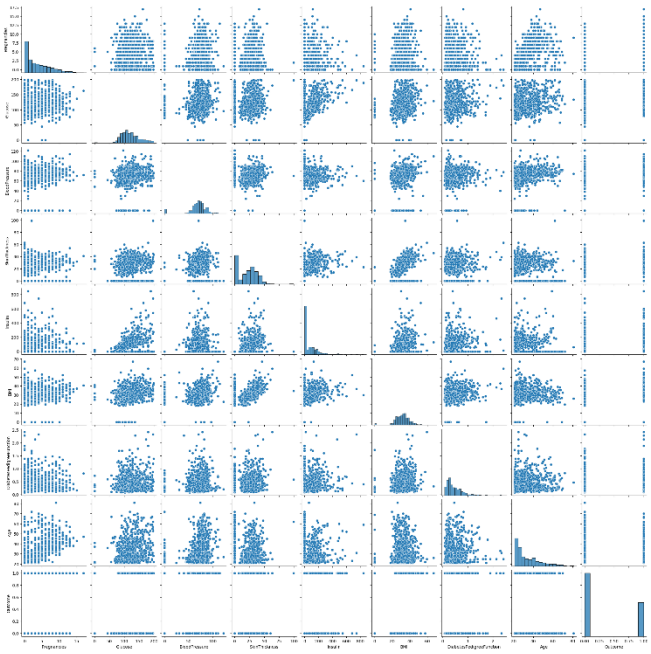

Fig. 2. Data analysis

II. TYPES OF DATA ANALYSIS

Data analysis is a broad field that encompasses a variety of ways and styles. There are several types of data analysis, including descriptive analysis, exploratory analysis, deducible analysis, prophetic analysis, conventional analysis, and individual analysis[4][7]. Descriptive analysis involves recapitulating and imaging data to describe its central tendency, variability, and distribution. Exploratory analysis involves exploring data to identify patterns and connections that may not be incontinently apparent. deducible analysis involves making consequences about a population grounded on a sample of data. Prophetic analysis involves using statistical or machine literacy models to make prognostications about unborn events grounded on literal data. conventional analysis involves using optimization and simulation ways to determine the stylish course of action given a set of constraints and objects. individual analysis involves relating the root cause of a problem or issue by

assaying data. Each type of data analysis has its strengths and sins and can be used in different surrounds depending on the exploration question and the nature of the data. Effective data analysis requires a combination of specialized chops, sphere knowledge, and critical thinking to insure that the perceptivity gained from the data are accurate and practicable.

III. NUMERICAL ANALYSIS

The Numerical analysis is an essential component of data analysis that involves analyzing numerical data to identify patterns and trends. This type of analysis typically begins with calculating summary statistics, such as means, medians, and standard deviations, to describe the central tendency, variability, and distribution of the data. Numerical analysis can also involve conducting tests of statistical significance, such as t-tests or ANOVA, to determine whether observed differences between groups are likely due to chance or are statistically significant[1][8]. This type of analysis is particularly useful when analyzing data that is quantitative in nature, such as measurements or counts. By analyzing numerical data, researchers can gain insights into patterns and trends that may not be immediately apparent, and use this information to inform decision-making or develop predictive models. Numerical analysis can be conducted using a variety of software tools, including Excel, Python, R, and MATLAB, and requires a strong foundation in statistical methods and data analysis techniques.

	Pregnanci	Glucose	BloodPres	SkinThickı	Insulin	BMI	DiabetesF	Age	Outcome
count	768	768	768	768	768	768	768	768	768
mean	3.845052	120.8945	69.10547	20.53646	79.79948	31.99258	0.471876	33.24089	0.348958
std	3.369578	31.97262	19.35581	15.95222	115.244	7.88416	0.331329	11.76023	0.476951
min	0	0	0	0	0	0	0.078	21	0
25%	1	99	62	0	0	27.3	0.24375	24	0
50%	3	117	72	23	30.5	32	0.3725	29	0
75%	6	140.25	80	32	127.25	36.6	0.62625	41	1
max	17	199	122	99	846	67.1	2.42	81	1

Table. 1. Statistical Summaries of the dataset

Numerical analysis is a complex and dynamic field that encompasses a wide range of ways and styles. In addition to mean, standard, mode, and standard divagation, there are several other important statistical measures that can be used to dissect numerical data[3][9]. One similar measure is friction, which is the normal of the squared differences between each data point and the mean. friction can give sapience into the spread of the data and is frequently used in confluence with standard divagation. Retrogression analysis is another important fashion used in numerical analysis that involves modeling the relationship between a dependent variable and one or further independent variables[9]. Retrogression analysis can be used to make prognostications or identify significant predictors in a dataset.

Time series analysis is another important fashion used in numerical analysis that involves assaying data over time. Time series analysis can help identify patterns or trends in data and can be used to read unborn values or events. numerical analysis is a complex and multifaceted field that requires a deep understanding of statistical proposition and styles. In addition to mean, standard, mode, and standard divagation, there are several other important statistical measures used in numerical analysis, including friction, correlation, retrogression analysis, and time series analysis. By using these ways, experimenters can gain precious perceptivity into numerical data and make informed opinions grounded on the results.

A. MEAN

The mean is the normal of a set of numerical values and is calculated by dividing the sum of the values by the total number of values[3][8]. The mean is a useful statistic when dealing with generally distributed data, but it can be affected by outliers or extreme values.

```
diabetes_ds['BMI'].mean()
```
```
31.992578124999977
```
```
diabetes_ds['Age'].mean()
```
```
33.240885416666664
```
```
diabetes_ds['DiabetesPedigreeFunction'].mean()
```
```
0.4718763020833327
```
```
diabetes_ds['BloodPressure'].mean()
```
```
69.10546875
```
```
diabetes_ds['Pregnancies'].mean()
```
```
3.8450520833333335
```

Fig. 3. Mean of the some features in dataset

B. MEDIAN

The standard is the middle value in a dataset when the values are arranged in order[9][10]. The standard is a robust statistic, which means it is not affected by outliers or extreme values. It's particularly useful when dealing with inclined ornon-typically distributed data.

```
diabetes_ds['BMI'].median()
```
```
32.0
```
```
diabetes_ds['Age'].median()
```
```
29.0
```
```
diabetes_ds['DiabetesPedigreeFunction'].median()
```
```
0.3725
```
```
diabetes_ds['BloodPressure'].median()
```
```
72.0
```
```
diabetes_ds['Pregnancies'].median()
```
```
3.0
```

Fig. 4. Median of the some features in dataset

C. MODE

The mode is the value that occurs utmost constantly in a dataset. The mode is a useful statistic when dealing with categorical or nominal data, but it may not be useful when dealing with continuous numerical data.

D. STANDARD DEVIATION

Standard divagation is a measure of the variability or spread of a dataset. A low standard divagation indicates that the data points are close to the mean, while a high standard divagation indicates that the data points are more spread out[8]. Standard divagation is particularly useful for comparing datasets with different means or for detecting outliers or extreme values.

Numerical analysis is a fundamental fashion used in data analysis to prize useful perceptivity from numerical data. Mean, standard, mode, and standard divagation are important summary statistics used in numerical analysis to describe the central tendency, variability, and distribution of the data.

IV. RELATIONSHIP ANALYSIS

Relationship analysis is a fundamental technique used in data analysis to identify patterns and associations between variables[4]. This type of analysis involves examining the correlation between two or more variables to determine the strength and direction of their relationship.

The most common method used for relationship analysis is correlation analysis, which measures the linear relationship between two variables. Correlation analysis produces a correlation coefficient that ranges from -1 to +1.

There are several other methods used for relationship analysis, including regression analysis, which models the relationship between a dependent variable and one or more independent variables[8][9]. Regression analysis can be used to predict values for the dependent variable based on the values of the independent variables.

Other methods include factor analysis, which identifies underlying factors that explain the correlation between a set of variables, and cluster analysis, which groups similar observations together based on their characteristics.

Effective relationship analysis requires a deep understanding of statistical theory and methods, as well as the ability to interpret and communicate results to stakeholders[2]. By identifying patterns and relationships between variables, researchers can gain valuable insights into the data and use this information to make informed decisions or develop predictive models.

Relationship analysis is a powerful technique that can be used in a wide range of fields, including finance, marketing, and healthcare, to name a few.

Fig. 5. Heat map on diabetic dataset.

V. QUANTITATIVE ANALYSIS

Quantitative analysis is a type of data analysis method that involves the use of numerical and statistical techniques to measure and quantify relationships between variables. It is often used in scientific research and other fields where data is collected through structured and objective methods. The goal of quantitative analysis is to provide a systematic and objective approach to data analysis, which can provide valuable insights into patterns, trends, and relationships in the data[2][3]. It involves several steps, including data collection, data cleaning, data transformation, statistical modeling, and interpretation of results. In data collection, researchers collect data using a structured approach, often through surveys or experiments. In data cleaning, researchers remove errors and inconsistencies in the data to ensure that the data is accurate and reliable. Data transformation involves converting the data into a suitable format for statistical analysis.

Statistical modeling is a critical step in quantitative analysis, which involves applying statistical methods to analyze the data and test hypotheses[10]. This includes techniques such as regression analysis, correlation analysis, and hypothesis testing. Finally, interpretation of results involves understanding the meaning of the statistical findings and drawing conclusions based on the analysis. Univariate, bivariate, and multivariate analysis are types of data analysis which comes under quantitative analysis methods used to analyze data based on the number of variables involved.

A. UNIVARIATE ANALYSIS

Univariate analysis involves analyzing a single variable at a time. This type of analysis is useful for describing the distribution of a single variable, such as calculating summary statistics like mean, median, and standard deviation.

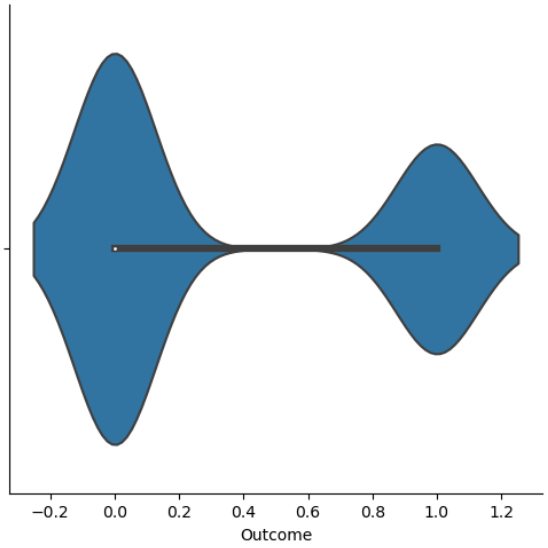

Fig. 6. Univariate analysis on diabetic dataset.

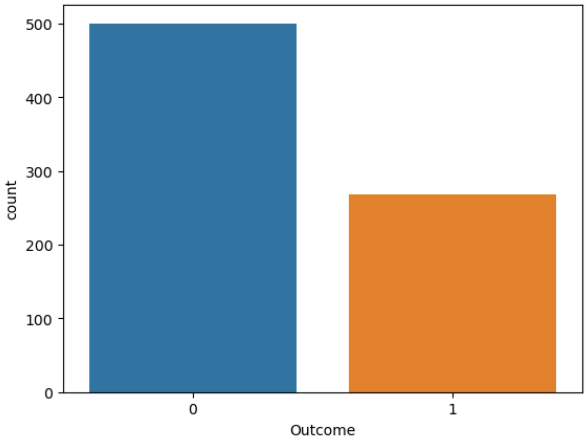

Fig. 7. Univariate analysis on diabetic dataset.

B. *BIVARIATE ANALYSIS*

Bivariate analysis involves analyzing the relationship between two variables[8]. This type of analysis is useful for investigating how one variable affects another and can be used to identify patterns or associations between variables. It often involves calculating correlation coefficients or regression models to investigate the relationship between variables.

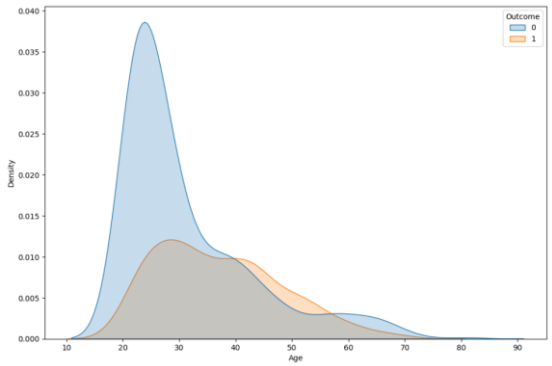

Fig. 8. Bivariate analysis on diabetic dataset.

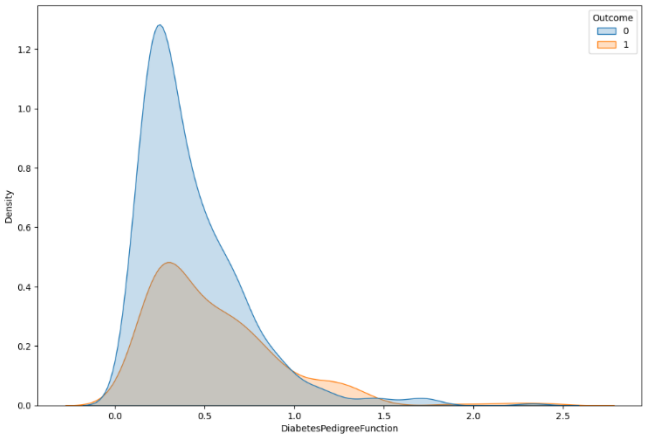

Fig. 9. Bivariate analysis on diabetic dataset.

C. *MULTIVARIATE ANALYSIS*

Multivariate analysis involves analyzing three or more variables at a time. This type of analysis is useful for investigating complex relationships between variables and identifying patterns that may not be apparent in bivariate analysis.

Multivariate analysis often involves techniques like factor analysis, cluster analysis, or structural equation modeling to identify underlying factors or relationships between variables.

Multivariate analysis involves several techniques, including factor analysis, cluster analysis, discriminant analysis, and structural equation modeling[1][9]. These techniques can be used to identify the underlying factors that contribute to the variation in the data and to explore the relationships between variables. Factor analysis is a multivariate technique that is used to identify the underlying factors that explain the variation in the data.

It involves reducing the number of variables in the data set to a smaller set of underlying factors that capture the essential information in the data[6][7]. Cluster analysis is another multivariate technique that is used to group similar observations in the data set. It involves identifying clusters of similar data points and grouping them together based on their similarities.

Discriminant analysis is a statistical technique used to identify the factors that differentiate two or more groups. It involves identifying the variables that are most significant in predicting the group membership[5]. Finally, structural equation modeling is a multivariate statistical technique used to analyze complex relationships between variables.

It involves constructing a model that represents the relationships between variables and testing the model's fit to the data.

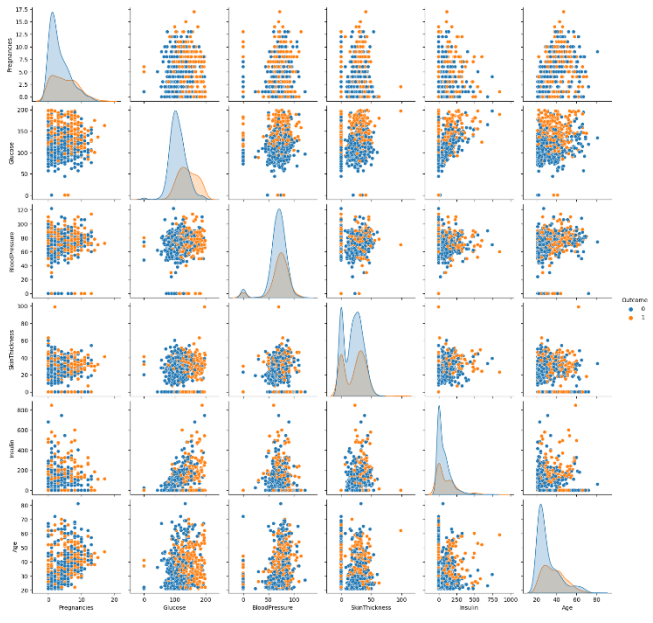

Fig. 1. Multivariate analysis on diabetic dataset.

VI. CONCLUSION

In conclusion, this journal presents an analysis of a diabetic dataset using various data analysis techniques. The study involved the application of univariate, bivariate, and multivariate analysis methods, along with numerical analysis, to understand the dataset's characteristics. The analysis identified key variables and relationships between them, providing insights into the underlying patterns and trends in the data. The univariate analysis revealed the distribution and summary statistics of individual variables, including measures of central tendency and variability. The bivariate analysis examined the relationship between two variables, while the multivariate analysis extended the analysis to three or more variables. The results of the analysis highlighted the importance of different variables in predicting the outcome and provided a basis for developing a predictive model. Overall, the analysis of the diabetic dataset demonstrates the importance of data analysis in gaining insights into complex data sets. By using various data analysis techniques, researchers can identify patterns and trends that may not be apparent in simple analysis. This information can be used to develop targeted interventions and improve patient outcomes. Furthermore, the study highlights the value of numerical analysis, which provides a basis for statistical modeling and decision-making.

.

REFERENCES

[1] M. Prakash, G. Padmapriy and M. V. Kumar, "A Review on Machine Learning Big Data using R", Second International Conference on Inventive Communication and Computational Technologies (ICICCT), pp. 1873-1877, 2018.

[2] T. D. Chung, R. Ibrahim, S. M. Hassan and N. S. Rosli, "Fast approach for automatic data retrieval using R programming language", 2nd IEEE International Symposium on Robotics and Manufacturing Automation (ROMA), pp. 1-4, 2016B. Rieder, *Engines of Order: A Mechanology of Algorithmic Techniques*. Amsterdam, Netherlands: Amsterdam Univ. Press, 2020.

[3] J. E. Cavanaugh and A. A. Neath, Akaike's Information Criterion: Background Derivation Properties and Refinements, Berlin, Heidelberg:Springer Berlin Heidelberg, pp. 26-29, 2011.

[4] . E. Kleiger, J. T. Bigger, M. S. Bosner, M. K. Chung, J. R.Cook, L. M. Rolnitzky, R. Steinman, J. L. Fleiss, Stability overtime of variables measuring heart rate variability in normal sub-jects, American Journal of Cardiology 68 (6) (1991) 626–630.

[5] U. R. Acharya, K. S. Vidya, D. N. Ghista, W. J. E. Lim, F. Moli-nari, M. Sankaranarayanan, Computer-aided diagnosis of di-abetic subjects by heart rate variability signals using discretewavelet transform method, Knowledge-based systems 81 (2015)56–64.

[6] Diabetes Prediction System based on Iridology using Machine Learning, IEEE Conference, Semarang, Indonesia, pages 1-9, 2011.

[7] Image augmentation methods by slideShare https://wwwpresentation-from-xperi, Accessed 18 May 2009.

[8] Diagnosis of Diabetes using Computer Methods: Soft Computing methods for Diabetes detection using Iris, World Academy of Science, Engineering and Technology Conference, pages 1-6, 2019.

[9] Christopher picture publications, https:// www. christopherpublications.com/Iridology.html, Accessed 13 Mar. 2011.

[10] Science direct research, https:// www. sciencedirect.com/science/article/abs/pii/S0045790618302556, Accessed 8 January, 2011.