PES UNIVERSITY

Data Analytics- EC campus

Section: F,G,I and J

Format for Literature Survey Report

1.Project Title	HEART DISEASE ANALYSIS USING MACHINE LEARNING ALGORITHMS.	
2.Team Name	Coconut water	
3.Team Members	SRN1: PES2UG20CS448	Name1: POOJA H
	SRN2: PES2UG20CS420	Name2: AYUSH DUDHE
	SRN3: PES2UG20CS453	Name3: RIA TREZA SERRAO
4.Dataset used	heart.csv	
5.Link for the Dataset	https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset	
6.Github link	https://github.com/ayushdudhe/heartdiesaseprediction	

7.Problem Statement:
This project covers manual exploratory data analysis and using pandas in Jupyter Notebook. Questions: 1. Import The Libraries And Dataset 2. Display Top 5 Rows of The Dataset 3. Check The Last 5 Rows of The Dataset 4. Find Shape of Our Dataset (Number of Rows And Number of Columns) 5. Get Information About Our Dataset Like Total Number Rows, Total Number of Columns, Datatypes of Each Column And Memory Requirement 6. Check Null Values In The Dataset 7. Check For Duplicate Data and Drop Them 8. Get Overall Statistics About The Dataset 9. Draw Correlation Matrix 10. How Many People Have Heart Disease, And How Many Don't Have Heart Disease In This Dataset? 11. Find Count of Male & Female in this Dataset 12. Find Gender Distribution According to The Target Variable 13. Check Age Distribution In The Dataset 14. Check Chest Pain Type 15. Show The Chest Pain Distribution As Per Target Variable 16. Show Fasting Blood Sugar Distribution According To Target Variable 17. Check Resting Blood Pressure Distribution 18. Compare Resting Blood Pressure As Per Sex Column 19. Show Distribution of Serum cholesterol 20. Plot Continuous Variables

8.EDA and Visualization

EDA : As at this preprocessing level, this is the important step; meaningful data is derived from the dataset of heart disease. This phase is compulsory because the raw data is not reliable and unfinished, so pre-processing is performed for more steps to render ready raw data. In this approach, the UCI Heart Disease Dataset, the data contains 75 attributes, and during pre-processing, 14 [6] attributes are extracted to understand the nature of patients' health better. The extricated 14 attributes include BP, sex, heart rate, chest, and others. The attribute's values are normalized and converted into numerical form. The quality of data plays an essential role, and the most carefully depicted thing to be. For this research, data cleaning has improved the quality of our dataset. Data cleaning is necessary as it removes unnecessary or irrelevant attributes of data from the dataset. This step of the model will make the dataset more precise and exact. In this part of approach, the Null (NaN) values are removed from the dataset to make it more useful as these values decrease the productivity of the algorithm [7]. At the data cleaning stage, the dataset is also normalized to not have any ambiguity after cleaning

VISUALIZATION: The dataset is in tabular form and it is hard to observe and understand the data in this or any other form. So the data is visualized Graphically below. It helps in knowing the trend of the data. Data visualization in this approach is a graphical representation of the data. In this analysis, using bar charts and scatter plots, the cleaned data acquired by pre-processing is visualized. It illustrates the actions of data attributes. It makes it easy to grasp the attribute's complicated relationship by graphical representation.

9. Summarize the Literature survey

Majority of researchers have used the Cleveland Heart Disease Dataset available from the UCI repository containing 76 attributes and 303 instances, of which only 14 attributes are used due to missing values [11]. There are huge benefits to having feature selection methods so as to minimize the number of attributes that one has to use in order to build an accurate model by checking the correlation between various attributes and their impact on the accuracy of the models. It can be seen from various research papers in the field that KNN and Neural Network works quite accurately in most cases for the prediction of heart diseases.

10. What is the specific problem your team is going to solve?

1. We did data visualization and data analysis of the target variable, age features, and whatnot along with its univariate analysis and bivariate analysis.

2. We also did a complete feature engineering part in this article which summons all the valid steps needed for further steps i.e.

model building.

11. References

1. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.302.2431&rep=rep1&type=pdf
2. A, A. S., & Naik, C. (2016). Different Data Mining Approaches for Predicting Heart Disease, 277–281. https://doi.org/10.15680/IJIRSET.2016.0505545
3. https://iopscience.iop.org/article/10.1088/1757-899X/1022/1/012072/pdf
4. Kiyasu J Y (1982). U.S. Patent No. 4,338,396. Washington, DC: U.S. Patent and Trademark Office
5. E.Taylor,P.s.Ezekiel,F.B.Deedam. (2019). “A Model to Detect Heart Disease using Machine Learning algorithm” International journal of Computer Science and engineering.vol-7,issue-11
6. R. Goel and A. Jain. (2018) "The Implementation of Image Enhancement Techniques on Color n Gray Scale IMAGEs," 2018 Fifth International Conference on Parallel, Distributed and Grid Computing (PDGC), , pp. 204-209, doi: 10.1109/PDGC.2018.8745782