PES UNIVERSITY

Data Analytics- EC campus

Section: F,G,I and J

Format for Literature Survey Report

1.Project Title	HEART DISEASE ANALYSIS	
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 Exploratory Data Analysis **:** As this preprocessing level helps us to understand the dataset. 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.

VISUALIZATION: The dataset is in tabular form and it is hard to observe and understand the data. 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

The algorithms that we used are more accurate, saves a lot of money i.e. it is cost efficient and faster than the algorithms that the previous researchers used.
Moreover, the maximum accuracy obtained by KNN and Logistic Regression are equal to 94.5% which is greater from previous researches.
So, we summarize that our accuracy is improved due to the increased medical attributes that we used from the dataset we took.
Our project also tells us that Logistic Regression and KNN outperforms Random Forest Classifier in the prediction of the patient diagnosed with a heart Disease.
This proves that KNN and Logistic Regression are better in diagnosis of a heart disease.

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

A very detailed, useful, and highly preferable Machine Learning based model in this paper that helps medical practitioners diagnose heart diseases at an early stage to enable patients to take precautionary measures in a rectification
Heart Disease prediction system assists a patient if he/she is diagnosed with cardiovascular disease and helps them prevent in the earlier stages with the accuracy of our model is 87.5%

11. References

- https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.302.2431&rep=rep1&type=pdf
- A, A. S., & Naik, C. (2016). Different Data Mining Approaches for Predicting Heart Disease, 277–281. https://doi.org/10.15680/IJIRSET.2016.0505545
- https://iopscience.iop.org/article/10.1088/1757-899X/1022/1/012072/pdf
- Kiyasu J Y (1982). U.S. Patent No. 4,338,396. Washington, DC: U.S. Patent and Trademark Office
- 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
- 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
- Archana Singh, Rakesh k. (2020). ”Heart disease Prediction Using machine Learning Algorithms” International Conferences On Electrical and electronics Engineering(ICE3)
- Shinde R, Arjun S, Patil P & Waghmare J (2015). An intelligent heart disease prediction system using k-means clustering and Naïve Bayes algorithm. International Journal of Computer Science and Information Technologies, 6(1), 637-9.
- Bashir S, Qamar U & Javed M Y (2014, November). An ensemble-based decision support framework for intelligent heart disease diagnosis. In International Conference on Information Society (i-Society 2014) (pp. 259-64). IEEE
- Jee S H, Jang Y, Oh D J, Oh B H, Lee S H, Park S W & Yun Y D (2014). A coronary heart disease prediction model: the Korean Heart Study. BMJ open, 4(5), e005025.