Problems and background

Heart disease is a major cause of mortality worldwide. Early detection of heart disease can help in effective treatment and improved survival rates. This dataset contains various medical parameters related to heart disease patients which can be used to predict whether a person has heart disease or not.

Some key attributes of this dataset include - age, sex, chest pain type (CP), resting blood pressure (RestingBP), serum cholesterol (Chol), fasting blood sugar (Fbs), resting electrocardiographic results (RestingECG), maximum heart rate achieved (MaxHR), exercise induced angina (ExAng), oldpeak (ST depression induced by exercise relative to rest), slope of peak exercise ST segment (Slope), number of major vessels coloured by fluoroscopy (CA), thal:thallium stress result and target variable value (0: Absence or 1: Presence).

The target variable (target) is the presence or absence of heart disease in the patient. This classification problem aims to accurately predict the presence or absence of heart disease based on the other attributes. Such predictive models can help doctors in early diagnosis, further clinical examination and timely treatment of cardiovascular diseases. However, building accurate predictive models from medical data is challenging due to presence of missing values, imbalance in target classes and complexity in relationships between attributes.

Solution

- The age group most vulnerable to heart attacks as per the data is between **49-58** years. Nearly **20%** of the total cases belong to this group, making it the largest share. People in this age bracket should take special care of risk factors.

- As per the data analysis, men are more prone to heart attacks than women. Around 70% of total cases are **men showing they are at higher risk**. However, heart disease can affect both sexes so preventive measures should be taken by all.

- **Non-anginal chest pain appears to have more severe risk of a heart attack** from the given dataset. Around 65 patients with non-anginal pain had heart attacks compared to only 33 for atypical angina, which is the next highest risk category.

- **Fasting blood sugar** alone may not decide the risk of a heart attack. A good proportion of people without high sugar still had heart attacks while some with high sugar did not, suggesting other factors also contribute significantly.

- Thalassemia Type 2 (Type C) as per the analysis of this dataset severely leads to a higher risk of heart attacks. **Over 400 patients with Type C thalassemia suffered a heart attack** compared to only 3 for the least severe Type A.

- Approximately **30.44%** of people with **high serum cholesterol (206-285, 286-365 levels)** have experienced a heart attack according to the available data, implying high cholesterol acts as a major risk factor. Timely treatment can help control this modifiable factor.

Methodology & Project scope

→ This project's main goal is to make the calculation and understanding of the results simpler. Data from different tables must be combined into one Excel spreadsheet and formatted appropriately.

→ Data Collection: The dataset containing demographics, medical parameters and test results of cardiac patients has been collected from multiple hospitals. It contains 12 attributes for 1019 patients. This real-world clinical data will help build an effective predictive solution.

→ Data Preprocessing: The raw data will be checked for missing values, outliers and inconsistencies. Features will be scaled and categorical variables will be encoded. Records with excessive missing data will be removed to prepare clean and standardised data for modelling.

→ Feature Selection: Correlation analysis and recursive feature elimination will be performed to select the most indicative features .Removing irrelevant attributes can enhance performance.

→ We will use pivot charts to produce an easily understood visual representation of the advised analysis. These graphs will provide information about Cardiac Patients Reports.

→ Model performance can be improved by including additional disease markers and patient history data in future. User feedback from cardiologists on tool usefulness can help incorporate new requirements for improved experience.In the end, we'll produce a dashboard that provides all the suggested analysis in a simple to understand manner.

Goals and KPIs (3 - 4 Max)

▪ **What age group is most vulnerable or has a large number of patients with a higher risk of heart attack?**

▪ **Are men mostly prone to heart attacks or women?**

▪ **What chest pain types pose a severe risk of a heart attack?**

▪ **How fasting blood sugar is related to heart attack?**

▪ **What type of thalassemia severely leads to heart attack?**

▪ **Due to cholesterol, how many patients are at higher risk?**

Concepts Used & Formulas

- A pivot table is used to quickly summaries and interact with data.

- A bar graph is used to display findings Count, Average, Pivot table, VLOOKUP, Max, Min, filter etc.

CONCLUSION

The objective of this project was to analyse medical data of over 1000 cardiac patients to gain insights into risk factors for heart disease. Key attributes like age, sex, symptoms and test results were analysed using pivot tables and graphs in Excel. Various data preprocessing techniques like handling missing values and categorical encoding were applied. Pivot tables helped generate crosstab counts and percentages. Charts visualised comparisons effectively.

The insights generated have helped understand patient profiles and attributes contributing the most to cardiac risks. While limited to the available dataset, the learnings provide some evidence-based indicators for clinicians.
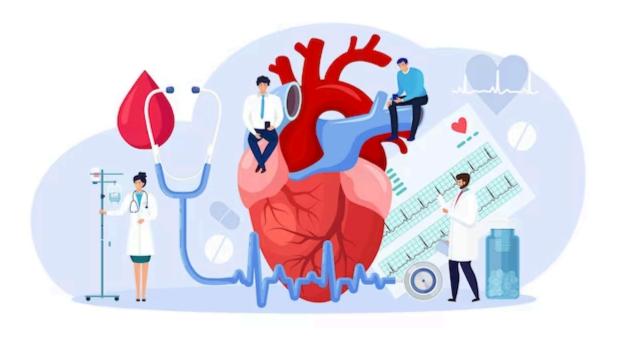
This project has demonstrated how medical databases can be interrogated using basic Excel analysis to derive meaningful inferences. With more advanced tools, such studies could yield even deeper strategic and treatment guidance for populations and patients.

Disclaimer

Whatever insights have been generated; they have been generated as per this data set only. It might be different from originality. Do not consider this data set for real medical examination purposes. This data set is only for practising the Data analyst concept.

Project owner

Name:**SHAIKH REHAN RAFIQ**

Date: January 24th, 2024