Data Quality Report

1. Introduction

This report provides an analysis of the data used to predict cardiovascular disease. The focus of this report is to assess the quality of the dataset by examining its accuracy, reliability, validity, redundancy, and other indicators of good quality data.

2. Data Description

The dataset contains medical records for individuals, with a set of features aimed at predicting the likelihood of cardiovascular disease. Key features include:

1. age - age in years
2. sex - (1 = male; 0 = female)
3. cp - chest pain type
    - 0: Typical angina: chest pain related to decreased blood supply to the heart
    - 1: Atypical angina: chest pain not related to heart
    - 2: Non-anginal pain: typically esophageal spasms (non-heart-related)
    - 3: Asymptomatic: chest pain not showing signs of disease
4. trestbps - resting blood pressure (in mm Hg on admission to the hospital) anything above 130-140 is typically cause for concern
5. chol - serum cholesterol in mg/dl
    - serum = LDL + HDL + .2 * triglycerides
    - above 200 is cause for concern
6. fbs - (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
    - '>126' mg/dL signals diabetes
7. restecg - resting electrocardiographic results
    - 0: Nothing to note
    - 1: ST-T Wave abnormality
        - can range from mild symptoms to severe problems
        - signals non-normal heart beat
    - 2: Possible or definite left ventricular hypertrophy
        - Enlarged heart's main pumping chamber
8. thalach - maximum heart rate achieved
9. exang - exercise-induced angina (1 = yes; 0 = no)
10. oldpeak—ST depression induced by exercise relative to rest looks at stress of heart during exercise; an unhealthy heart will stress more
11. slope - the slope of the peak exercise ST segment
    - 0: Upsloping: better heart rate with exercise (uncommon)
    - 1: Flatsloping: minimal change (typical healthy heart)
    - 2: Downsloping: signs of unhealthy heart
12. ca - number of major vessels (0-3) colored by fluoroscopy
    - colored vessel means the doctor can see the blood passing through
    - the more blood movement, the better (no clots)

13. thal - thallium stress result
    - 1,3: normal
    - 6: fixed defect: used to be defect but ok now
    - 7: reversible defect: no proper blood movement when exercising
14. target - have disease or not (1=yes, 0=no) (= the predicted attribute)

Data Accuracy refers to the degree to which the values in the dataset are correct and free from error. The dataset was reviewed for inconsistencies:

- **Outliers**: Upon visualising the data, certain features like systolic and diastolic blood pressure displayed extreme values. For example, blood pressure values greater than 300 mmHg are physiologically unlikely. These entries may need further verification or removal to ensure accurate model predictions.
- **Categorical Accuracy**: Categorical data, such as gender, appears to be consistently labelled with no obvious misclassification.
- **Measurement Precision**: Features such as age are recorded with reasonable precision (integer or up to one decimal place), suggesting the data captures fine details required for accurate prediction.

4. Reliability

Data Reliability assesses the consistency of the dataset over time or across different measurements.

- **Consistency of Measurements**: Features like blood pressure and glucose levels are standard medical measurements, which suggests that the dataset likely has high reliability, as these measurements are taken in controlled conditions using standard methods.
- **No Temporal Data**: The dataset does not contain time-series data, which means we cannot assess temporal consistency or track the evolution of certain health conditions over time. However, the lack of this information does not directly impact the reliability of the snapshot data.

5. Validity

Data Validity refers to the degree to which the dataset measures what it is intended to measure.

- **Internal Validity**: The dataset contains features that are well-established risk factors for cardiovascular disease, such as age, blood pressure, cholesterol, stress, and BMI, which enhances the internal validity of the data. These are medically verified predictors, making the dataset valid for its purpose.
- **Face Validity**: The dataset appears to cover all necessary and relevant variables that one would expect for predicting cardiovascular disease, contributing to its overall validity.

For example, features like smoking, cholesterol levels, and physical activity align with established medical research on cardiovascular health.

6. Redundancy

Data Redundancy refers to unnecessary duplication or overlap in the data.

- **Feature Redundancy**: The dataset does not appear to have redundant features. All features included provide unique information relevant to cardiovascular health. However, features like systolic and diastolic blood pressure may be highly correlated, though both are medically significant and justified in their inclusion.
- **Data Duplication**: There were no duplicate entries observed during data exploration, indicating the dataset has no redundancy in patient records.

7. Completeness

Data Completeness reflects the presence or absence of missing data:

- **Missing Data**: The dataset showed no missing values across any of the key features. This indicates that the data collection process was thorough and complete, reducing the need for imputation or deletion of records.
- **Coverage**: The dataset appears to cover a wide range of values for most features, including age, blood pressure, and cholestrol, making it suitable for generalisation across different patient demographics.

8. Uniformity

Data Uniformity examines whether data is presented in a consistent format:

- **Standardised Measurement Units**: Features like blood pressure (mmHg) and BMI are recorded in standard, globally accepted units. This uniformity ensures that models can interpret the data correctly without requiring conversions or additional preprocessing.
- **Encoding Consistency**: Categorical variables such as gender, cholesterol, and glucose levels are consistently encoded throughout the dataset, ensuring that machine learning algorithms will handle them correctly.

9. Data Distribution

Data Distribution analysis provides insight into the balance and spread of data:

- **Class Distribution**: The target variable (presence of cardiovascular disease) appears to be balanced between the two classes, meaning that there is no significant class imbalance. This is important for building reliable models, as imbalanced datasets can lead to biased predictions.

- **Feature Distributions**: Continuous variables like age are distributed with a slight skew, indicating that most patients in the dataset fall into middle-aged and older age groups, which aligns with the typical population at risk for cardiovascular disease.

10. Conclusion

The dataset used for cardiovascular disease prediction is of high quality, with strong indicators of accuracy, reliability, and validity. There is no redundancy or missing data, ensuring a comprehensive and consistent dataset for model training. However, care should be taken when dealing with potential outliers, especially in blood pressure measurements, to avoid skewing model predictions. Additionally, some categorical encodings (e.g., gender) may benefit from finer granularity to increase the predictive power of models.

This dataset is well-suited for machine learning tasks aimed at predicting cardiovascular disease and is expected to yield reliable and valid predictions when used appropriately.