Heart Disease Prediction Project: Data Analysis and Modeling Phases

ALAOUI Nouhaila & FRIDI Fatima Zahra
1st Master ISI
SUP MTI

***Abstract*—This paper presents a structured approach to a heart disease prediction project, following eight phases of the data science pipeline: data collection, problem understanding, data understanding, data preparation, exploratory data analysis, modeling, model evaluation, and interpretation. Each phase is detailed with objectives, methodology, and insights, providing a comprehensive understanding of the project workflow.**

***Index Terms*—Heart Disease, Machine Learning, Data Analysis, Classification, Data Preparation, Feature Engineering**

I. Introduction

Heart disease remains a leading cause of mortality worldwide. Predicting its occurrence using machine learning can aid early diagnosis and improve patient care. This paper presents the systematic approach followed to build a predictive model for heart disease risk. The methodology is divided into eight phases covering the complete data science pipeline, from data collection to interpretation of results.

II. Collecte de Données

A. Objective

Select a standard, high-quality dataset suitable for analysis and modeling, ensuring completeness and reliability.

B. Methodology

The dataset was sourced from Kaggle, containing patient demographic and clinical metrics, including age, sex, blood pressure, cholesterol levels, and heart disease diagnosis. Data quality checks were performed to:

- Verify completeness and absence of missing values.
- Remove duplicates.
- Ensure logical consistency of features.
- Optionally, explore combining multiple sources or using APIs to enrich the dataset.

C. Insights

The dataset was clean, complete, and ready for analysis, making it suitable for predictive modeling.

III. Compréhension du Problème

A. Objective

Define the business and clinical context, identify the target variable, and determine the type of machine learning problem.

B. Methodology

- Context: Support clinical decisions and early detection of heart disease.
- Objective: Predict the presence or absence of heart disease in patients.
- Problem type: Binary classification (categorical target variable: heart disease present or absent).

C. Insights

Understanding the problem clarified the project goals and guided the selection of appropriate models and evaluation metrics.

IV. Compréhension des Données

A. Objective

Gain familiarity with dataset characteristics and identify potential issues.

B. Methodology

- Describe dataset dimensions, variables, and types.
- Detect missing values and duplicates.
- Identify outliers and assess target variable balance.

C. Insights

The dataset contained no missing values or duplicates. Outliers were present in some numeric features, and the target variable was relatively balanced.

V. Préparation des Données

A. Objective

Transform raw data into a structured format suitable for modeling.

B. Methodology

- Encode categorical variables using one-hot or label encoding.
- Normalize or standardize numeric variables to a comparable scale.
- Split data into training and testing sets.
- Remove duplicates and correct any inconsistencies.

C. Insights

Data was fully prepared, transformed, and partitioned, ensuring models could learn effectively.

VI. Analyse Exploratoire des Données (EDA)

A. Objective

Explore dataset patterns, correlations, and distributions to inform modeling decisions.

B. Methodology

- Compute descriptive statistics for each variable.
- Visualize distributions, outliers, and relationships using histograms, boxplots, scatterplots, and heatmaps.
- Analyze correlations between variables and formulate initial hypotheses.

C. Insights

EDA revealed key feature relationships and guided feature selection for modeling. Outliers were documented, and correlations suggested which variables were most predictive.

VII. Modélisation et Analyse

A. Objective

Apply suitable machine learning algorithms to predict heart disease risk.

B. Methodology

- Models considered: Logistic Regression, Random Forest Classifier, and K-Nearest Neighbors (KNN).
- Models trained on the prepared dataset.
- Cross-validation performed to evaluate stability and generalizability.

C. Insights

Random Forest and KNN captured nonlinear relationships effectively. Logistic Regression provided interpretability, offering insight into feature contributions.

VIII. Évaluation des Modèles

A. Objective

Assess model performance using appropriate metrics and select the best-performing model.

B. Methodology

- Classification metrics: Accuracy, Precision, Recall, F1-score, and ROC-AUC.
- Compare models based on these metrics and their ability to generalize.
- Visualize performance using confusion matrices and ROC curves.

C. Insights

Random Forest outperformed other models across most metrics. Logistic Regression provided baseline performance and interpretability. Model evaluation confirmed the chosen models' reliability.

IX. Interprétation et Discussion

A. Objective

Interpret model results, identify key predictors, discuss limitations, and propose improvements.

B. Discussion

- Feature importance highlighted age, cholesterol, and maximum heart rate as critical predictors.
- Limitations included dataset size, potential biases, and reliance on static clinical data.
- Recommendations: expand dataset, test additional algorithms, tune hyperparameters, and validate with real-world clinical data.

C. Insights

The analysis provided actionable insights into factors influencing heart disease risk and highlighted avenues for improving predictive performance.

X. Conclusion

This project successfully implemented a structured heart disease prediction pipeline. Following the eight phases ensured data quality, comprehensive exploration, informed feature selection, model training, evaluation, and interpretation. Random Forest was identified as the most effective model. Future work should focus on dataset expansion, advanced modeling techniques, and clinical validation to enhance predictive accuracy and real-world applicability.

Acknowledgment

We would like to thank our instructor and colleagues for guidance and support throughout this project.

References

[1] Kaggle Heart Disease Dataset. *https://www.kaggle.com/ronitf/heart-disease-uci*
[2] G. James, D. Witten, T. Hastie, R. Tibshirani, *An Introduction to Statistical Learning*, Springer, 2013.
[3] scikit-learn documentation, *https://scikit-learn.org/stable/*