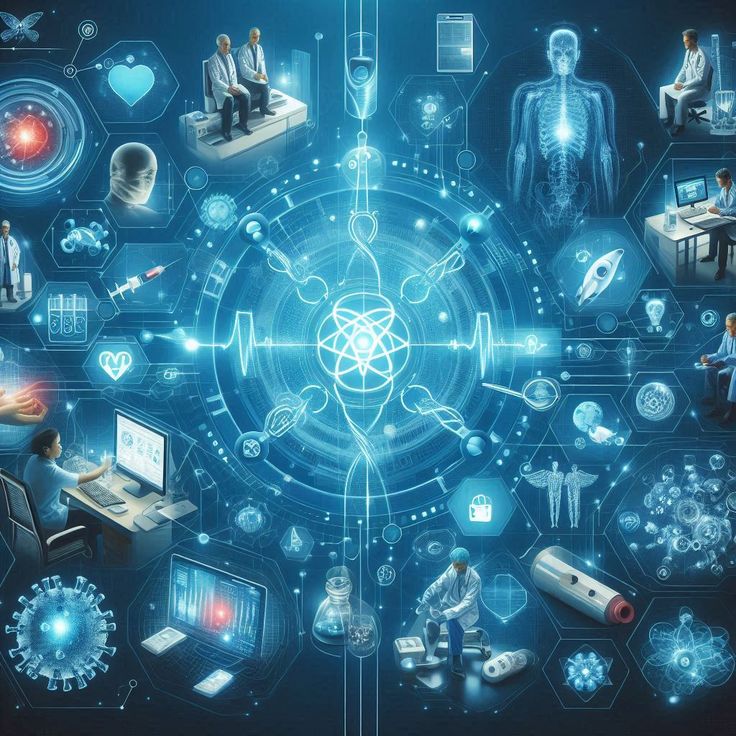
A logo of a globe and a graduation cap

AI-generated content may be incorrect.A black text on a white background

AI-generated content may be incorrect.

IBM Data Science Project

Heart Disease Risk Prediction System Using Machine Learning



|  |  |
| --- | --- |
| Names | Eman Abdelfatah  Hazem Ibrahim  Hussein Mohamed  Kareem fekry  Mohamed Mostafa |
| Group Code | **CLS GIZ2\_AIS4\_S1** |
| Technical Instructor | **Eng. Eslam Adel** |

Table of Contents

[Project Proposal 3](#_Toc197728904)

[Problem Statement: - 3](#_Toc197728905)

[Objectives: 3](#_Toc197728906)

[Scope: - 3](#_Toc197728907)

[Project Gantt Chart: 4](#_Toc197728908)

[Task Assignments & Roles: 4](#_Toc197728909)

[Dataset Exploration (EDA) 5](#_Toc197728910)

[Introduction: - 5](#_Toc197728911)

[System’s Behavior: - 5](#_Toc197728912)

[Dataset Overview: - 5](#_Toc197728913)

[Data Preprocessing: - 6](#_Toc197728914)

[Exploratory Data Analysis (EDA): - 6](#_Toc197728915)

[Key Findings: - 6](#_Toc197728916)

[Data Ingestion: - 7](#_Toc197728917)

[Data Preprocessing: - 7](#_Toc197728918)

[Feature Selection: - 7](#_Toc197728919)

[Model Training: - 7](#_Toc197728920)

[Prediction: - 8](#_Toc197728921)

[Visualization: - 8](#_Toc197728922)

[Backend: - 9](#_Toc197728923)

[Frontend: - 9](#_Toc197728924)

[Machine Learning Model: - 9](#_Toc197728925)

[Database: - 10](#_Toc197728926)

[Model Summary and Deployment 11](#_Toc197728927)

[Input Features: - 11](#_Toc197728928)

[Deployment Approach: - 11](#_Toc197728929)

[Preprocessing & Encoding: - 11](#_Toc197728930)

[Evaluation Result 12](#_Toc197728931)

# Project Proposal

## Problem Statement: -

In the healthcare industry, timely and accurate predictions of patient health outcomes are essential for improving patient care, optimizing resource management, and supporting clinical decision-making. However, the complexity and volume of healthcare data make it challenging to derive actionable insights through traditional methods. Predictive analytics, powered by machine learning, offers a potential solution by enabling the analysis of large datasets to identify patterns and forecast patient outcomes.

This project aims to build a predictive system that estimates heart disease risk based on lifestyle and health-related factors. to predict patient risks, identify trends, and support decision-making processes. This system aims to enhance patient care by providing accurate predictions that aid in early diagnosis, treatment planning, and resource allocation.

# Objectives:

1. **Heart Disease Risk Prediction**: Build a machine learning model that accurately predicts an individual’s risk of heart disease based on lifestyle and health-related inputs such as BMI, physical activity, sleep time, and chronic conditions.
2. **User-Friendly Interaction**: Design a simple and intuitive web-based interface using Streamlit that allows users to input personal health data and receive immediate feedback.
3. **Real-Time Evaluation**: Enable real-time prediction by integrating the trained model into a live web application without requiring backend infrastructure or databases.
4. **Educational Insight**: Help users understand how certain health and lifestyle factors impact heart disease risk through clear, visual, and personalized output.
5. **Portable & Lightweight Deployment**: Deliver the application as a lightweight tool that can run locally or on the web with minimal setup, making it accessible to a wider audience.

## Scope: -

1. **Data Collection:** Utilize a structured dataset containing individual health and lifestyle information, including age group, BMI, general health rating, chronic conditions, physical activity, and sleep time.
2. **Exploratory Data Analysis (EDA):** Analyze feature distributions, correlations, and class balance to understand the most significant predictors of heart disease.
3. Model Development: Train and evaluate a Random Forest Classifier to predict the likelihood of heart disease based on preprocessed input features.
4. **Web-Based Deployment:** Deploy the model in a user-friendly web application using Streamlit, enabling real-time risk prediction from user input.
5. **User Accessibility & Portability:** Ensure the application is lightweight, easily executable on local machines, and does not require extensive infrastructure or backend integration.

## A screenshot of a computer AI-generated content may be incorrect.Project Gantt Chart:

## Task Assignments & Roles:

|  |  |
| --- | --- |
| Team Member | Responsibilities |
| Mohamed Mostafa | Data collection, preprocessing, and feature engineering |
| Hazem Ibrahem | Exploratory data analysis, visualization, and model training |
| Hussein Mohamed | Model development, optimization, and deployment |
| Kareem Fekry | API development, web dashboard, and integration |
| Eman Abdelfatah | Coordination, documentation, and final reporting |

# Dataset Exploration (EDA)

## Introduction: -

This report explores a structured health dataset used to train a machine learning model for predicting the risk of heart disease. The dataset includes a range of features describing individuals' demographics, physical health, lifestyle habits, and chronic conditions. The goal of this analysis is to identify patterns and feature relationships that contribute to predicting heart disease risk.

## System’s Behavior: -

The system follows a structured and streamlined flow:

1. **Data Collection**: Health and lifestyle data was gathered in a tabular format (e.g., CSV), including features such as BMI, sleep time, physical activity, diabetes, and general health.
2. **Data Preprocessing**: Categorical features were manually encoded into numerical values. No normalization or missing value handling was required due to dataset completeness and model type.
3. Model Training: A RandomForestClassifier was trained on the processed data to classify individuals as "at risk" or "not at risk" of heart disease.
4. **Deployment**: The model was deployed using Streamlit, allowing users to interactively input data and receive predictions.
5. **Prediction**: Real-time predictions are provided based on user inputs.
6. **Visualization & Feedback**: Users receive immediate feedback in the form of success or warning messages based on the prediction outcome.

## Dataset Overview: -

The dataset contains various health and lifestyle-related features, including:

* **Demographic Information**: Age category, Sex, Race
* **Health Indicators**: BMI, General Health rating, Physical and Mental Health days, Difficulty walking
* **Chronic Conditions**: Diabetes, Asthma, Kidney Disease, Skin Cancer
* **Lifestyle Factors**: Smoking, Alcohol Consumption, Physical Activity, Sleep Time

## Data Preprocessing: -

The following steps were performed to prepare the data for machine learning:

* **Manual Encoding**: Categorical features (e.g., gender, race, general health) were manually encoded using custom dictionaries.
* **Binary Mapping**: Yes/No responses were converted to 1/0 values for features such as smoking, physical activity, diabetes, and chronic diseases.
* **Ordinal Encoding**: General health ratings and age categories were mapped to ordinal numeric scales.
* No Scaling or Imputation: Since the dataset was clean and the model was ensemble-based, no scaling or missing value imputation was required.

## Exploratory Data Analysis (EDA): -

The EDA focused on understanding feature distributions and their relation to heart disease:

* **Distribution Plots**: Used to visualize key numeric variables such as BMI, physical and mental health days, and sleep time.
* **Count Plots**: Analyzed frequency of categorical attributes like sex, age category, general health, and diabetes.
* **Correlation Analysis**: Explored the relationships between features and the target variable (heart disease presence).
* **Class Balance Check**: Evaluated how balanced the dataset was between high-risk and low-risk cases.

## Key Findings: -

* Individuals with **poor general health ratings** and **high BMI** are more likely to be predicted as high-risk.
* **Diabetes**, **lack of physical activity**, and **difficulty walking** were strongly associated with heart disease risk.
* **Age** plays a major role, with older age groups showing higher predicted risk.
* The **combination of multiple lifestyle and chronic condition indicators** proved more predictive than any single feature.

Machine Learning Pipeline for Heart Disease Risk Prediction System

The purpose of this pipeline is to develop a lightweight and capable of providing feature importance insights model that predicts the likelihood of heart disease based on user-provided health and lifestyle information. The pipeline is designed for real-time interaction through a web application built with Streamlit.

## Data Ingestion: -

* The dataset is loaded from a structured CSV file containing features such as BMI, general health, smoking habits, physical activity, diabetes status, sleep time, and more.
* Python libraries like **Pandas** are used to read and manage the data.

## Data Preprocessing: -

1. **Manual Encoding**: Categorical values (e.g., Yes/No, General Health, Age Category) are manually encoded using predefined dictionaries.
2. **Binary Conversion**: Simple binary mapping is applied for many features (e.g., 1 = Yes, 0 = No).
3. No Scaling or Imputation: Random Forest models are generally robust to unscaled and complete data, so no feature scaling or imputation was required.
4. **Data Splitting**: The dataset is divided into training and testing sets to evaluate performance during development.

## Feature Selection: -

* Features are selected based on domain knowledge and basic correlation analysis.
* All features included in the input form were chosen for their clinical relevance and predictive potential, not through automated feature elimination.

## Model Training: -

1. A RandomForestClassifier is trained on the processed dataset due to its simplicity, interpretability, and compatibility with categorical data.
2. Hyperparameters such as tree depth and minimum samples are tuned for better performance.
3. The model is evaluated using standard metrics like Accuracy and F1-Score (external to the Streamlit app).

## Prediction: -

1. The trained RandomForestClassifier is used to predict whether a user is at risk of heart disease based on new, real-time input.
2. Predictions are binary:

* 1 = At risk of heart disease
* 0 = No significant risk

1. In the deployed app, users receive immediate feedback through visual messages (e.g., "No risk" or "At risk").
2. Evaluation metrics such as Accuracy and F1-Score were considered during development, although they are not displayed in the Streamlit interface.

## Visualization: -

1. The deployed application provides user-level feedback rather than detailed visualizations.
2. Instead of performance graphs, the app displays simple, understandable messages tailored to each user's prediction.
3. During development, internal visualization techniques (outside the app) such as **bar plots**, **correlation heatmaps**, and **feature importance graphs** were used for EDA and model understanding.

Software Architecture for Heart Disease Prediction System

The architecture of the system is designed to be simple, lightweight, and interactive, making it ideal for real-time prediction tasks. It consists of three main components: **Frontend**, **Backend**, and the **Machine Learning Model**. The system is built entirely in Python and deployed using **Streamlit**, requiring no complex database or infrastructure.

## Backend: -

* **Purpose**: Processes user input, applies manual feature encoding, and generates predictions using a trained model.
* **Technologies Used**: Python (Pandas, Scikit-Learn, Joblib, Streamlit).
* **Tasks Performed**:
* Input Handling: Receives form input via Streamlit.
* Feature Encoding: Manually encodes categorical variables using predefined mappings.
* Prediction: Applies the pre-trained model to input data and returns a result in real time.
* No model training or retraining is performed at runtime.

## Frontend: -

* **Purpose**: Provides a clean, interactive interface for users to input data and receive prediction results.
* **Technologies Used**: Streamlit (Python framework).
* **Tasks Performed**:
* Input Collection: Uses form elements like sliders, radio buttons, and dropdowns to gather user health data.
* Prediction Display: Shows output messages such as “No risk” or “At risk” based on model results.
* User Experience: Styled interface using custom CSS to enhance visual clarity and interaction.

## Machine Learning Model: -

* **Purpose:** The core component that predicts heart disease risk based on 17 health and lifestyle features.
* Technologies Used: Scikit-Learn (RandomForestClassifier), Joblib.
* **Tasks Performed:**
  + Model Building: Selecting appropriate algorithms (Logistic Regression, Random Forests, etc.).
  + Model Training: Training with preprocessed data.
  + Model Evaluation: Assessing performance using metrics like Accuracy, Precision, Recall, F1-score.

## Database: -

* **Purpose:** Stores datasets, trained models, and prediction results.
* **Technologies Used:** CSV files for data storage, with potential use of SQLite or other databases for scalability.
* **Tasks Performed:**
  + Data Storage: Keeping raw and processed datasets.
  + Model Storage: Saving trained models for future predictions.
  + Result Storage: Recording prediction outcomes for analysis.

The system will be deployed using **Streamlit**, ensuring simplicity and accessibility for end-users without needing complex installations.

# Model Summary and Deployment

The machine learning model used in this project is a RandomForestClassifier from the scikit-learn library. It was trained on structured health and lifestyle data to classify individuals as either at risk (1) or not at risk (0) of heart disease.

## Input Features: -

The model considers 17 features, including:

* **Demographics**: Age category, sex, race
* **Health indicators**: BMI, general health, physical and mental health days, difficulty walking
* **Chronic conditions**: Diabetes, asthma, kidney disease, skin cancer
* **Lifestyle factors**: Smoking, alcohol use, physical activity, sleep time

## Deployment Approach: -

The model is deployed using **Streamlit**, providing a clean and interactive web-based user interface. Key characteristics of the deployment include:

* **Input Mechanism**: Users enter their health and lifestyle data via sliders, radio buttons, and dropdown menus.
* **Real-Time Prediction**: Upon form submission, the system encodes inputs and instantly predicts the heart disease risk.
* **Feedback**: The result is shown in a clear visual format (e.g., green checkmark for no risk, red warning for at-risk status).

## Preprocessing & Encoding: -

* Categorical features are encoded using manual mapping (e.g., Yes/No → 1/0, ordinal mapping for age and general health).
* No feature scaling or imputation is used due to the model’s compatibility with raw categorical input.

This streamlined pipeline makes the system highly capable of providing feature importance insights, lightweight, and ideal for public use with minimal technical setup.

# Evaluation Result

he Random Forest model was evaluated using standard classification metrics on the test dataset. The following results summarize the model's performance:

* **Precision:**
  + **Class 0 (No risk):** 0.91
  + **Class 1 (At risk):** 0.89Precision measures how many of the positive predictions were actually correct. The model demonstrates high precision for both classes, particularly for non-risk cases, indicating a low rate of false positives.
* **Recall (Sensitivity):**
  + **Class 0 (No risk):** 0.88
  + **Class 1 (At risk):** 0.92Recall indicates how well the model identifies all actual positive cases. The high recall for the at-risk class (0.92) is especially valuable in a healthcare context, where failing to identify someone at risk could have serious consequences.
* **F1-Score:**
  + **Class 0:** 0.90
  + **Class 1:** 0.90The F1-score represents the harmonic mean of precision and recall. Balanced F1-scores for both classes indicate that the model maintains a good trade-off between correctly identifying cases and avoiding false predictions.

These metrics confirm that the Random Forest model provides robust and balanced performance across both risk classes. Its high recall for at-risk individuals makes it well-suited for preventive health screening, where early identification of heart disease risk is critical for timely intervention and better patient outcomes.