DA 204o: **Data Science in Practice**

Course Project

Utilizing CDC Data to Predict Heart Disease in the US

Akhzar Farhan, DSBA , akhzarfarhan@iisc.ac.in
Arun Kumar A, DSBA, aruna@iisc.ac.in
Gajam Venu Gopal, AI, gajamvenu@iisc.ac.in
Venkata Ajay Kolla, AI, ajaykolla@iisc.ac.in

Problem Definition

- ## Background of the problem

  Heart disease is one of the leading causes of mortality worldwide, resulting in high medical costs and significant loss of life. For healthcare providers and insurers, this leads to increased hospitalization costs, resource strain, and economic burdens.

- ## Why is it important?

  By identifying high-risk patients in advance, healthcare organizations can better allocate resources, reduce treatment costs, and ultimately improve the quality of life for patients proactively.

- ## Objectives of the project

  Objective is to create a reliable ML model that predicts the likelihood of heart disease in individuals based on medical history, and factors like high blood pressure, high cholesterol, smoking, diabetes status, obesity (high BMI), physical inactivity, and excessive alcohol consumption.

- ## How can Data Science solve the problem?

  Data science helps us in processing the historical data available to us and develop a reliable ML model which can be used by healthcare providers/patients/insurers to better assess the patient for heart disease.

Data Collection and Preparation

- Data source(s) (where it's from, how it was collected)
  - Indicators of Heart Disease (2022 UPDATE) (kaggle.com)
  - The dataset originally comes from the CDC and is a major part of the Behavioral Risk Factor Surveillance System (BRFSS), which conducts annual telephone surveys to collect data on the health status of U.S. residents
- Description of the data (features, size, format)
  - **Format:** csv
  - **Size:**

    A version of the file with missing data (NaNs): 445,132 rows. = 139.87MB

    A version of the file without missing data (NaNs): 246,022 rows. = 81.98 MB
  - **Features:**

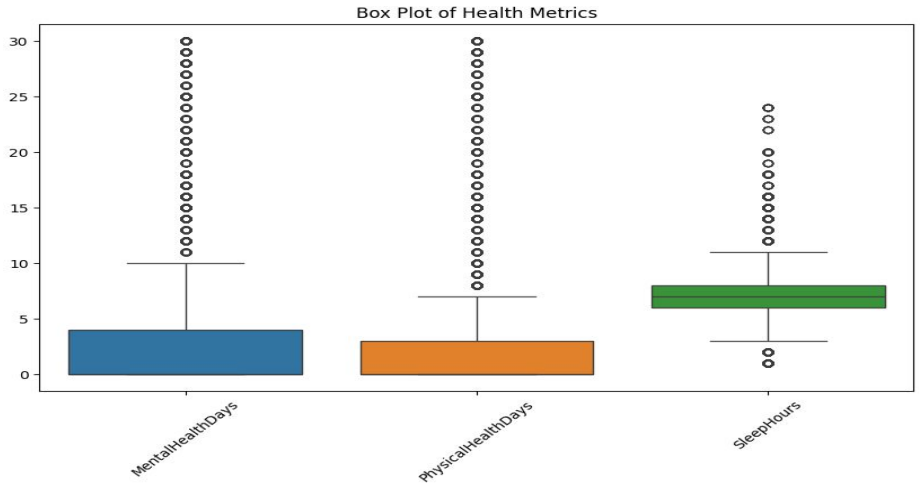
    There are 40 features available in dataset including *PhysicalActivities, SleepHours, HadStroke*, etc.
- Any preprocessing steps required
  - Removing/fixing missing data.
  - Removing features which are not correlated to our prediction class.
  - Data transformation – Data type consistency, Ordinal Encoding, OHE, etc.

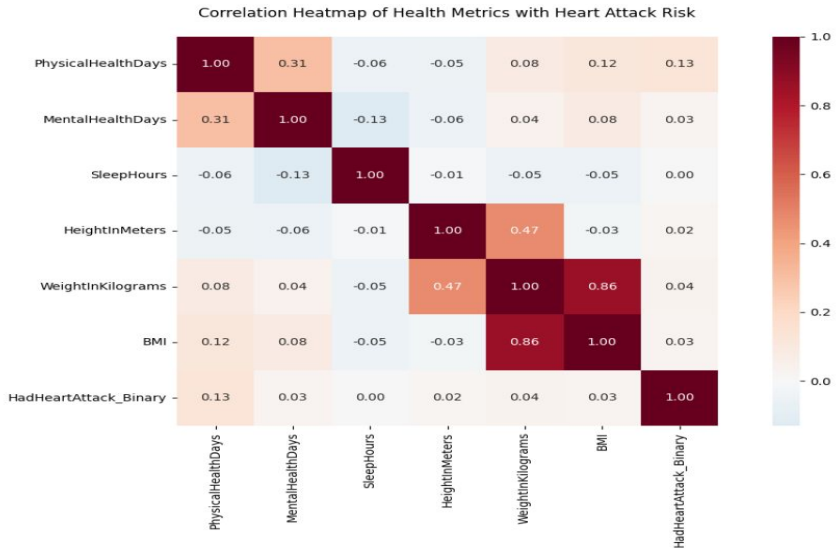
Data Preprocessing

- **Imputation for missing values:** Removed rows with NaN values
- **Unbalanced Dataset:** Down-Sampling of majority class
- **One Hot Encoding:** For binary features
- **Binning:** Categorised BMI into smaller intervals
- **Ordinal Encoding:** For categorical data with some inherent order or ranking

EDA

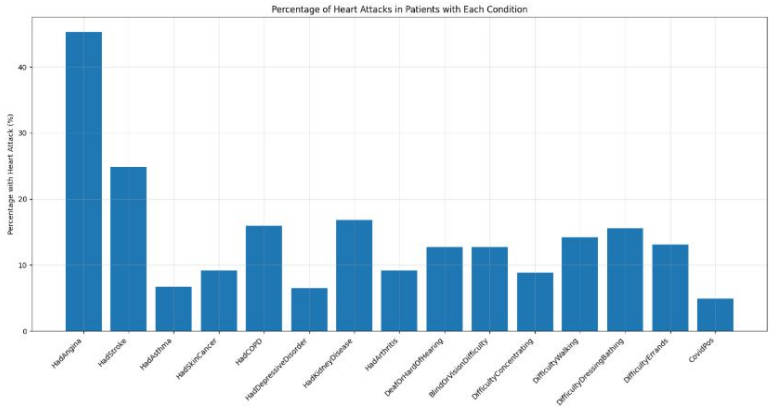
Numerical Variables Analysis

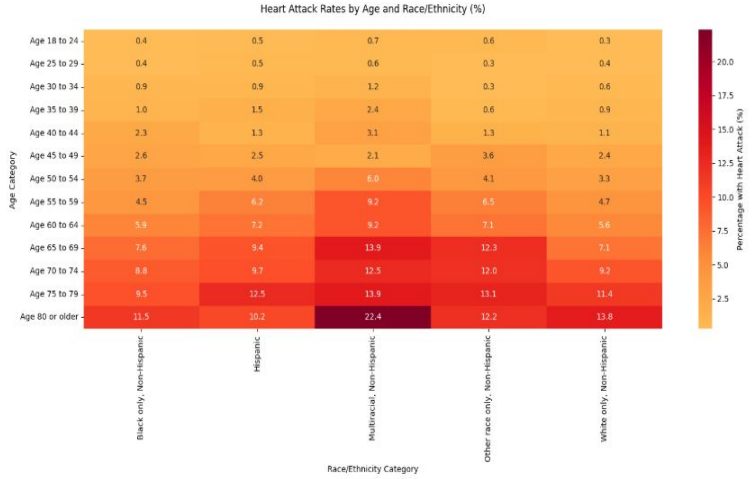


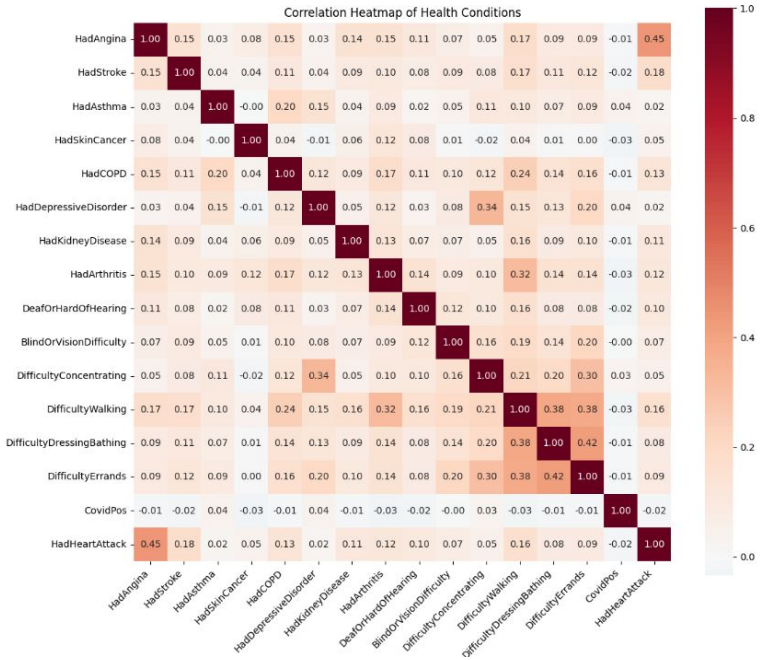


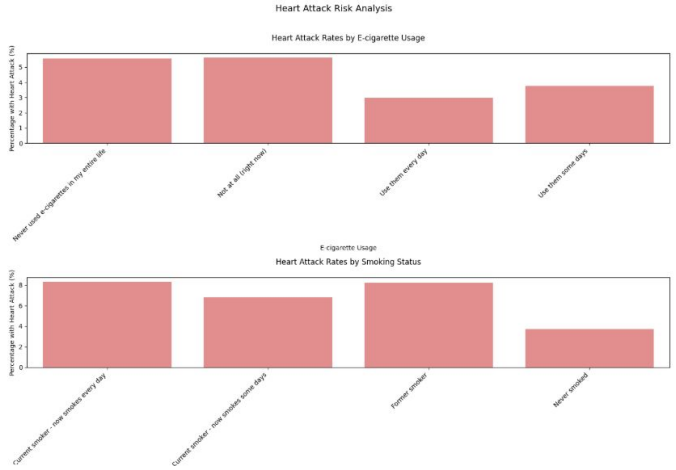
EDA

Categorical Variable Analysis









Model training

Feature Selection: Using the correlation matrix, selected necessary features for training

Train-Test split: 30% test, 70% train with stratify w.r.t. class.

Baseline model training:

Algorithm	F1 score
Logistic regression	0.85
Decision Tree	0.87
AdaBoost Classifier	0.90
Gradient Boosting	0.90
Random Forest	0.89
XGBoost	0.90

Model training

Model trained with Hyperparameters

Random Forest

Best RandomForestClassifier:
Accuracy: 0.9108459933589962

	precision	recall	f1-score	support
0	0.93	0.97	0.95	30000
1	0.68	0.47	0.55	4031
accuracy			0.91	34031
macro avg	0.81	0.72	0.75	34031
weighted avg	0.90	0.91	0.90	34031

XGBoost

Best XGBClassifier:
Accuracy: 0.9112279980018219

	precision	recall	f1-score	support
0	0.93	0.97	0.95	30000
1	0.69	0.46	0.55	4031
accuracy			0.91	34031
macro avg	0.81	0.72	0.75	34031
weighted avg	0.90	0.91	0.90	34031

DNN Classifier

Classification Report:

	precision	recall	f1-score	support
0	0.92	0.98	0.95	30000
1	0.70	0.41	0.51	4031
accuracy			0.91	34031
macro avg	0.81	0.69	0.73	34031
weighted avg	0.90	0.91	0.90	34031

Hyperparameter tuning with cross validation

04_model-creation.ipynb

Random Forest	RandomizedSearchCV with 10 Iter. and CV = 5 Scoring Metric = F1-score `param_grid = { 'n_estimators': [50, 100, 200], 'max_depth': [None, 10, 20], 'min_samples_split': [2, 5, 10], 'max_features': ['sqrt', 'log2', None] }`	Accuracy = 0.91 F1 Score = 0.90 `'n_estimators': 100, 'max_depth': 20, 'min_samples_split': 10, 'max_features': 'sqrt',`
XGBoost	RandomizedSearchCV with 10 Iter. and CV = 5 Scoring Metric = F1-score `param_dist = { 'n_estimators': [100, 200, 300], 'learning_rate': [0.01, 0.1, 0.2], 'max_depth': [3, 5, 7], 'subsample': [0.5, 0.7, 0.9] }`	Accuray = 0.91 F1 Score = 0.90 `'n_estimators': 300, 'learning_rate': 0.01, 'max_depth': 3, 'subsample': 0.7,`
Deep Neural Network (DNN)	GridSearchCV with CV = 3 Scoring Metric = F1-score `param_grid = { 'model__neurons': [128, 256], 'batch_size': [512, 1024], 'epochs': [10, 20], }`	Accuracy = 0.91 F1 Score = 0.90 `'model__neurons': 256, 'batch_size': 1024, 'epochs': 10,`

Results

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.83	0.89	0.83	0.85
Decision Tree	0.87	0.86	0.87	0.87
Random Forest	0.90	0.88	0.90	0.89
AdaBoost	0.91	0.90	0.91	0.90
Gradient Boosting	0.91	0.90	0.91	0.90
Random Forest (Hyper-parameter tuned)	0.91	0.90	0.91	0.90
XGBoost (Hyper-parameter tuned)	0.91	0.90	0.91	0.90
Deep Neural Network (Hyper-parameter tuned)	0.91	0.90	0.91	0.90

Challenges and Future work

Challenges	Solutions
Missing/Incomplete values	Dropped data points with missing values
Imbalance Dataset	Under-Sampling of majority class
Large set of features	Feature selection using EDA techniques such as correlation matrix
Bias in Dataset	Used Ensemble Learning such as Random Forest, Adaboost, Gradient Boosting, XGBoost,
Over-Fitting	Multiple models with multiple metrics such as accuracy, precision, recall and F1-score
Model Improvement	Hyperparameter tuning techniques such as Randomised Search and Grid Search with cross-validation

Future Work:

- Collect more data from different sources to make rich dataset
- Deploy with UI

Data Science Canvas

Project: Utilizing CDC Data to Predict Heart Disease in the US

Team: *Akhzar Farhan, Arun Kumar A, Gajam Venu Gopal, Venkata Ajay Kolla*

Problem Statement

Business Case & Value Added
Which business case should be analyzed and what added value does it generate?

Predict heart disease in US population by examining key previous illness and lifestyle factors. It will support healthcare providers in making more informed, proactive clinical decisions.

Data Landscape
Which data is required for this and which is already available? Which additional data has to be collected?

Health records of US population. We have taken the dataset from Kaggle https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease/data

Model Selection
Which analysis methods can be considered on the basis of the specific data landscape and the business case?

Classification models are used.

- Logistic regression
- Decision Tree Classifier
- AdaBoost Classifier
- Gradient Boosting Classifier
- Random Forest Classifier
- XGBoost Classifier
- DNN KerasClassifier

Model Requirements
Which model requirements must be complied with in order to obtain a valid model?

Handle large datasets efficiently. Evaluate multiple models based on their performance metrics and recommend the best of it.

Software & Libraries
Which software should be used? Is there already a standard solution? Which libraries are used?

Software and libraries include: Python, Pandas, NumPy, Scikit-learn, Matplotlib, Seaborn, Keras,etc

Skills
What skills are needed to provide the data and model development?

Proficiency in Python, pandas, visualization libraries. Good understanding of machine learning algorithms. Understanding of evaluation metrics

Execution & Evaluation

Model Evaluation
Which indicators require quality control and validation and how should they be interpreted? Is real-time monitoring necessary?

Build base models with different classification algorithms and choose suitable algorithms based on their performance metrics like accuracy, precision, recall and F1 score.
Use hyperparameter tuning to find the best parameters of these chosen algorithms.
Build and verify the performance improvement after building the models with these hyper parameters.

Data Storytelling
What requirements does the target group have for the presentation of the results and how do I effectively communicate this data?

Target group can be business stakeholders of Hospitals, Insurance providers.
Insights such as leading indicators of heart disease, feasibility of using this model for their business can be explained

Data Collection & Preparation

Data Selection & Cleansing
Which of the available data is relevant? Do the data have to be cleaned up?

Data is selected from publicly available forums. Remove missing data. Change the data types if necessary.

Data Integration
In which system should the data from different sources be migrated?

Integrate the data wherever necessary. Use necessary functions to for binning and encoding. And remove the original columns after these feature engineering process.

Data Collection
How and with which methods should additionally required data be collected? What properties has this data to fulfil?

Use encoders like label and ordinal encoders to ease model building. Use binning for columns like age, cigarette usage, etc. Downscale the majority class data, if needed to reduce the imbalance datasets for prediction class

Exploratory Data Analysis
Are there outliers or structures to be considered? Creation of descriptive key figures for the first assessment of the data.

Handle missing fields, remove outliers and remove irrelevant columns.
Visualize the data distribution. Use correlation matrix to find the relationship

Thank you