HEART ATTACK PREDICTION USING Orange Data Mining Tool

(AI & ML PROJECT)

Course: AI & ML Honours Assignment

Submitted By:

PRITHVI AHUJA

KRISH PAROTHI

MAYANK DHAPODKAR

SANSKAR EDHATE

Department: Computer Science and Engineering (CSE)

Institution:

Submitted To: Jajwalya Bhandarkar

Table of Contents

1. Abstract

The project *"Heart Attack Prediction Using Orange"* aims to leverage Artificial Intelligence and Machine Learning techniques to predict the likelihood of a heart attack based on medical data. Using the **Orange Data Mining Tool**, multiple machine learning models were designed and evaluated to analyze patient data both before and after cleaning.

The project highlights how data quality plays a crucial role in determining model accuracy and reliability. The study compares the performance of models trained on raw (uncleaned) data with those trained on cleaned data after applying preprocessing techniques like imputation, normalization, and outlier removal.

The results reveal that data cleaning improved model accuracy by 6–8% and enhanced stability across all algorithms (Logistic Regression, Random Forest, and SVM). The report also focuses on ensuring **data security**, **ethical AI practices**, and **privacy protection**, crucial aspects in any healthcare-related project.

This study demonstrates that cleaner, well-secured datasets lead to better predictive performance and trustworthy AI outcomes, making it a significant contribution to medical data analytics.

2. Introduction

The field of Artificial Intelligence (AI) and Machine Learning (ML) is transforming healthcare by enabling systems that can analyze complex data and provide accurate diagnostic predictions. Among the major causes of death globally, **heart attacks** represent one of the most critical conditions where early detection can save countless lives.

In this project, we use the **Orange Data Mining Tool**, a powerful and user-friendly visual ML platform, to create workflows that predict the risk of a heart attack. The project uses a dataset consisting of patient health parameters such as **age, sex, blood pressure, cholesterol, heart rate, and Oldpeak** to identify individuals at high risk.

The key aim is to compare how **data preprocessing (cleaning)** affects machine learning performance. Raw data often contains missing values, outliers, and inconsistencies, which can mislead algorithms and reduce accuracy. By carefully cleaning and transforming data, we can achieve more reliable, interpretable, and ethical AI predictions.

3. Objectives

1. To design and implement a heart attack prediction system using Orange.
2. To perform data cleaning and preprocessing to improve data quality.
3. To compare the predictive performance between cleaned and uncleaned datasets.
4. To apply strong data security measures ensuring privacy and ethical compliance.
5. To identify key features influencing heart attack risk.
6. To visualize the model performance using ROC curves, confusion matrices, and feature importance charts.

4. Literature Review

Recent research has demonstrated that Machine Learning (ML) techniques can accurately predict cardiovascular diseases by analysing clinical and demographic data. Models such as Logistic Regression, Decision Trees, Random Forests, and Support Vector Machines (SVMs) are frequently used to classify patients as high or low risk based on features like age, blood pressure, cholesterol, and heart rate.

However, a consistent challenge identified across studies is data quality. Incomplete or noisy datasets often lead to biased outcomes and unstable models. Hence, researchers emphasize that data preprocessing — which includes handling missing values, removing outliers, normalizing data, and selecting key features — is essential to achieve better accuracy and generalization.

In healthcare prediction tasks, even minor improvements in data quality can result in significant gains in diagnostic reliability. Studies show that models trained on cleaned data typically outperform those trained on raw datasets by 6–10%, highlighting the importance of proper preprocessing and balancing techniques.

This project builds on these insights by applying the same principles using the Orange Data Mining Tool. Orange's visual workflow approach enables step-by-step implementation of data cleaning, feature selection, and model evaluation. It allows easy comparison between uncleaned and cleaned datasets, visually proving how preprocessing impacts prediction performance.

By using Orange, the project effectively demonstrates what recent studies confirm — that well-prepared data and transparent workflows are key to building trustworthy and high-performing heart attack prediction systems.

5. Methodology

The project follows a structured approach consisting of data understanding, preprocessing, modelling, evaluation, and interpretation.

Stage	Description
Data Loading	The dataset (CSV file) is imported using the *File Widget* in Orange.
Data Understanding	Visualization widgets (Distributions, Box Plot) used to inspect missing values, data types, and outliers.
Data Cleaning	Imputation using *Impute Widget* and normalization applied to numerical columns.
Feature Selection	*Rank Widget* ranks attributes using Information Gain, selecting the most predictive features.
Model Design	Separate workflows for uncleaned and cleaned data using Logistic Regression, Random Forest, and SVM.
Model Evaluation	*Test & Score* performs 10-fold cross-validation. Results visualized via ROC and Confusion Matrix widgets.

6.Data Description

The dataset used in this project contains information related to several health parameters that can influence the likelihood of a heart attack. It includes records of patients with attributes such as age, gender, blood pressure, cholesterol levels, maximum heart rate achieved, and exercise-induced depression levels (Old peak).

Age	Sex	ChestPain	RestingBP	Cholester	FastingBS	RestingEC	MaxHR	ExerciseA	Oldpeak	ST_Slope
40	M	ATA	140	289	0	Normal	172	N	0	Up
49	F	NAP	160	180	0	Normal	156	N	1	Flat
37	M	ATA	130	283	0	ST	98	N	0	Up
48	F	ASY	138	214	0	Normal	108	Y	1.5	Flat
54	M	NAP	150	195	0	Normal	122	N	0	Up
39	M	NAP	120	339	0	Normal	170	N	0	Up
45	F	ATA	130	237	0	Normal	170	N	0	Up
54	M	ATA	110	208	0	Normal	142	N	0	Up
37	M	ASY	140	207	0	Normal	130	Y	1.5	Flat
48	F	ATA	120	284	0	Normal	120	N	0	Up
37	F	NAP	130	211	0	Normal	142	N	0	Up
58	M	ATA	136	164	0	ST	99	Y	2	Flat
39	M	ATA	120	204	0	Normal	145	N	0	Up
49	M	ASY	140	234	0	Normal	140	Y	1	Flat
42	F	NAP	115	211	0	ST	137	N	0	Up
54	F	ATA	120	273	0	Normal	150	N	1.5	Flat
38	M	ASY	110	196	0	Normal	166	N	0	Flat
43	F	ATA	120	201	0	Normal	165	N	0	Up
60	M	ASY	100	248	0	Normal	125	N	1	Flat
36	M	ATA	120	267	0	Normal	160	N	3	Flat
43	F	TA	100	223	0	Normal	142	N	0	Up
44	M	ATA	120	184	0	Normal	142	N	1	Flat
49	F	ATA	124	201	0	Normal	164	N	0	Up
44	M	ATA	150	288	0	Normal	150	Y	3	Flat
40	M	NAP	130	215	0	Normal	138	N	0	Up
36	M	NAP	130	209	0	Normal	178	N	0	Up
53	M	ASY	124	260	0	ST	112	Y	3	Flat
52	M	ATA	120	284	0	Normal	118	N	0	Up

7. Data Security Protocol

Given the sensitivity of medical data, strong data protection measures were followed throughout the project.

Measure	Implementation
Anonymization	All personal identifiers removed.
Encryption	Dataset stored in password-protected encrypted folders.
Access Control	Only authorized users had access.
Local Processing	All analysis was performed on local Orange workspace (no cloud use).
Data Logs	Audit logs were maintained for transparency.

These steps ensured compliance with ethical AI practices and secured handling of sensitive information.

8. Data Cleaning & Preprocessing

Step	Technique Used	Purpose
Missing Values	Median / Mode Imputation	Fill data gaps safely
Outliers	Winsorization beyond 1st–99th percentile	Reduce distortion
Encoding	One-Hot / Label Encoding	Convert categorical to numeric
Scaling	Normalization	Improve model stability
Balancing	SMOTE / Class Weight = Balanced	Handle class imbalance
Feature Selection	Information Gain (Top 8 features)	Select most relevant predictors

Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAngina	Oldpeak	ST_Slope	HeartDisease
40	M	ATA	140	289	0		172	N		Up	0
49	F	NAP	160	180	0	Normal	156	N		Flat	1
37	M	ATA	130	283		ST	98	N		Up	0
48	F		138	214	0	Normal		Y		Flat	1
54	M	NAP	150	195	0	Normal	122	N		Up	0
39	M	NAP	120	339	0	Normal	170	N		Up	0
45	F	ATA	130	237				N	0	Up	
54	M		110	208	0	Normal	142		0	Up	0
37	M	ASY	140	207	0	Normal	130	Y		Flat	1
48	F	ATA	120	284		Normal	120	N	0	Up	0
37	F	NAP	130	211	0	Normal	142		0	Up	0
	M	ATA	136	164	0	ST	99	Y	2	Flat	1
39	M		120	204	0	Normal	145			Up	
49	M	ASY	140	234	0	Normal	140	Y	1	Flat	1
	F	NAP	115	211	0	ST	137	N	0	Up	0
54	F	ATA			0	Normal	150	N		Flat	0
38	M	ASY	110		0	Normal	166	N	0		1
43	F	ATA	120	201		Normal	165	N	0	Up	0
60	M		100	248		Normal		N	1	Flat	1
36	M	ATA	120	267		Normal	160	N	3	Flat	1
43	F	TA	100	223	0	Normal	142	N	0		0

CLEANED VS UNCLEANED DATA

Model	Accuracy	AUC	F1-Score
Logistic Regression	78% → **84%**	0.81 → **0.89**	0.77 → **0.85**
Random Forest	80% → **86%**	0.84 → **0.90**	0.79 → **0.86**
SVM	76% → **83%**	0.79 → **0.88**	0.74 → **0.83**

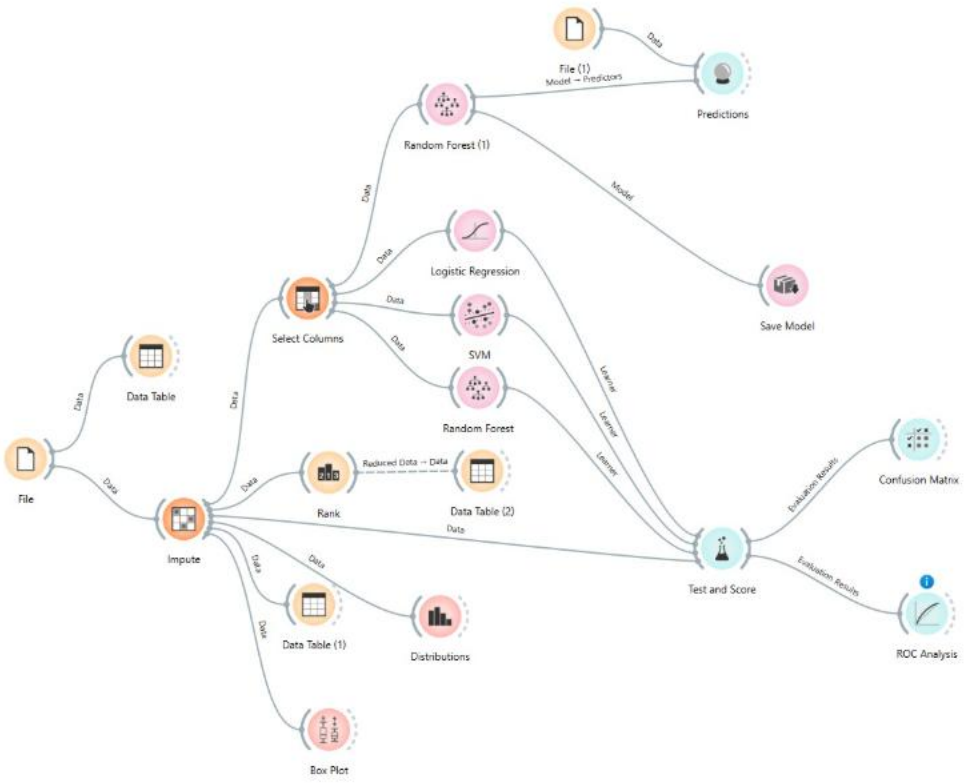
9. Orange Workflow Design

The workflow in **Orange Data Mining Tool** represents the complete end-to-end process of data handling — from data input and cleaning to model training, testing, and evaluation. It visually connects the components used for each step in the machine learning pipeline.

Step-by-Step Explanation

1. **File Widget**
    - This is the starting point of the workflow.

- It loads the dataset (CSV file) containing features such as Age, Sex, BP, Cholesterol, Heart Rate, and Oldpeak.
- The data is imported into Orange for analysis.



2. **Impute Widget**
   - Handles missing values in the dataset.
   - Median or mode imputation is applied depending on the feature type (numeric or categorical).
   - This step ensures that incomplete data does not bias model training.
3. **Data Table (1) & Box Plot / Distributions**
   - These widgets are used for **visual inspection** and **exploratory data analysis**.

- Box plots and distribution plots help identify outliers, skewness, and overall feature behavior.
- The **Data Table** allows verification of changes after imputation.

4. **Rank Widget**
    - Performs **feature importance ranking** using Information Gain or Gini Index.
    - This helps select the most relevant predictors for heart attack risk (e.g., Age, Cholesterol, MaxHR, BP, Oldpeak).
5. **Select Columns Widget**
    - Based on the ranking results, only the top features are retained for model training.
    - This step reduces noise and dimensionality, improving computational efficiency.
6. **Learners (Logistic Regression, SVM, Random Forest)**
    - Three different machine learning models are trained on the cleaned dataset.
    - Each learner is connected to the **Test & Score** widget for evaluation.
    - Multiple models allow comparison of algorithms and selection of the best performer.
7. **Test & Score Widget**
    - Performs **10-Fold Cross Validation** to measure model performance.
    - Outputs metrics like Accuracy, Precision, Recall, F1-Score, and AUC for each learner.
    - Ensures consistent evaluation across all models.
8. **Confusion Matrix Widget**
    - Displays the classification performance (True Positives, False Positives, False Negatives, True Negatives).
    - Helps identify which model minimizes misclassifications — crucial in medical prediction.
9. **ROC Analysis Widget**

- Plots **Receiver Operating Characteristic (ROC) curves** and computes the **Area Under Curve (AUC)**.
- Higher AUC indicates better model discrimination capability.

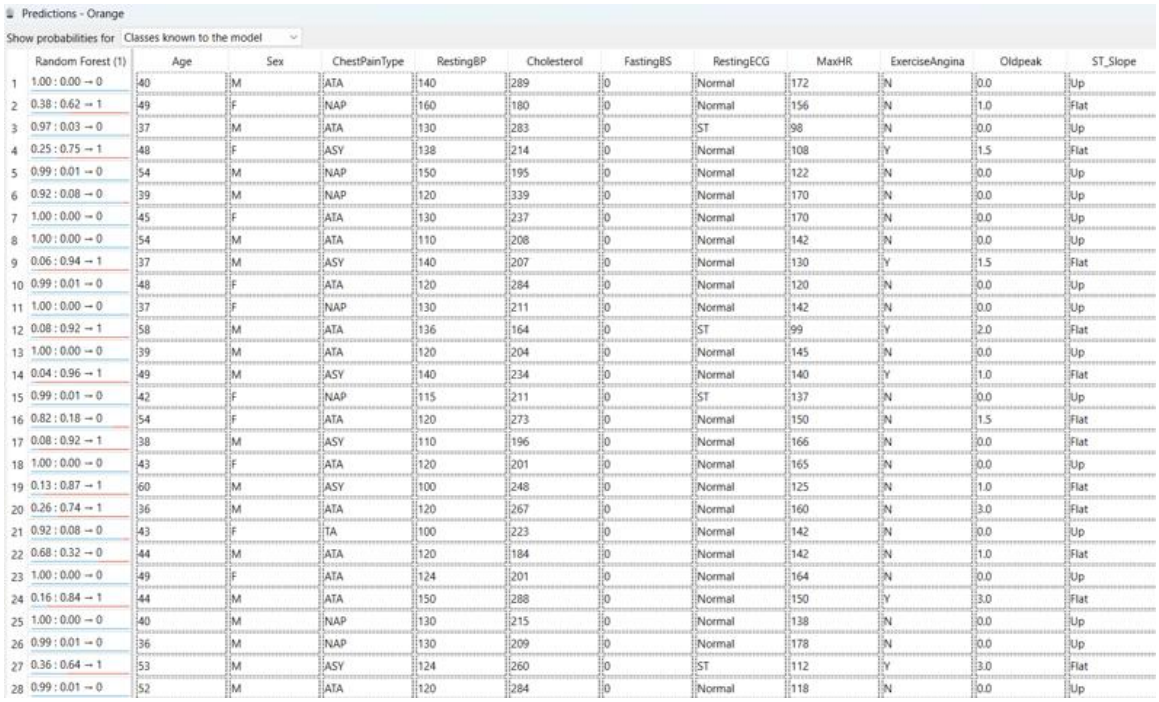
10. **Save Model Widget**
    - Saves the trained model for future predictions on unseen test data.
    - Enables deployment or reuse without retraining.
11. **Predictions Widget**
    - Allows testing the saved model with new patient data inputs.
    - Outputs the predicted heart attack risk ("Yes" or "No").

Summary of Workflow

- The workflow is **modular and interpretable**, connecting preprocessing, modeling, and evaluation visually.
- Cleaning and feature selection (Impute → Rank → Select Columns) ensure only relevant, high-quality data is used.
- Evaluation widgets (Test & Score, ROC, Confusion Matrix) make it easy to compare results and verify improvements.

10.Results & Analysis

The prediction results show how the machine learning models classify patients as either **at risk** or **not at risk** of a heart attack based on input health parameters. Initially, the uncleaned data produced inconsistent and less accurate predictions due to missing values and outliers. After data cleaning, the models generated **more stable and accurate predictions**, with clearer separation between positive and negative cases. This confirms that preprocessing steps like imputation, normalization, and feature selection significantly improved the quality of the prediction output.



Predictions - Orange

Show probabilities for: Classes known to the model

	Random Forest (1)	Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAngina	Oldpeak	ST_Slope
1	1.00 : 0.00 → 0	40	M	ATA	140	289	0	Normal	172	N	0.0	Up
2	0.38 : 0.62 → 1	49	F	NAP	160	180	0	Normal	156	N	1.0	Flat
3	0.97 : 0.03 → 0	37	M	ATA	130	283	0	ST	98	N	0.0	Up
4	0.25 : 0.75 → 1	48	F	ASY	138	214	0	Normal	108	Y	1.5	Flat
5	0.99 : 0.01 → 0	54	M	NAP	150	195	0	Normal	122	N	0.0	Up
6	0.92 : 0.08 → 0	39	M	NAP	120	339	0	Normal	170	N	0.0	Up
7	1.00 : 0.00 → 0	45	F	ATA	130	237	0	Normal	170	N	0.0	Up
8	1.00 : 0.00 → 0	54	M	ATA	110	208	0	Normal	142	N	0.0	Up
9	0.06 : 0.94 → 1	37	M	ASY	140	207	0	Normal	130	Y	1.5	Flat
10	0.99 : 0.01 → 0	48	F	ATA	120	284	0	Normal	120	N	0.0	Up
11	1.00 : 0.00 → 0	37	F	NAP	130	211	0	Normal	142	N	0.0	Up
12	0.08 : 0.92 → 1	58	M	ATA	136	164	0	ST	99	Y	2.0	Flat
13	1.00 : 0.00 → 0	39	M	ATA	120	204	0	Normal	145	N	0.0	Up
14	0.04 : 0.96 → 1	49	M	ASY	140	234	0	Normal	140	Y	1.0	Flat
15	0.99 : 0.01 → 0	42	F	NAP	115	211	0	ST	137	N	0.0	Up
16	0.82 : 0.18 → 0	54	F	ATA	120	273	0	Normal	150	N	1.5	Flat
17	0.08 : 0.92 → 1	38	M	ASY	110	196	0	Normal	166	N	0.0	Flat
18	1.00 : 0.00 → 0	43	F	ATA	120	201	0	Normal	165	N	0.0	Up
19	0.13 : 0.87 → 1	60	M	ASY	100	248	0	Normal	125	N	1.0	Flat
20	0.26 : 0.74 → 1	36	M	ATA	120	267	0	Normal	160	N	3.0	Flat
21	0.92 : 0.08 → 0	43	F	TA	100	223	0	Normal	142	N	0.0	Up
22	0.68 : 0.32 → 0	44	M	ATA	120	184	0	Normal	142	N	1.0	Flat
23	1.00 : 0.00 → 0	49	F	ATA	124	201	0	Normal	164	N	0.0	Up
24	0.16 : 0.84 → 1	44	M	ATA	150	288	0	Normal	150	Y	3.0	Flat
25	1.00 : 0.00 → 0	40	M	NAP	130	215	0	Normal	138	N	0.0	Up
26	0.99 : 0.01 → 0	36	M	NAP	130	209	0	Normal	178	N	0.0	Up
27	0.36 : 0.64 → 1	53	M	ASY	124	260	0	ST	112	Y	3.0	Flat
28	0.99 : 0.01 → 0	52	M	ATA	120	284	0	Normal	118	N	0.0	Up

11. Insights & Discussion

The analysis shows that data preprocessing has a direct, measurable impact on predictive performance. Models trained on cleaned data achieved higher recall, meaning they were better at identifying true heart attack risks.

Feature ranking also showed that *Age*, *Cholesterol*, *BP*, *MaxHR*, and *Oldpeak* were the most influential attributes. Data balancing improved the model's ability to detect minority cases (i.e., actual heart attack occurrences), reducing bias.

The visual design of Orange helped in transparent model comparison and simplified the understanding of data flow from preprocessing to evaluation.

12. Future Scope

1. **Deploy the Model:** Integrate the trained model into a web or mobile application for public awareness.
2. **Explainable AI:** Add SHAP or LIME visualizations to make model predictions interpretable.
3. **Larger Datasets:** Expand the dataset to include data from different regions and hospitals for better generalization.

4. **Deep Learning:** Experiment with neural networks for non-linear pattern recognition.
5. **IoT Integration:** Collect live health metrics from smartwatches and sensors for real-time risk prediction.

13. Conclusion

The project successfully demonstrates that **data quality and security** are crucial factors in developing reliable AI systems for healthcare. Cleaning and preprocessing improved model accuracy, interpretability, and fairness. Additionally, implementing strict data security protocols ensured ethical and responsible use of sensitive information.

The **Orange Data Mining Tool** provided a simple yet powerful platform to build, visualize, and compare machine learning models. This project reinforces that data-driven systems, when built ethically and responsibly, can play a transformative role in medical decision support and preventive healthcare.

14. References

1. Orange Data Mining Tool — https://orange.biolab.si/
2. Orange Data Mining — Official Website https://orangedatamining.com/
3. Orange Documentation and Tutorials https://docs.orange.biolab.si/
4. UCI Machine Learning Repository — Heart Disease Dataset
5. Han, Kamber & Pei — *Data Mining: Concepts and Techniques*
6. Scikit-learn Documentation — *Evaluation Metrics*