EMINENCE5.0

Artificial Intelligence-Based Task

Tasks Details

By
Women in Engineering (WIE) Affinity Group,
IEEE Student Branch,
University of Ruhuna

1 Task 01 – Heart Attack Risk Prediction (Classification)

1.1 Problem Statement

In 2025, an international medical consortium launched the Global Heart Health Study to understand why heart attacks remain one of the leading causes of death worldwide. Researchers collected detailed information from thousands of patients across different continents to build a dataset that could help train machine learning models for predicting heart attack risk.

For every patient, a wide range of factors were recorded:

- Demographics: Age, Sex, Country, Continent, Hemisphere, Income.
- Lifestyle: Smoking, Alcohol Consumption, Diet, Exercise Hours, Physical Activity Days, Sedentary Hours, Sleep Hours.
- Medical History: Diabetes, Obesity, Family History, Previous Heart Problems, Medication Use.
- Clinical Tests: Cholesterol, Triglycerides, BMI, Stress Level.
- Vital Signs: Heart Rate and blood pressure readings — one taken when the patient was relaxed (Resting Blood Pressure) and another after mild exercise (Active Blood Pressure).

The idea was that looking at only one type of data might not fully reflect a patient's cardiovascular stress. Including all the relevant information may give researchers a more complete picture, helping them better identify hidden risks and predict heart attack likelihood with greater accuracy.

Your task is to build a **Machine Learning Classification Model** that can predict the **risk of a heart attack** based on these features.

1.2 Dataset Information

The dataset has been divided into two parts:

1. **Train Dataset** – This includes all patient features and the target column Heart Attack Risk.
   - You can use this dataset to train and validate your models.
   - https://drive.google.com/file/d/1OQWfTaTaRatB8gT3sg723LRtYYKBm5tU/view?usp=sharing
2. **Test Dataset** – This includes all patient features but **does not include the target column**.
   - You need to predict heart attack risk for these patients.
   - https://drive.google.com/file/d/1MznJKj9_X4TBgnkbMWfBSBv9jyCQ6krd/view?usp=sharing

1.3 Columns in the Dataset

The dataset contains detailed health, lifestyle, and demographic information for each patient, aiming to predict the risk of a heart attack. It includes basic patient information such as Patient ID, Age, and Sex, which provides context about the individual. Clinical measures like Cholesterol, Blood Pressure, Heart Rate, Triglycerides, and BMI capture the patient’s cardiovascular health, while medical history indicators such as Diabetes, Family History, Previous Heart Problems, and Medication Use provide insights into existing health conditions and hereditary risks.

Lifestyle and behavioral factors are also included, offering a comprehensive view of the patient’s habits. Columns like Smoking, Obesity, Alcohol Consumption, Exercise Hours Per Week, Physical Activity Days Per Week, Sedentary Hours Per Day, Diet, and Sleep Hours Per Day reflect daily routines and health-related behaviors that can influence heart attack risk. Additionally, socioeconomic indicators such as Income, as well as geographic information like Country, Continent, and Hemisphere, provide context for environmental and lifestyle differences across populations.

Finally, the dataset contains the target variable, Heart Attack Risk, which indicates whether a patient is at risk of a heart attack (0 = No, 1 = Yes). By combining clinical measurements, lifestyle habits, medical history, and demographic information, this dataset provides a rich foundation for building a predictive model. The diverse set of features allows for both statistical and machine learning approaches to understand the complex relationships that contribute to heart attack risk.

Here are all the columns with respective descriptions in this dataset.

- **patient_id**: Unique ID of each patient
- **age**: Age of the patient
- **sex**: Gender of the patient (Male/Female)
- **chol**: Cholesterol level
- **bp**: Blood pressure in systolic/diastolic format
- **hr**: Resting heart rate
- **diabetes**: Whether the patient has diabetes (0 = No, 1 = Yes)
- **family_history**: Family history of heart disease (0 = No, 1 = Yes)
- **smoking**: Smoking status (0 = No, 1 = Yes)
- **obesity**: Obesity status (0 = No, 1 = Yes)
- **alcohol**: Alcohol consumption (0 = No, 1 = Yes)
- **exercise_hr_wk**: Hours of exercise per week
- **diet**: Diet type (e.g., Poor, Average, Healthy)
- **prev_heart_prob**: History of previous heart problems (0 = No, 1 = Yes)

- **med_use**: Whether the patient uses medication (0 = No, 1 = Yes)
- **stress_lvl**: Self-reported stress level (scale 0–10)
- **sedentary_hr**: Hours spent sedentary per day
- **income**: Annual income of the patient
- **bmi**: Body Mass Index
- **triglycerides**: Triglyceride level
- **phys_act_days**: Number of days per week with physical activity
- **sleep_hr**: Average hours of sleep per day
- **country**: Country of residence
- **continent**: Continent of residence
- **hemisphere**: Hemisphere of residence
- **heart_attack_risk**: Target variable: 0 = No, 1 = Yes

1.4 What You Need to Do

- Data Understanding & Preparation
  - Explore and clean the dataset using appropriate preprocessing techniques.
  - Handle missing values, categorical variables, and scaling as required.
- Model Building
  - Train a **classification model** (or multiple models) using the train dataset.
  - Try different algorithms, tuning, or feature selection to get better results.
  - Perform feature selection and hyperparameter tuning to improve performance.
- Model Evaluation
  - Report **training performance** using the following metrics:
    - Accuracy
    - Precision, Recall, F1-score
    - ROC-AUC Score
  - For **competition scoring**, we will use **only Recall** for this task. (Higher Recall = better performance).
- Prediction on Test Data
  - Use your final model to predict Heart Attack Risk for the **test dataset**.
  - Save the output in a **CSV file** with the following format.

- Use the same format for the CSV headings. Otherwise, your submission will not be valid.
- We will use this for evaluate the **testing performance** of your model.

Table 1: Structure of Evaluation CSV for Task 01

patient_id	**heart_attack_risk**
12345	0
67890	1

1.5 What You Need to Submit

All submissions must be uploaded to a public GitHub repository. Private repositories or missing files will be considered invalid.

1. **Link for Notebook**
   - Submit your complete notebook (.ipynb) in your GitHub repo.
   - The notebook must include your code, training process, evaluation metrics (Accuracy, Precision, Recall, F1-score, ROC-AUC Score) and final model predictions generation.
   - File Naming Rule: Rename your notebook as:
     - TeamCode_TeamName_Task1_HeartAttack.ipynb
2. **Training Metrics**
   - Clearly display the evaluation metrics from your training (Accuracy, Precision, Recall, F1-score, ROC-AUC Score) in the notebook.
   - You must also include a **screenshot (one picture)** in your repository showing these metric values for verification.
   - Save it as:
     - TeamCode_TeamName_Task1_Metrics.png
3. **Predictions File**
   - CSV file with columns: **patient_id, heart_attack_risk** (Table 1**).**
   - This file should contain predictions for the **test dataset** provided.
   - File Naming Rule:
     - TeamCode_TeamName_Task1_Predictions.csv

Important: If you do not follow the above rules (file format, naming conventions, or required submissions), your submission will be considered invalid and will not be evaluated.

2 Task 02 - House Price Prediction (Regression)

2.1 Problem Statement

In 2025, a large real estate company decided to build a smart system to estimate house prices more accurately. They collected detailed information about thousands of homes that were recently sold, including their size, number of rooms, age, location, and special features like views and renovations. The goal of this study is to train machine learning models that can predict the selling price of a house based on its characteristics. By analyzing different features such as bedrooms, bathrooms, living area, lot size, neighborhood, and renovation history, researchers hope to understand how each factor influences property value. This can help buyers, sellers, and agents make better decisions in the housing market.

Your task is to build a **Machine Learning Regression Model** that can predict the **price of a house** based on its features.

2.2 Dataset Information

The dataset has been divided into two parts:

3. **Train Dataset** – This includes all house features and the price column.
    - You can use this to train and validate your models.
    - https://drive.google.com/file/d/128XLT44uu9bVSBPd6JiiYyv_M4LJP-3x/view?usp=sharing
4. **Test Dataset** – This includes all house features, but without the price column.
    - You need to predict the prices for these houses.
    - https://drive.google.com/file/d/1JMUa5MIwyx3QEXk3KKt_u0AeMIDUjJUx/view?usp=sharing

2.3 Columns in the Dataset

- **house_id** – Unique ID of each home
- **sale_date** – Date the home was sold
- **target_price** – Price of the home (only in training set)
- **num_bedrooms** – Number of bedrooms
- **num_bathrooms** – Number of bathrooms
- **living_area** – Square footage of living area
- **lot_area** – Square footage of land lot
- **num_floors** – Number of floors

- **is_waterfront** – Whether the house has a waterfront view (0 = No, 1 = Yes)
- **view_rating** – Rating of view quality (0–4)
- **condition_index** – Condition of the house (1–5)
- **construction_grade** – Grade of construction (1–13)
- **above_area** – Square footage above ground
- **basement_area** – Square footage of basement
- **built_year** – Year the house was built
- **renovated_year** – Year the house was last renovated (0 if never)
- **zip_area** – Zip code of the house
- **latitude** – Latitude coordinate
- **longitude** – Longitude coordinate
- **neighbor_living_area** – Average living area of 15 nearest neighbors
- **neighbor_lot_area** – Average lot area of 15 nearest neighbors

2.4 What You Need to Do

- Data Understanding & Preparation
  - Explore and clean the dataset if necessary using data preprocessing techniques.
- Model Building
  - Train a regression model (or multiple models) using the **train dataset**.
  - Try different algorithms, tuning, or feature selection to get better results.
- Model Evaluation
  - Report training performance using at following regression metrics
    - RMSE (Root Mean Squared Error)
    - MAE (Mean Absolute Error)
    - R^2 score.
  - For **competition scoring,** we will use **only RMSE** for this task. (Lower RMSE = better performance).
- Prediction on Test Data
  - Use your final model to predict prices for the **test dataset**.
  - Save the output in a **CSV file** with the following format.
  - Use the same format for the CSV headings. Otherwise, your submission will not be valid.

- We will use this for evaluate the **testing performance** of your model.

Table 2: Structure of Evaluation CSV for Task 02

house_id	predicted_price
12345	452000
67890	785000

2.5 What You Need to Submit

All submissions must be uploaded to a public GitHub repository. Private repositories or missing files will be considered invalid.

4. **Link for Notebook**
   - Submit your complete notebook (.ipynb) in your GitHub repo.
   - The notebook must include your code, training process, evaluation metrics (RMSE, MAE, R^2), and final model predictions generation.
   - File Naming Rule: Rename your notebook as:
     - TeamCode_TeamName_Task2_HousePrice.ipynb
5. **Training Metrics**
   - Clearly display the evaluation metrics from your training/validation (RMSE, MAE, R^2) in the notebook.
   - You must also include a **screenshot (one picture)** in your repository showing these metric values for verification.
   - Save it as:
     - TeamCode_TeamName_Task2_Metrics.png
6. **Predictions File**
   - CSV file with columns: **house_id, predicted_price** (Table 2**).**
   - This file should contain predictions for the **test dataset** provided.
   - File Naming Rule:
     - TeamCode_TeamName_Task2_Predictions.csv

Important: If you do not follow the above rules (file format, naming conventions, or required submissions), your submission will be considered invalid and will not be evaluated.