Homework 3 Project Proposal

Predict whether a patient has heart diseases or not based on binary classification.
Goal: The objective of this research is to create a predictive model that, using specific health markers from the Behavioral Risk Factor Surveillance System (BRFSS) dataset, would categorize people into risk groups for heart disease. We decided to construct a model to categorize persons who may have heart disease based on their living behavior and present health status using XGBoost, a decision-tree-based with gradient boosting technique.
Dataset Name: Heart Disease Dataset
URL source: https://www.cdc.gov/brfss/index.html
Dataset Size: The BRFSS at the Centers for Disease Control and Prevention (CDC) will be the source of the data for this project. This dataset contains a wide range of different factors (questions) that either directly or indirectly influence heart disease. Based on this, we choose the most significant variables and prepare them for usage in machine learning applications. The preprocessed dataset has fewer variables, which makes its size and complexity more tolerable. This eliminates the requirement for extra GPU resources while training and evaluating models with conventional CPU resources. The project's data preprocessing stage will include a notation on the dataset's precise size.
Outcome of model: Using a confusion matrix, calculate:

1. Accuracy: It measures the overall correctness of the model, how often the model correctly predicts both the presence and absence of heart disease.
2. Precision: It measures the reliability of the model in predicting heart disease. For example, How often the model correctly predicts that a patient has heart disease.
3. Recall: It measures the model's ability to detect all relevant cases of heart disease. It assesses how well the model identifies patients with heart disease.
4. F-1 score: It provides a single score that balances both the precision and recall of the model, which is particularly useful if there is an uneven class distribution.

MORE INFO ON XGBoost:
https://medium.com/@ajuruvictor/how-xgboost-can-save-your-heart-with-94-accuracy-and-8-parameters-a-simple-and-effective-method-7396060c3a35

OTHER SOURCES:
A Gentle Introduction to XGBoost for Applied Machine Learning, which gives an overview of XGBoost and how to use it in Python.
XGBoost Simply Explained (With an Example in Python), which shows how XGBoost works step by step with an example.
XGBoost — GeeksforGeeks, which explains the main concepts and features of XGBoost with diagrams and code snippets.

To efficiently complete your task of creating a predictive model for heart disease risk categorization using the Behavioral Risk Factor Surveillance System (BRFSS) dataset and XGBoost, follow these steps:

1. Data Collection:
   a. Download the Heart Disease Dataset from the CDC's BRFSS website.
2. Data Preprocessing:
   a. Clean and preprocess the dataset, handling missing values and outliers.
   b. Identify relevant health markers such as BMI, blood pressure, and cholesterol levels.
3. Feature Engineering:
   a. Create new features or transform existing ones to enhance model performance.
4. Model Training:
   a. Implement an XGBoost model using libraries like Scikit-Learn or XGBoost itself.
5. Model Evaluation:
   a. Split the dataset into training and testing sets.
   b. Evaluate the model's performance using metrics like accuracy, precision, recall, and F1 score.
6. Hyperparameter Tuning:
   a. Optimize XGBoost hyperparameters for better model performance.
7. Validation:
   a. Validate the model on new data to ensure generalizability.
8. Documentation:
   a. Document the entire process, including data preprocessing, feature selection, model architecture, and evaluation metrics.
9. Comparative Analysis:
   a. Compare the performance of your XGBoost model with other relevant models or algorithms.

Sources:

CDC - BRFSS Dataset
Prediction of cardiovascular disease risk based on major…
An optimized XGBoost based diagnostic system for…
A Heart Disease Prediction Model Based on Feature…
Early Prediction of Heart Disease via LSTM-XGBoost
XGBoost, A Novel Explainable AI Technique, in the…
Cardiovascular Risk Prediction Based on XGBoost

To start a predictive model using Python in Jupyter Notebook:

1. Import Necessary Libraries:
    a. Begin by importing essential libraries such as NumPy, pandas, and scikit-learn for data manipulation and machine learning functionalities.
2. Read and Explore the Dataset:
    a. Use pandas to read your dataset into a DataFrame and explore its structure and features.
3. Feature Selection:
    a. Identify relevant features that will contribute to your predictive model.
4. Build the Model:
    a. Utilize scikit-learn to create and train your predictive model. Choose the appropriate algorithm based on your dataset and task.
5. Evaluate the Model:
    a. Assess the performance of your model using metrics like accuracy, precision, recall, and F1 score
6. Visualization:
    a. Use matplotlib or seaborn for data visualization to better understand the model's insights.
7. Iterate and Improve:
    a. If necessary, iterate on the model, adjust hyperparameters, and enhance feature selection to improve performance.

Sources:

The Definitive Guide to Building a Predictive Model in Python
An Introduction to Predictive Modeling in Python
How to Build a Predictive Model in Python?
Step-by-Step Guide — Building a Prediction Model in Python
Jupyter for Data Science: Making a Prediction Using scikit ...
Build a Predictive Model in 10 Minutes (using Python)

1. Find TARGET variable (y) AKA HeartAttack var.
2. Using remaining variables as independent variables (x)
    a. Then find the most distinct features that mostly affect our target variable to help support our outcome/results
3. Splitting the dataset to training sets and testing sets
4. Model the datasets for comparison evaluation
5. Make a prediction on our test data and evaluate; reduce the parameters and build the new model
6. Commit a conclusion.