**Final Research Report**

**Objective:**  
The objective of this assignment is to apply the machine learning methods learned in class to a new research topic of your choosing. This project will include data selection and analysis, drafting a manuscript-style report, and preparing a video presentation.

**Project Selection:**  
You are required to select a unique research topic. This can be anything of interest to you that involves the application of machine learning. Consider topics that have a direct societal, scientific, or industrial application for maximum impact.

**Data Selection:**  
Identify and choose appropriate datasets to work with for your research project. Your data can be sourced from public repositories or collected directly if it adheres to ethical guidelines and data privacy regulations. Be sure to provide detailed information about your dataset, including where it came from, how it was collected, and what it represents.

**Data Analysis:**  
Carry out a detailed analysis of your dataset using one or more machine learning methods covered in the course. Be sure to document your preprocessing steps, including how you dealt with missing or outlier data. Your analysis should be thorough and aim to answer the research question you've proposed.

**Tables and Figures:**  
Illustrate your analysis with tables and figures to summarize your findings. These can include things like graphs of your data, tables of your results, or diagrams of your machine learning models. All figures and tables should be appropriately labeled, and any statistical analysis should be clearly explained.

**Manuscript-style Report:**  
Your report should follow the structure of a scientific paper, including sections for **Introduction, Literature Review, Methods, Results, and Discussion**. Be sure to clearly describe your research question, the machine learning methods used, your results, and what your results mean in the context of your research question. Make sure your report is thoroughly proofread and formatted consistently.

**Video Presentation:**  
Create a 10 to 12-minute video presentation of your research project. Your presentation should cover the key points from your report, including your research question, the methods you used, your main findings, and why these findings are significant. The video can be a simple slide presentation with voiceover. It should be engaging, easy to understand, and uploaded in a common video format for review.

**Submission:**  
Submit your final manuscript-style report, tables, figures, and video presentation through the course's online platform. Be sure to check that all files are correctly formatted and named appropriately.

**Evaluation:**  
Your project will be evaluated on the selection and application of machine learning methods, the quality of your data analysis, the clarity and completeness of your report, and the presentation of your findings. Be sure to demonstrate a clear understanding of the machine learning methods you've used, and to clearly communicate your findings and their significance in your report and presentation.

**Deadline:**  
All components of your research project must be submitted by the date required on the syllabus. Late submissions may not be accepted or may incur a penalty.

**Note:**  
While collaboration and discussion with classmates are encouraged during the data analysis and method selection stages, the final report and video presentation should be your own original work. Plagiarism will not be tolerated and will result in a failing grade for this assignment.

Good luck, and feel free to reach out with any questions or for help with your projects!

**Introduction**

Heart disease the number one cause of death in the United States. This provides great motivation to discover ways to predict and prevent its development. Heart disease is typically diagnosable via a doctor’s exam, with x-rays and other readings to support symptoms. The factors that contribute to a positive diagnosis are numerous and often interact in complex ways. This presents itself well as a problem to be tackled by machine learning, which can find complex trends within data.

The research question to be answered is: Can machine learning algorithms effectively predict the risk of heart disease based on comprehensive examination data, and what factors contribute significantly to the accuracy and reliability of such predictive models? This will be evaluated in a series of machine learning methods, including logistic regression, random forest, gradient boosting, and artificial neural networks.

**Literature Review**

**Methods**

Data was collected from the CDC NHES I, conducted from 1959-1962 and focused on chronic disease. The research question was addressed using scikit-learn, a public library containing machine learning algorithms used to evaluate the data.

A graph of a network

Description automatically generated The best parameters are {'activation': 'relu', 'alpha': 0.1, 'hidden\_layer\_sizes': (60,), 'learning\_rate': 'constant', 'solver': 'adam'}

Sensitivity: 0.80

Specificity: 0.91

Accuracy: 0.80

A graph of a tree

Description automatically generated

The training set has 5337 samples and the test set has 1335 samples.

Before OverSampling, counts of label '1 (No HD)': 224

Before OverSampling, counts of label '2 (HD)': 5113

After OverSampling, the shape of train\_X: (10226, 163)

After OverSampling, the shape of train\_y: (10226,)

After OverSampling, counts of label 'No Heart Disese': 5113

After OverSampling, counts of label 'Heart Disease': 5113

Fitting 6 folds for each of 280 candidates, totalling 1680 fits

The best parameters are {'criterion': 'entropy', 'max\_depth': None, 'max\_features': None, 'splitter': 'random'}

Sensitivity: 0.98

Specificity: 0.43

Accuracy: 0.95

Gradient Boosting

A graph of a function

Description automatically generated with medium confidence

The training set has 5337 samples and the test set has 1335 samples.

Data split.

Elapsed time: 0.1648714542388916 seconds

Unique values in the column to predict:[2 1]

Before OverSampling, counts of label 1:'Heart Disease': 224

Before OverSampling, counts of label 2:'No Heart Disease': 5113

After OverSampling, counts of label 'Heart Disease': 5113

After OverSampling, counts of label 'No Heart Disease': 5113

Before OverSampling, the shape of train\_X: (5337, 163)

Before OverSampling, the shape of train\_y: (5337,)

After OverSampling, the shape of train\_X: (10226, 163)

After OverSampling, the shape of train\_y: (10226,)

Fitting 6 folds for each of 24 candidates, totalling 144 fits

The best parameters are {'learning\_rate': 0.01, 'max\_depth': 50, 'max\_features': 'log2', 'min\_impurity\_decrease': 0.3, 'min\_samples\_split': 2, 'n\_estimators': 50, 'subsample': 0.3}

Sensitivity: 0.98

Specificity: 0.58

Accuracy: 0.96

Random Forest

A graph of a logistic curve

Description automatically generated

The best parameters are {'max\_depth': 200, 'max\_features': 5, 'min\_samples\_leaf': 3, 'min\_samples\_split': 5, 'n\_estimators': 300}

Sensitivity: 0.99

Specificity: 0.47

Accuracy: 0.96

References:

<https://wwwn.cdc.gov/nchs/nhanes/nhes1/default.aspx>

<https://wwwn.cdc.gov/nchs/data/nhes123/1004.pdf>