Herald College

Concepts and Technologies of AI

5CS037

Final Portfolio Project

Regression and Classification.

January 26, 2024

Contents

1 Assignment Details and Submission Guidelines

1.1 Assignment Details:

Due	Marks	Submission
Feb-13.	40	code notebook.

1.2 Plagiarism and AI Generated Content

Plagiarism of more than 20% and any AI generated content found in the report will be reported for academic misconduct. Thus we highly encourage you to submit your original work.

1.3 Submission Guidelines:

This is an individual task.

1. **Deliverables:**
   - There would be small viva after your presentation, where individual be asked question.
   - You are supposed to submit a code of your task in .ipynb format.

The Final Date for submission is: Feb 13.

1.3.1 Naming Conventions:

You are supposed to follow naming conventions strictly any file not following the naming conventions will be marked "0".

`File Name: WLVID_FullName(firstname+last).ipynb`
`Example: 00000_ABC Sharma.ipynb`

1.3.2 Where to Submit?

Designated Portal Opened at Canvas or as instructed by your instructor.

Please consult with your instructor for more details (very important!).

2 Assignment Overview

2.1 About Assignment:

In this assignment, you will perform a series of task (explained in section 3) for Regression and Classification for a dataset and provide a rigorous rationale for your solutions. We will determine scores by judging both the soundness and cleanliness of your **code**, the quality of the **write-up(report)** and your ability to answer the question during **viva**. Here are examples of aspects that may lead to **point deductions:**

- Use of misleading, unnecessary, or unmotivated graphic elements.
- Unreadable code.
- Missing or incomplete design rationale in write-up.
- Ineffective encoding for your stated goal (e.g., distracting colors, improper data transformation).

Tools and Python Package which can be used for this assignments (listed but not limited to):

1. **Pandas library(pd)**
2. **Numpy library(np)**
3. **Matplotlib library(plt)**
4. **Seaborn library(sns)**
5. **sickit Learn(sklearn)**

2.2 Learning Outcomes:

Learning outcomes can be following but not limited to:

1. Use Pandas as the primary tool to process structured data in Python with CSV files,
2. Extract various information from a given dataset using statistical and visualizing techniques.
3. To be able to build a Machine Learning Model, interpret the design choices for the model.
4. Be able to conduct various experiment on the model and interpret the result of the same.

2.3 Data Selection:

1. Please feel free to pick any structured datasets in csv format that matches the task requirements. But please take pre-approval from your respected instructor and Module leader.
2. If you are not sure about which dataset to pick, select one from the options provided.

The best source to find datasets are but not limited to:

1. Kaggle Datasets:
   Kaggle provides a high-quality dataset in different formats that we can easily find and download.
2. UCI Machine Learning Repository:
   This repository contains databases, domain theories, and data generators that are widely used by the machine learning community for the analysis of ML algorithms.

3 Tasks and Marks Division

3.1 Classification [20]

3.1.1 Choose; Load; Inspect and Explore your Data [5].

- **Pick a Domain and Dataset you are interested in:**
  For this assignment, you will need to find a dataset of your choosing (interest) and load into dataframe object with PANDAS library.

  Perform a initial observation regarding the dataset, while doing that try to answer following question:
  - Detailed description about the dataset:
    1. When and Who created the dataset?
    2. How did you get acess to the dataset?
    3. List out the attributes (columns) of a dataset.
  - Guess some probable question that dataset could answer.
  - Assess the basic fitness of the dataset.
- **Load;Inspect and Explore your Data:**
  Understanding the characteristics of Data beforehand allow us to build a better model with acceptable performance. Before you begin the quest of **building, training and testing** of model, You must write some code in chunks to **check, preview, summarize, explore and visualize** your data.
  1. Load and Check the dataset: After loading the data, it is a good practise to run some checks on it. You must perform the following:
     (a) **Data Cleaning and find the summary statistics of the data.**
     (b) **Explore the data with Visualization and chart.**
     {Do not forget to explain and summarize the chart you opt to build.}

3.1.2 Build Primary Model [5]

Once you have assembled your dataset and gained insights into the key characteristics of your data, it's time to **Build**; **Train**; and **Evaluate** your model. For this task you must do the following:

1. **Split the Dataset into Train and Test set.**
2. **Built at least two machine learning model for Regression Task.**
3. **Evaluate both model on Test Dataset.**
4. **Conclude: Which Model best performed in your dataset?**

3.1.3 Hyper-parameter Optimization with Cross-Validation.[2.5]

Hyper-parameter optimization (aka Hyper-parameter Tuning) is the process of finding the best hyperparameters value for your selected model. In this step you must perform the following:

1. Identify the various hyper-parameters of the model you used in section **3.1.2** {For both the model}.
2. Used any cross-validation techniques to find the best value of hyper-parameters selected above. {**Hint: You can use grid searchCV or randomized searchCV.**
3. **Conclude: The best Hyper-parameters for both the model.**

3.1.4 Feature Selection [2.5]:

In this section you must any one of the feature selection technique discussed on Week-11 Tutorial to select and identify the best features.

3.1.5 Final Model [2.5]:

With the best Hyper-parameters from section **3.1.3** and selected features from **3.1.4**.rebuild both the model from section **3.1.2**.

3.1.6 Conclusion [2.5]:

Please write a brief summary about the outcomes of your experiment. You can explain the following questions:

1. **What was your model performance in section 3.1.3.?**
2. **Did any of the methods you applied {Cross Validation and Feature selection} increased or decreased.**
3. **What did you learn and what could be the future direction?**

3.2 Regression [20]

3.2.1 Choose; Load; Inspect and Explore your Data [5].

- **Pick a Domain and Dataset you are interested in:**
  For this assignment, you will need to find a dataset of your choosing (interest) and load into dataframe object with PANDAS library.

  Perform a initial observation regarding the dataset, while doing that try to answer following question:

  - Detailed description about the dataset:
    1. When and Who created the dataset?
    2. How did you get acess to the dataset?
    3. List out the attributes (columns) of a dataset.
  - Guess some probable question that dataset could answer.
  - Assess the basic fitness of the dataset.

- **Load;Inspect and Explore your Data:**
  Understanding the characteristics of Data beforehand allow us to build a better model with acceptable performance. Before you begin the quest of **building, training and testing** of model, You must write some code in chunks to **check, preview, summarize, explore and visualize** your data.

  1. Load and Check the dataset: After loading the data, it is a good practise to run some checks on it. You must perform the following:
     (a) **Data Cleaning and find the summary statistics of the data.**
     (b) **Explore the data with Visualization and chart.**
     {Do not forget to explain and summarize the chart you opt to build.}

3.2.2 Build Primary Model: [5]

Once you have assembled your dataset and gained insights into the key characteristics of your data, it's time to **Build**; **Train**; and **Evaluate** your model. For this task you must do the following:

1. **Split the Dataset into Train and Test set.**
2. **Built at least two machine learning model for Regression Task.**
3. **Evaluate both model on Test Dataset.**
4. **Conclude: Which Model best performed in your dataset?**

3.2.3 Hyper-parameter Optimization with Cross-Validation: [2.5]

Hyper-parameter optimization (aka Hyper-parameter Tuning) is the process of finding the best hyperparameters value for your selected model. In this step you must perform the following:

1. Identify the various hyper-parameters of the model you used in section **3.1.2** {For both the model}.
2. Used any cross-validation techniques to find the best value of hyper-parameters selected above. {**Hint: You can use grid searchCV or randomized searchCV.**
3. **Conclude: The best Hyper-parameters for both the model.**

3.2.4 Feature Selection [2.5]:

In this section you must any one of the feature selection technique discussed on Week-11 Tutorial to select and identify the best features.

3.2.5 Final Model [2.5]:

With the best Hyper-parameters from section **3.1.3** and selected features from **3.1.4**.rebuild both the model from section **3.1.2**.

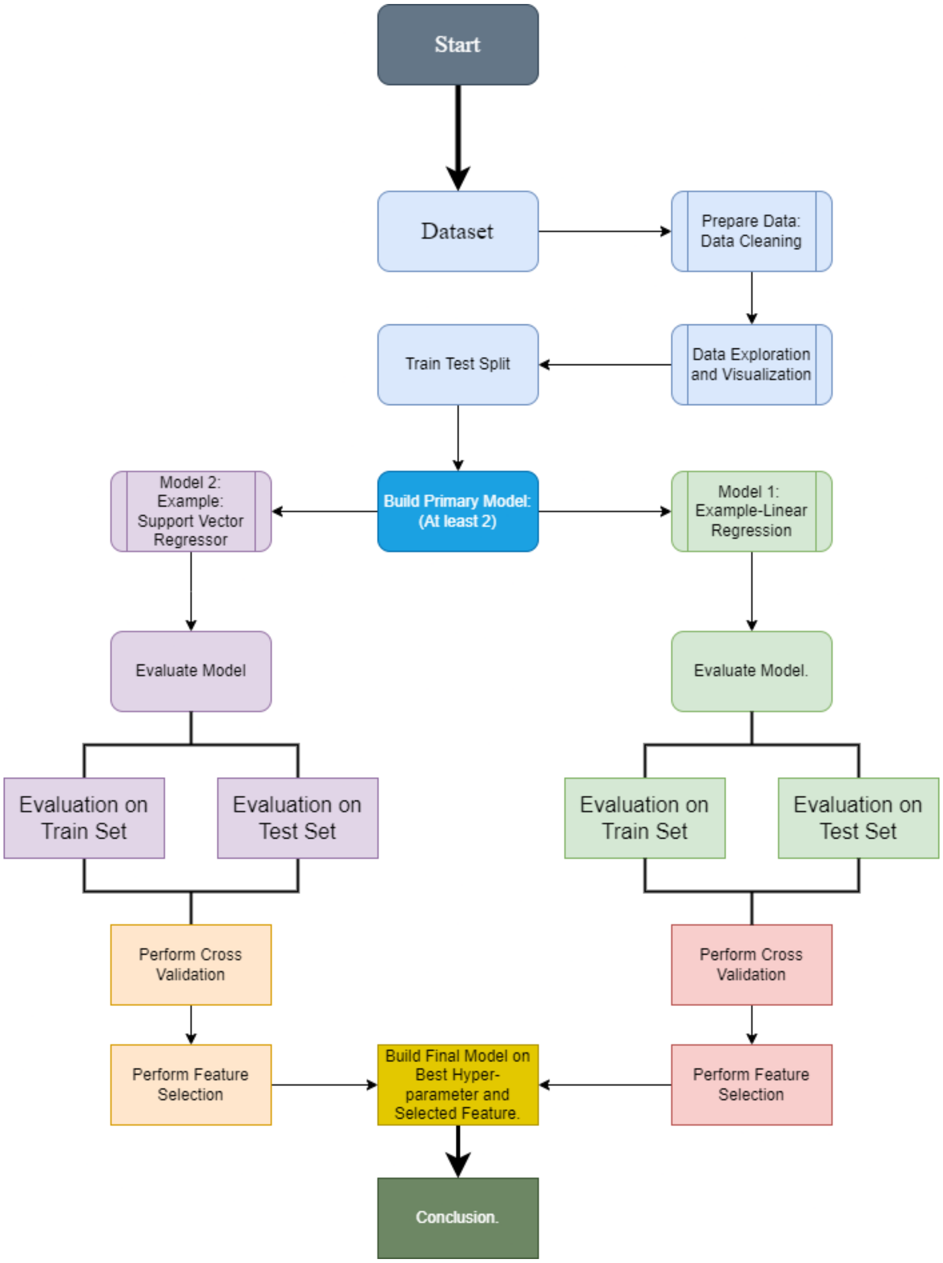
3.2.6 Conclusion [2.5]:

Please write a brief summary about the outcomes of your experiment. You can explain the following questions:

1. **What was your model performance in section 3.1.3.?**
2. **Did any of the methods you applied** {**Cross Validation and Feature selection**} **increased or decreased.**
3. **What did you learn and what could be the future direction?**

4 Task-Flow Diagram:

Repeat for both the Regression and Classification.


Start
Dataset
Prepare Data: Data Cleaning
Data Exploration and Visualization
Train Test Split
Build Primary Model: (At least 2)
Model 2: Example: Support Vector Regressor
Model 1: Example-Linear Regression
Evaluate Model
Evaluate Model.
Evaluation on Train Set
Evaluation on Test Set
Evaluation on Train Set
Evaluation on Test Set
Perform Cross Validation
Perform Cross Validation
Perform Feature Selection
Build Final Model on Best Hyper-parameter and Selected Feature.
Perform Feature Selection
Conclusion.