Herald College, Kathmandu

Concepts and Technologies of AI

5CS037

Final portfolio Project.

An End - to - End Machine Learning Project
on Regression and Classification Task.

Jan 21, 2025

Contents

1 Assignment Details and Submission Guidelines

1.1 Assignment Details:

Due	Marks	Submission
Feb - 11.	50	A Report and 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 project must be completed individually.
- The data set used for this project must be preapproved from your respected Module Leader.

- What to Submit?
  - You are expected to submit a report based on the task { One for Regression and One for Classification Task} and associated code file.
  - For Code:
    1. All solutions - Code must be written in the Jupyter notebook.
    2. All codes must be pushed to `GitHub` before the deadlines.
  - For report:
    1. Please follow the APA format; for a sample, see Section 3 of this document.
  - Where to submit?
    Designated portal opened on Canvas.
- After Submission:
  - You are expected defend your work after the submission.Please consult with your respected instructor for date and time of the viva.

The Final Date for submission is: **`11 Feb.`**

1.3.1 Naming Conventions:

You are supposed to strictly follow the naming conventions, and any file that does not follow the naming conventions will be marked as "0".
`File Name: WLVID_FullName(firstname+last).ipynb`, `File Name: WLVID_FullName(firstname+last).p`

2 Assignment Overview

2.1 About Assignment:

In this assignment, you will undertake a comprehensive end-to-end machine learning project designed to deepen your understanding of the entire ML pipeline, from data preprocessing to model building and evaluation. This assignment integrates knowledge from all the lessons covered during this module, challenging you to apply it to both regression and classification tasks using real-world datasets.

2.2 Cautions!!!

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)**

2.3 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.

2.4 Dataset 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.

3 Tasks - To - Do:

Please Complete all the Tasks as instructed.

3.1 For Classification Task [25]:

1. Exploratory Data Analysis and Data Understanding [5]:

1. Choosing a Dataset:
   - Select a dataset of your choice that interests you and aligns with one of the United Nations Sustainable Development Goals (UNSDG).
   - Load the dataset into a DataFrame object using the Pandas library.
   - Perform an initial analysis to gather a detailed description of the dataset. For example:
     (a) When and by whom was the dataset created?
     (b) How did you access the dataset?
     (c) How does it align with the chosen UNSDG?
     (d) List all the attributes (columns) present in the dataset.
   - Identify potential questions that the dataset could help answer.
   - Assess the dataset's suitability for analysis (e.g., data completeness, relevance, and quality).
2. Conducting Exploratory Data Analysis (EDA):
   - Understanding the characteristics of the data beforehand is crucial for building a model with acceptable performance. Before proceeding to **build, train, and test** the model, write code to **inspect, preview, summarize, explore, and visualize** your data. For example:
     (a) Perform data cleaning and compute summary statistics for the dataset.
     (b) Explore the data through visualizations and charts. Ensure you explain and summarize the insights gained from each chart.

2. Build a Model from Scratch [5]:

For Classification Task, build a Appropriate Logistic Regression {Sigmoid or Softmax} from scratch, and report the appropriate evaluation metrics on train and test set.

3. Build a Primary Model [5]:

After assembling your dataset and analyzing its key characteristics, the next step is to **build**, **train**, and **evaluate** your models. Follow the steps below to complete this task:

1. **Split the dataset into training and testing sets.**
2. **Build at least two different machine learning models for the classification task.**

3. **Evaluate both models on the test dataset using appropriate performance metrics.**
4. **Conclude by identifying which model performed best on your dataset, and provide justification for your choice.**

4. Hyper-parameter Optimization with Cross-Validation [2.5]:

Hyper-parameter optimization, also known as hyper-parameter tuning, is the process of identifying the best hyper-parameter values for your selected models. Follow the steps below to perform this task:

1. Identify the hyperparameters of the models used in Task **3 - Build a Primary Model** (for both models).
2. Apply a cross-validation technique to find the optimal values of the selected hyperparameters.
   - **Hint:** You can use techniques like **GridSearchCV** or **RandomizedSearchCV**.
3. Conclude by summarizing the best hyperparameters for both models.

5. Feature Selection [2.5]:

In this section, apply one of the feature selection techniques discussed in the Week-10 tutorial to identify and select the most relevant features for your models. Clearly document your process and justify your choice of features.

6. Final Model [2.5]:

Using the optimal hyperparameters identified in **Task - 4** and the selected features from **Task - 5**, rebuild both models from **Task - 3**. Evaluate the performance of the final models and provide a summary of your findings.

7. Conclusion [2.5]:

Provide a brief summary of the outcomes of your experiment by addressing the following points:

1. **Model Performance:** How did your models perform in? Discuss the key results and metrics.
2. **Impact of Methods:** Analyze the effect of the techniques you applied, such as **Cross-Validation** and **Feature Selection**. Did these methods improve or reduce model performance? Provide a brief explanation.
3. **Insights and Future Directions:** Reflect on what you learned from the experiment. What insights can be drawn from your analysis, and what potential improvements or extensions could be explored in future work?

3.2 For Regression Task [25]:

1. Exploratory Data Analysis and Data Understanding [5]:

1. Choosing a Dataset:
   - Select a dataset of your choice that interests you and aligns with one of the United Nations Sustainable Development Goals (UNSDG).
   - Load the dataset into a DataFrame object using the Pandas library.
   - Perform an initial analysis to gather a detailed description of the dataset. For example:
     (a) When and by whom was the dataset created?
     (b) How did you access the dataset?
     (c) How does it align with the chosen UNSDG?
     (d) List all the attributes (columns) present in the dataset.
   - Identify potential questions that the dataset could help answer.
   - Assess the dataset's suitability for analysis (e.g., data completeness, relevance, and quality).
2. Conducting Exploratory Data Analysis (EDA):
   - Understanding the characteristics of the data beforehand is crucial for building a model with acceptable performance. Before proceeding to **build, train, and test** the model, write code to **inspect, preview, summarize, explore, and visualize** your data. For example:
     (a) Perform data cleaning and compute summary statistics for the dataset.
     (b) Explore the data through visualizations and charts. Ensure you explain and summarize the insights gained from each chart.

2. Build a Model from Scratch [5]:

For the Regression task, build an appropriate model from scratch, such as Linear Regression, and report the appropriate evaluation metrics on the train and test sets (e.g., Mean Squared Error, R-squared).

3. Build a Primary Model [5]:

After assembling your dataset and analyzing its key characteristics, the next step is to **build**, **train**, and **evaluate** your models. Follow the steps below to complete this task:

1. **Split the dataset into training and testing sets.**
2. **Build at least two different machine learning models for the regression task.**
3. **Evaluate both models on the test dataset using appropriate performance metrics (e.g., Mean Absolute Error, Root Mean Squared Error, R-squared).**
4. **Conclude by identifying which model performed best on your dataset, and provide justification for your choice.**

4. Hyper-parameter Optimization with Cross-Validation [2.5]:

Hyper-parameter optimization, also known as hyper-parameter tuning, is the process of identifying the best hyper-parameter values for your selected models. Follow the steps below to perform this task:

1. Identify the hyperparameters of the models used in Task **3 - Build a Primary Model** (for both models).
2. Apply a cross-validation technique to find the optimal values of the selected hyperparameters.
   - **Hint:** You can use techniques like **GridSearchCV** or **RandomizedSearchCV**.
3. Conclude by summarizing the best hyperparameters for both models.

5. Feature Selection [2.5]:

In this section, apply one of the feature selection techniques discussed in the Week-10 tutorial to identify and select the most relevant features for your models. Clearly document your process and justify your choice of features.

6. Final Model [2.5]:

Using the optimal hyperparameters identified in **Task - 4** and the selected features from **Task - 5**, rebuild both models from **Task - 3**. Evaluate the performance of the final models and provide a summary of your findings.

7. Conclusion [2.5]:

Provide a brief summary of the outcomes of your experiment by addressing the following points:

1. **Model Performance:** How did your models perform? Discuss the key results and metrics (e.g., Mean Squared Error, R-squared, etc.).
2. **Impact of Methods:** Analyze the effect of the techniques you applied, such as **Cross-Validation** and **Feature Selection**. Did these methods improve or reduce model performance? Provide a brief explanation.
3. **Insights and Future Directions:** Reflect on what you learned from the experiment. What insights can be drawn from your analysis, and what potential improvements or extensions could be explored in future work?

4 Task - Flow for Final Portfolio Project:

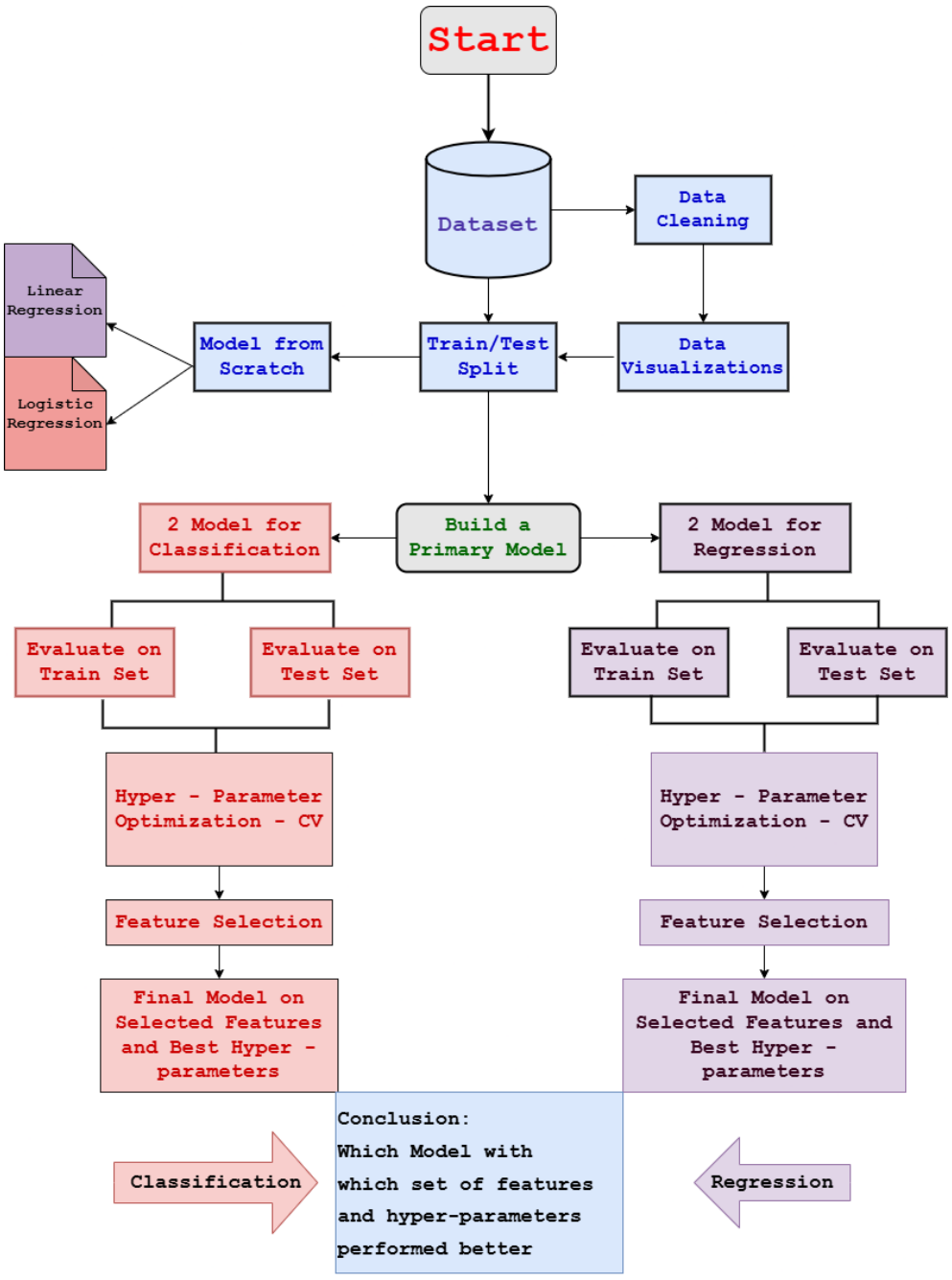

Figure 1: Task Flow Diagram

5 Report Guidelines.

There are no specific format of what report should be, feel free to use your imagination to make it better. Before you make a submission please make sure following are covered:

1. General Guidelines

- Include the College and University approved Cover and Title Page.
- Formatting: Use clear headings for each section and subsection.
- Visualizations: Ensure all plots are appropriately labeled and titled, with concise captions.
- Language: Maintain a formal, academic tone throughout the report.
- Submission Requirements:
  - Submit a well-organized PDF report.
  - Attach the Jupyter notebook with all code, comments, and rendered outputs.
  - Ensure all tasks align with the report content.
  - Ensure you save or screenshot a copy of plagiarism report and ask your respective instructor to verify with a signature and keep your report save till the end of semester.{Please be reminded Plagiarism allowed is 20% only and any AI detected content will be not accepted for submission.}

Classification Analysis Report

Warning: This is just a template. Please modify it as needed for your report.

Abstract

Purpose: The purpose of this report is to predict a categorical variable using classification techniques.
Approach: The dataset chosen for this analysis is [Dataset Name], which contains [Dataset Description]. The steps involved include Exploratory Data Analysis (EDA), model building with [e.g., Logistic Regression or Decision Trees], hyper-parameter optimization, and feature selection.
Key Results: The performance of the models was evaluated using accuracy, precision, recall, and F1-score. The models showed [Model Performance].
Conclusion: The classification models performed [Outcome of Model], and key insights include [Conclusion].

1 Introduction

1.1 Problem Statement

The goal of this project is to predict a categorical target variable. For instance, the problem could involve classifying email messages as spam or not spam, predicting disease presence, or any other classification task based on the available dataset.

1.2 Dataset

The dataset used in this analysis is [Dataset Name], which was obtained from [Source]. It contains [Brief Description of Data]. This dataset aligns with the United Nations Sustainable Development Goals (UNSDG) by [Link to UNSDG].

1.3 Objective

The objective of this analysis is to build a predictive classification model that estimates the target categorical variable based on the given features in the dataset.

2 Methodology

2.1 Data Preprocessing

Before building the model, the data was cleaned by [Handling Missing Values, Outliers, or Inconsistent Data]. Additionally, transformations such as [Scaling/Normalization] were performed to prepare the data for analysis.

2.2 Exploratory Data Analysis (EDA)

EDA was performed using visualizations such as histograms, bar charts, and correlation matrices to better understand the data. Key insights from EDA include [Key Insights].

2.3 Model Building

Two classification models were considered for this task: [Model 1, e.g., Logistic Regression] and [Model 2, e.g., Decision Trees]. The model was built by [Steps Taken to Build the Model, e.g., splitting the data into training and testing sets, training the model].

2.4 Model Evaluation

The model's performance was evaluated using the following metrics:

- **Accuracy**: The proportion of correct predictions over the total predictions.
- **Precision**: The proportion of positive predictions that are actually correct.
- **Recall**: The proportion of actual positive cases that are correctly identified.
- **F1-Score**: The harmonic mean of precision and recall.

These metrics were selected because they are commonly used to evaluate classification models, especially when dealing with imbalanced classes.

2.5 Hyper-parameter Optimization

To improve the performance of the model, hyper-parameter optimization was conducted using [GridSearchCV or RandomizedSearchCV]. The optimal parameters for the model were found to be [List of Optimal Parameters].

2.6 Feature Selection

Feature selection was performed using [Technique Used, e.g., Recursive Feature Elimination (RFE)] to identify the most important features for predicting the target variable. The selected features were [List of Selected Features].

3 Conclusion

3.1 Key Findings

The model's performance on the test dataset was evaluated using [Evaluation Metrics]. The results showed [Key Findings of the Model].

3.2 Final Model

The final model, based on [Selected Model], was the most effective at predicting the target variable. The model achieved [Key Result from the Evaluation Metrics].

3.3 Challenges

During the project, several challenges were encountered, including [Challenges Faced, e.g., data quality issues, feature selection difficulties].

3.4 Future Work

To improve the model, future work could involve [Suggestions for Improvement, e.g., using more advanced classification algorithms, optimizing feature selection].

4 Discussion

4.1 Model Performance

The model's performance was evaluated using [Metrics Used]. The results indicate that the model [Performance Discussion, e.g., performed well or poorly on the test data].

4.2 Impact of Hyperparameter Tuning and Feature Selection

Hyperparameter tuning and feature selection played a crucial role in improving model performance. After applying these techniques, [Results of Improvement].

4.3 Interpretation of Results

The chosen features and model performed in a manner consistent with expectations. The key insights from the results suggest that [Interpretation of Results].

4.4 Limitations

Despite the successful modeling, some limitations remain in the approach, such as [Limitations, e.g., limited dataset size, assumptions in the model].

4.5 Suggestions for Future Research

Future research could explore [Suggestions for Future Research, e.g., experimenting with different classification algorithms, increasing the dataset size, implementing feature engineering].

Regression Analysis Report

Warning: This is just a template. Please modify it as needed for your report.

Abstract

Purpose: The purpose of this report is to predict a continuous variable using regression techniques.
Approach: The dataset chosen for this analysis is [Dataset Name], which contains [Dataset Description]. The steps involved include Exploratory Data Analysis (EDA), model building with [e.g., Linear Regression], hyper-parameter optimization, and feature selection.
Key Results: The performance of the model was evaluated using R-squared and Mean Squared Error (MSE). The model showed [Model Performance].
Conclusion: The regression model performed [Outcome of Model], and key insights include [Conclusion].

1 Introduction

1.1 Problem Statement

The goal of this project is to predict a continuous target variable. For instance, the problem could involve predicting house prices, stock prices, or any other continuous variable based on the available dataset.

1.2 Dataset

The dataset used in this analysis is [Dataset Name], which was obtained from [Source]. It contains [Brief Description of Data]. This dataset aligns with the United Nations Sustainable Development Goals (UNSDG) by [Link to UNSDG].

1.3 Objective

The objective of this analysis is to build a predictive regression model that estimates the target variable based on the given features in the dataset.

2 Methodology

2.1 Data Preprocessing

Before building the model, the data was cleaned by [Handling Missing Values, Outliers, or Inconsistent Data]. Additionally, transformations such as [Scaling/Normalization] were performed to prepare the data for analysis.

2.2 Exploratory Data Analysis (EDA)

EDA was performed using visualizations such as scatter plots, histograms, and summary statistics to better understand the data. Key insights from EDA include [Key Insights].

2.3 Model Building

Two regression models were considered for this task: [Model 1, e.g., Linear Regression] and [Model 2, e.g., Decision Trees]. The model was built by [Steps Taken to Build the Model, e.g., splitting the data into training and testing sets, training the model].

2.4 Model Evaluation

The model's performance was evaluated using the following metrics:

- **R-squared**: To measure the proportion of the variance in the dependent variable explained by the independent variables.
- **Mean Squared Error (MSE)**: To measure the average squared difference between the actual and predicted values.

These metrics were selected because they are commonly used to evaluate regression models.

2.5 Hyper-parameter Optimization

To improve the performance of the model, hyper-parameter optimization was conducted using [GridSearchCV or RandomizedSearchCV]. The optimal parameters for the model were found to be [List of Optimal Parameters].

2.6 Feature Selection

Feature selection was performed using [Technique Used, e.g., Recursive Feature Elimination (RFE)] to identify the most important features for predicting the target variable. The selected features were [List of Selected Features].

3 Conclusion

3.1 Key Findings

The model's performance on the test dataset was evaluated using [Evaluation Metrics]. The results showed [Key Findings of the Model].

3.2 Final Model

The final model, based on [Selected Model], was the most effective at predicting the target variable. The model achieved [Key Result from the Evaluation Metrics].

3.3 Challenges

During the project, several challenges were encountered, including [Challenges Faced, e.g., data quality issues, feature selection difficulties].

3.4 Future Work

To improve the model, future work could involve [Suggestions for Improvement, e.g., using more advanced regression algorithms, optimizing feature selection].

4 Discussion

4.1 Model Performance

The model's performance was evaluated using [Metrics Used]. The results indicate that the model [Performance Discussion, e.g., performed well or poorly on the test data].

4.2 Impact of Hyperparameter Tuning and Feature Selection

Hyperparameter tuning and feature selection played a crucial role in improving model performance. After applying these techniques, [Results of Improvement].

4.3 Interpretation of Results

The chosen features and model performed in a manner consistent with expectations. The key insights from the results suggest that [Interpretation of Results].

4.4 Limitations

Despite the successful modeling, some limitations remain in the approach, such as [Limitations, e.g., limited dataset size, assumptions in the model].

4.5 Suggestions for Future Research

Future research could explore [Suggestions for Future Research, e.g., experimenting with different regression algorithms, increasing the dataset size, implementing feature engineering].