Table of Contents

[Project Overview 1](#_Toc160091380)

[Problem Statement 1 1](#_Toc160091381)

[Question 1](#_Toc160091382)

[Code Snippet 1](#_Toc160091383)

[Output 1](#_Toc160091384)

[Problem Statement 2 2](#_Toc160091385)

[Question 2](#_Toc160091386)

[Code Snippet 2](#_Toc160091387)

[Output 2](#_Toc160091388)

[Problem Statement 3 3](#_Toc160091389)

[Question 3](#_Toc160091390)

[Design Process Procedure 3](#_Toc160091391)

[Considered MLPClassifier tradeoffs 4](#_Toc160091392)

[Problem Statement 4 4](#_Toc160091393)

[Question 4](#_Toc160091394)

[Code Snippet 4](#_Toc160091395)

[Output 4](#_Toc160091396)

[Problem Statement 5, 6 & 7 5](#_Toc160091397)

[Question 5](#_Toc160091398)

[Solution 5](#_Toc160091399)

[Observations 6](#_Toc160091400)

[Problem Statement 8 6](#_Toc160091401)

[Question 6](#_Toc160091402)

[Solution 6](#_Toc160091403)

[Observations 7](#_Toc160091404)

[Project 2 Conclusion 8](#_Toc160091405)

[Appendix 9](#_Toc160091406)

[Code 1 9](#_Toc160091407)

[Code 2 10](#_Toc160091408)

**List of Tables**

[Table 1: Code 1- Split of Data into Training, Validation and Testing with values. 2](#_Toc160092506)

[Table 2: Code 1 - Cross-Validation Scores for each iteration and the average. 3](#_Toc160092507)

[Table 3: Code 1 - Input vs Output values comparison along with red highlight (worst performing scenario) and green highlight (Best performing scenario). 6](#_Toc160092508)

[Table 4: Code 2 - Split of Data into Training, Validation and Testing with values. 8](#_Toc160092509)

[Table 5: Input vs Output values comparison along with red highlight (worst performing scenario) and green highlight (Best performing scenario). 8](#_Toc160092510)

# Project Overview

The project 3 aims to expand learning through practice upon the concepts of Neural Network multilayer perceptron learning algorithms with focus upon MLPClassifier and the MLPRegressor.

# Problem Statement 1

## Question

The expectation is to write a python code that splits the Wisconsin Breast Cancer data set available in the *Sklearn* dataset library into Training/Validation and Testing subsets through a split ratio of 80% and 20% respectively. The key objective is to mention and describe how the split is conducted along with providing the random value used for the split for the purpose of recreation.

## Code Snippet

The following is the code snippet for the question given:



The entire code block for the Problem Statement 1 can be found in the appendix within the Code 1 of this document.

## Output



The above output clearly identifies the total number of samples in the Wisconsin Breast Cancer Dataset which is 569 and the number of samples divided for Training and Validation is 455 i.e. 80% (79.96%). Additionally, the Training and Validation is further split in half with 227 samples shared with Training and 228 samples shared with Validation. The number of samples divided for Testing is 114 i.e. 20% (20.03%). The random state value provided for the split is given as 19.

|  |  |  |
| --- | --- | --- |
| **Category** | **Percentage split** | **Number of rows (samples)** |
| Training and Validation | 80% | 455 (227+228) |
| Training | 50% of 80% | 227 |
| Validation | 50% of 80% | 228 |
| Testing | 20% | 114 |

Table 1: Code 1- Split of Data into Training, Validation and Testing with values.

# Problem Statement 2

## Question

The requirement here is to either create an additional split for the validation sub-dataset or to use cross validation technique to tune the overall program’s MLP Hyperparameters.

## Code Snippet

The following is the code snippet for the question given:



The entire code block for the Problem Statement 2 can be found in the appendix within the Code 1 of this document.

## Output



The output at first displays the cross-validation scores with a cross validation count of number of folds is given as 10. This means that the validation subset is divided into 10 equal folds and each fold is compared with the remaining 9 other to evaluate the model’s performance based on 10 iterations (in this case). The average value of these 10-fold iterations is also displayed and the value obtained here is 0.9164. This shows that the model is already leaning on a good learning curve.

|  |  |
| --- | --- |
| **Iteration Count** | **Cross Validation Scores** |
| 1 | 0.97826087 |
| 2 | 0.91304348 |
| 3 | 0.93478261 |
| 4 | 0.82608696 |
| 5 | 0.93478261 |
| 6 | 0.91111111 |
| 7 | 0.91111111 |
| 8 | 0.95555555 |
| 9 | 0.86666667 |
| 10 | 0.93333333 |
| **Average** | **0.9164734299516908** |

Table 2: Code 1 - Cross-Validation Scores for each iteration and the average.

# Problem Statement 3

## Question

The expectation with this question is to provide a procedure to document the design process and to identify and mention the tradeoffs considered in using a MLPClassifier.

## Design Process Procedure

The following are the general steps involved (along with short descriptions in relation to the Project 3 fulfillment) in the establishment of design as a precursor to the MLPClassifier:

1. Problem Definition:

In this case, the problems are clearly defined in what is to be achieved along with the necessary inputs (Ex. Choice of Dataset), procedural steps and the necessary outputs (Ex. MLP and Cross Validation scores).

1. Data Collection and Preprocessing:

This step involved prepping the data (cleaning, normalization etc.) before operations can be done using it. However, in this case, since we are already using a prepped dataset such as the Wisconsin Breast Cancer Dataset from the Sklearn dataset library, it is directly used for operations.

1. Model Architecture:

The model architecture focusses on the finer details of how the data is to be analyzed based on a certain set of hyperparameters such as number of layers with the number of neurons in each layer along with the number of epochs that is to be conducted et al.

1. Training and Hyperparameter tuning:

This stage of the process trains the model from the split training sub-dataset and tuning the hyperparameters such as the number of neurons, learning rate etc. The usage of cross validation technique is also considered for hyperparameter to achieve the best score for the model.

1. Model Evaluation:

This step focusses on the model’s performance upon the unseen data i.e. testing sub-dataset. This evaluation is done by using metrics such as scores and the expected efficiency is obtained based on the problem statement’s requirement.

1. Documentation

This stage is a vital step of the process since it involves documenting the entire process procedure, the decisions made and the tuning of the hyperparameters to achieve desired results and to satisfy the problem statement. This step also provides an insight upon the entire activity conducted from the start till the end.

## Considered MLPClassifier tradeoffs

1. Increasing the complexity of a model does improve the model’s efficiency and its ability to solve complex problems, however, tuning the complexity of a model greater than a limit causes overfitting in the model.
2. Increasing the number of layers or the number of neurons improves the performance of the model but also increases the training time for any dataset.
3. MLP Classifiers are black box systems due to their complexity and thus leaves little room for interpretation and decision making in complex scenarios. However, the replacement for MLP Classifiers can either be linear classifiers or decision trees which can interpret much easily. But the performance of an MLP Classifier is unmatched.
4. Data Availability affects the model complexity as well i.e. with abundant labelled examples from the dataset, complex models like MLP Classifiers can be used else there is a risk of overfitting when the dataset isn’t large and unlabeled.

# Problem Statement 4

## Question

The requirement here is to use an MLP Classifier to train, validate and test an MLP Model. Any number of features from the dataset can be used. The random state value is to be mentioned for the purpose of model recreation.

## Code Snippet



The entire code block for the Problem Statement 4 can be found in the appendix within the Code 1 of this document.

## Output



It can be observed that the output prints the score of all the all the three subsets separately.

# Problem Statement 5, 6 & 7

## Question

The requirement for problem statement 5 is to provide a list of hyperparameters that were used as default values as input along with the new values that were modified from the original values. The requirement for problem statement 6 is to provide a score values of each dataset (Training, Validation and Testing). The requirement for problem statement 7 is to note the observations along with pointers for improvement.

## Solution



Table 3: Code 1 - Input vs Output values comparison along with red highlight (worst performing scenario) and green highlight (Best performing scenario).

## Observations

The model shows the consistency in the scores among average cross validation, training, validation, and testing scores with the varying input hyperparameters accordingly. The best-case scenario can be observed in Table 3 where the output testing score is 0.947. The worst-case scenario can be observed in Table 3 where the output testing score is 0.754.

The feedback on pointers for improvement would be to increase the number of iterations in the cross validation as that would impact the average cross validation and the validation scores, which in turn would make the model robust. Along with this, an increase to the ratio of number of neurons per layer in the MLP Classifier would significantly make the mode further complex but also enhance the chances for a better case scenario.

It was also observed that the increase to the number of epochs would not directly impact towards a best-case scenario, rather it would streamline the results and push the output towards a further normalized data as seen in case 4 with 10000 epochs from Table 3.

# Problem Statement 8

## Question

The requirement for problem statement 8 is to repeat the steps performed for Code 1 with change in use of MLP Regressor in place of a MLP Classifier along with adding a shuffle to the dataset before the operations are done on it. Furthermore, the same inputs, outputs and observations are to be recorded.

## Solution

The following is the output obtained for using a MLP Regressor upon the Diabetes dataset present in SK Learn dataset following the same procedure as Code 1:



The entire code block for the Problem Statement 8 can be found in the appendix within the Code 2 of this document.

The output shows that the *Diabetes* dataset contains a total of 442 samples of rows of data among which 353 i.e. 80% (79.86%) of the original data is further divided into Training and Validation at 50% each i.e. 176 and 177 samples respectively. The testing sub-dataset contains a sample of 89 rows of data which is 20% (20.13%) of the original data.

|  |  |  |
| --- | --- | --- |
| **Category** | **Percentage split** | **Number of rows (samples)** |
| Training and Validation | 80% | 353 (176+177) |
| Training | 50% of 80% | 176 |
| Validation | 50% of 80% | 177 |
| Testing | 20% | 89 |

Table 4: Code 2 - Split of Data into Training, Validation and Testing with values.

The output also displays the cross-validation scores with a cross validation count of number of folds is given as 4. This means that the validation subset is divided into 4 equal folds and each fold is compared with the remaining 3 other to evaluate the model’s performance based on 4 iterations (in this case). The average value of these 4-fold iterations is also displayed and the value obtained here is -2.9827. This shows that the model is either failing due to the mismatch in the decision chosen for the original subset used for the pertaining methodology of approach of operations.

|  |  |
| --- | --- |
| **Iteration Count** | **Cross Validation Scores** |
| 1 | -3.23383008 |
| 2 | -3.1745197 |
| 3 | -2.36219969 |
| 4 | -3.16064332 |
| **Average** | **-2.982798197130794** |

Table 5: Code2 - Cross-Validation Scores for each iteration and the average.

The Table 5 shows the cross-validation scores of each iteration performed.



Table 6: Code 2 - Input vs Output values comparison along with red highlight (worst performing scenario) and green highlight (Best performing scenario).

## Observations

The model shows the consistency in the poor scores among average cross validation. The scores of training, validation, and testing scores with the varying input hyperparameters accordingly with the Testing scores not aligning with the model’s learning and validation in all the scenarios. The best-case scenario can be observed in Table 6 where the output testing score is 0.262. The worst-case scenario can be observed in Table 6 where the output testing score is -1.162.

The feedback on pointers for improvement would be to change the model from the source as clearly the model shows that MLP Regressor isn’t the best fit for the dataset chosen. Along with this, an increase to the ratio of number of neurons per layer in the MLP Classifier would significantly make the mode further complex but also enhance the chances for a better case scenario.

It was also observed that the increase to the number of epochs would not directly impact towards a best-case scenario, rather it would streamline the results and push the output towards a further normalized data as seen in case 4 with 10000 epochs from Table 3.

# Project 3 Conclusion

Project 2 has been the most challenging task that has been endured so far. The convolution in understanding and further continued learnings from each problem statement has not only been frustrating but also fruitful towards understanding each model and the concept of splitting data for Training, Validation and Testing in Machine Learning. The problems revolve only around this splitting concept and helps one solidify their reasoning upon the knowledge of these Machine Learning concepts.

# Appendix

## Code 1



## Code 2

