HW2

Data Engineering Step

1. Loading the training dataset and test dataset
2. Review training dataset and test dataset then find issues.

   Sex and cp are characters in training dataset, numbers in test dataset.

   Null data in training dataset.

   Cp, Ca, oldpeak, and target std are larger then mean. But I will not remove the Cp, Ca, and oldpeak, and target because those data are still information.
3. Understand what is an unreasonable value.

   年齡 (age) 醫學合理範圍：20 至 100 歲 不合理數值： 小於 20 歲：一般而言，此類心臟病資料集多針對成人，小於 20 歲較少見。 大於 100 歲：超過 100 歲的患者較罕見且可能為資料錯誤。

   靜息血壓 (trestbps) 醫學合理範圍：80 至 220 mmHg 不合理數值： 小於 80 mmHg：顯著低血壓，很可能資料輸入錯誤。 大於 220 mmHg：極高的血壓值少見於臨床常態，可能為量測或輸入錯誤。

   膽固醇 (chol) 醫學合理範圍：100 至 600 mg/dl 不合理數值： 小於 100 mg/dl：過低的膽固醇值極為少見，可能表示資料錯誤或檢驗失誤。 大於 600 mg/dl：極端高膽固醇值在臨床上少見，應視為異常數據。

   最大心跳率 (thalach) 醫學合理範圍：60 至 220 bpm 不合理數值： 小於 60 bpm：極低的最大心跳率在一般狀況下不太可能，需檢查資料正確性。 大於 220 bpm：超過人類生理極限，為明顯不合理資料。

   心電圖 ST 段下降 (oldpeak) 醫學合理範圍：0 至 6（一般臨床數值） 不合理數值： 小於 0：不合理，心電圖 ST 下降不可能為負數。 大於 6：臨床上極端少見且幾乎不合理，可能為錯誤數據。

   主要血管數量 (ca) 醫學合理範圍：0 至 4 不合理數值： 小於 0 或 大於 4：解剖結構上不可能，因冠狀動脈主要數量通常為 0 到 4 之間。

   thal 欄位（地中海貧血基因型） 醫學合理數值：1（正常）, 2（固定缺損）, 3（可逆缺損） 不合理數值： 不在 1、2、3 範圍內（例如負數、0 或大於 3）：不符醫學定義，視為錯誤。
4. Transfer data type.

   train_data['sex'].map({'Male': 0, 'Female': 1})

   train_data['cp'].map({'low': 0, 'medium': 1, 'high': 2, 'severe': 3})

5. Fill the training/test dataset numerical data with the median.
6. The training/test dataset category data is filled with the majority.
7. Defining medically reasonable limits and removing unreasonable or abnormal data.
8. Check whether there are still null values in the cleaned data.
9. Confirm that the data types are all numerical and consistent with the test data.
10. Review the statistics of each training/test dataset field after data cleaning.

Training

```
RangeIndex: 272 entries, 0 to 271
Data columns (total 14 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   age       272 non-null    int64
 1   sex       272 non-null    float64
 2   cp        272 non-null    float64
 3   trestbps  272 non-null    float64
 4   chol      272 non-null    float64
 5   fbs       272 non-null    int64
 6   restecg   272 non-null    float64
 7   thalach   272 non-null    float64
 8   exang     272 non-null    int64
 9   oldpeak   272 non-null    float64
 10  slope     272 non-null    float64
 11  ca        272 non-null    int64
 12  thal      272 non-null    float64
 13  target    272 non-null    float64
dtypes: float64(10), int64(4)
```

Defining Neural Networks

11. Define NN model and considering of:

    Using larger NN model for better capability.

    Use LeakyReLU avoids the dead ReLU.

    Use dropout to avoid overfitting and maintains the appropriate ratio to avoid underfitting.

    Use batchnorm with order preservation after nonlinear activation.

    Use Kaiming initialization method.

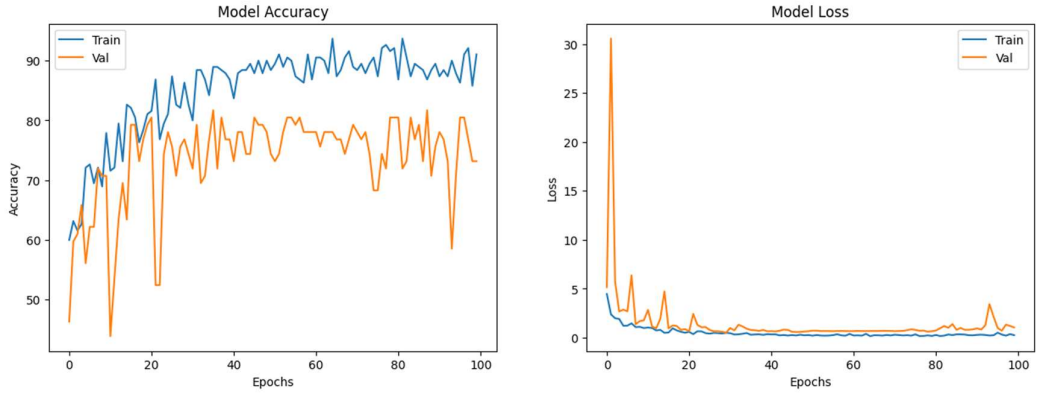
Defining Training Parameters

12. Keep epochs = 100.
13. Keep criterion = nn.CrossEntropyLoss()
14. Modify optimizer = torch.optim.Adam(model.parameters(), lr=1e-2, weight_decay=1e-4) for faster learning.
15. Change lr_scheduler = torch.optim.lr_scheduler.CosineAnnealingLR(optimizer, T_max=60) for make sure model will not in local optima and can fined the global optima.
16. Add L2 Regularization to avoid exploding gradient problem.
17. Add gaussian noise for data augmentation.

Training Model Try And Error Experience

18. Larger NN model will cause more overfitting.
19. Higher dropout rate will cause model training unstable.
20. Lower dropout rate will cause overfitting.
21. Adding batchnorm can help with increase accuracy and avoid overfitting.
22. Larger learning rates will increase training speed.
23. Add L2 Regularization can avoid overfitting.
24. Add gaussian noise can increase accuracy.

Training Result



Test accuracy is 80.64516129032258%

Conclusion

It still overfits after 40 epochs, and test accuracy is not satisfied. Training model's performance is not stable if retrain several times. Maybe the cause of Cp, Ca, oldpeak, and target std are larger than the mean. Or maybe we can try another kind of models for this kind of data.