# Asthma Disease Prediction

Group\_33: Huaqing Jiang (A20482610, [hjiang30@hawk.iit.edu](mailto:hjiang30@hawk.iit.edu));

Chieh Jui Le (A20523922, [clee114@hawk.iit.edu](mailto:clee114@hawk.iit.edu) ).

### 1. Dataset Introduction

The "Asthma Disease Prediction" dataset is a comprehensive collection of anonymous health records and patient data, carefully planned for predictive modeling and research purposes. It includes important patient information, environmental factors, and medical history, enabling the development of advanced machine learning models to predict the onset, severity, and treatment outcomes of asthma. This dataset is a valuable resource for improving early diagnosis and management of asthma, ultimately enhancing the quality of care for affected individuals.

This dataset has the following attributes:

Fatigue: Fatigue may not be a direct symptom of asthma, but it may be a secondary effect. Asthma patients may experience fatigue due to breathing difficulties and other symptoms.

Dry cough: Dry cough is a common symptom of asthma. It can be triggered or worsened by asthma related inflammation and airway irritation.

Throat pain: Throat pain is usually not the main symptom of asthma. However, if asthma patients cough frequently due to asthma symptoms, they may experience throat pain.

None\_ Symptom and None\_ Experience: It is possible that some people in the dataset did not experience any of the listed symptoms. The symptoms of asthma vary from person to person, and some people may have asthma, but the symptoms may not persist.

Pain: "Pain" is somewhat universal and can refer to different types of pain. This is not a typical symptom of asthma, but asthma patients may experience chest discomfort or tightness during asthma attacks.

Nasal congestion and runny nose: Nasal congestion and runny nose are not typical symptoms of asthma itself. However, they may exist in patients with allergic asthma, as allergies can trigger asthma symptoms.

Ages 0-9 to 60+: Asthma can affect individuals of all ages. Generally speaking, the severity and frequency of asthma symptoms vary among different age groups. Children and elderly people may have different symptoms of asthma.

Gender\_ Female and Gender\_ Male: Asthma can affect individuals of any gender. Gender may not directly cause asthma symptoms, but it can affect the management of asthma and the prevalence of asthma in different populations.

Severity\_ Mild, Severity\_ Moderate and Severity\_ None: These lists indicate that the severity of asthma varies. Asthma can indeed range from mild to moderate to severe, depending on its level of control and individual response to treatment. The existence of these severity levels indicates that asthma symptoms have different impacts on different individuals.

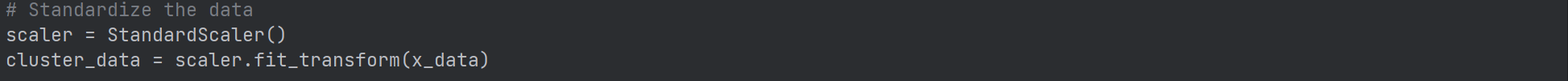
Difficulty breathing: This is a typical symptom of asthma. The characteristic of asthma is narrowing of the airways, leading to difficulty breathing. At the end of this project, we will use this data as the target dataset and train based on different models.

### 2. Data Clustering Analysis

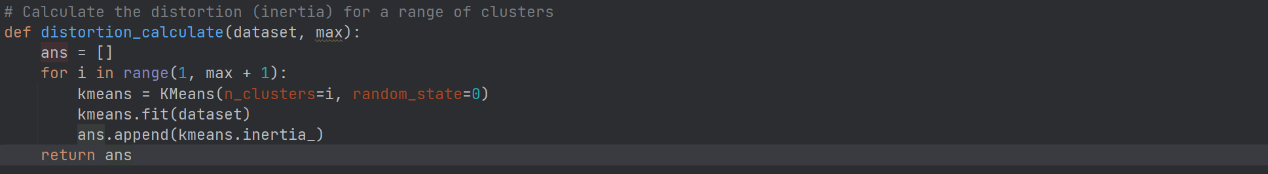
After importing the data, we will use clustering to analyze the data and features. Based on the clustering heatmap, we can intuitively feel the classification and influence of features.

#### 2.1 Determination of the optimal number of clusters

Firstly, we use ‘StandardScaler‘ to standardize the data. This is to ensure that all features have the same proportion in cluster analysis. The process of standardization usually involves converting data into a distribution with a mean of 0 and a standard deviation of 1：



Then, we defined a function to calculate the distortion (inertia) of different numbers of clusters：

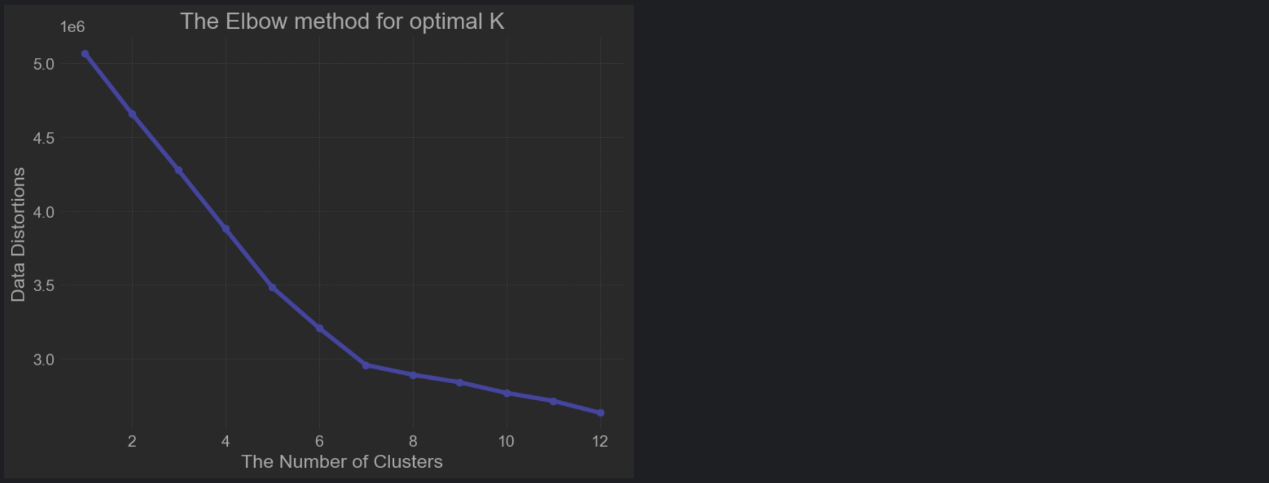


This function calculates the distortion of clusters from 1 to max (represented by inertia in the K-Means algorithm). Inertia refers to the sum of squared distances from each point to its cluster center, and the smaller this value, the better the clustering effect.

Finally, we use the elbow method to draw a chart to determine the optimal number of clusters:



Due to the presence of a total of 19 features in this example, the maximum clustering we tested was 12, and the resulting elbow image is:



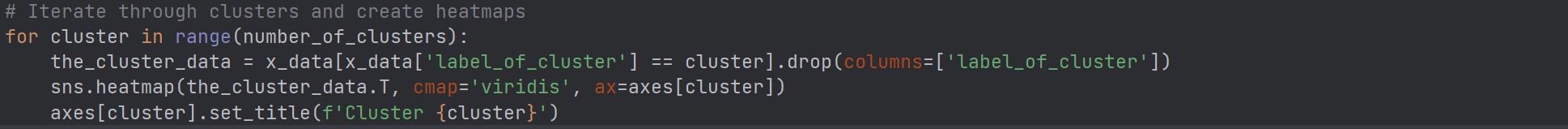
From the graph, it can be intuitively seen that the inflection point of clustering distortion is 7, so the optimal number of clusters is 7.

#### 2.2 Draw clustering heatmap

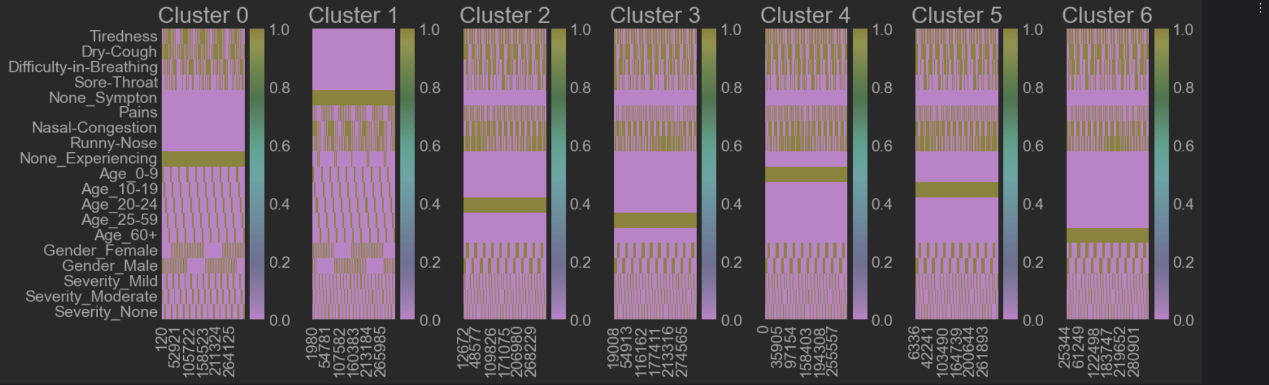
In the above steps, we have learned that the optimal number of clusters is 7. Next, we will create subgraphs for each cluster:



Then traverse the clusters and create heatmaps for each cluster:



The final clustering heatmap obtained is:



From this heat map, we can see that:

Cluster 0: This cluster seems to be related to influenza or colds. Further domain knowledge is needed to thoroughly examine it.

Cluster 1: This cluster seems to be related to influenza. These functions are usually balanced.

Cluster 2: Refers to the age range of 20 to 24 years old. They also show symptoms.

Cluster 3: Refers to the age range of 25-59 years old. Their gender distribution is balanced and they exhibit symptoms, usually with mild or no severity of asthma.

Cluster 4: Refers to the age range of 0-9 years old. Their gender distribution is balanced and they exhibit symptoms, usually with moderate or no asthma severity.

Cluster 5: Refers to the age range of 10 to 19 years old. Their gender distribution is balanced and they exhibit symptoms, with asthma severity ranging from moderate to mild.

Cluster 6: It seems to refer to the age group of 60 and above. They exhibit symptoms, usually with mild or no severity of asthma.

At the same time, it can be seen that the distribution of males and females in each cluster is relatively balanced and there is no significant difference. Therefore, we can discard these two features in the following model training to improve the accuracy of the model.

### 3. Model Training

Before training the model, we first use ‘MinMaxScaler’ to normalize features and targets, ensuring that all features are on the same scale.

Then use ‘train\_test\_split’ divides the data into training and testing sets, with the testing set accounting for 20% and a random seed set to 42 to ensure the reproducibility of the results.

#### 3.1 Neural network model

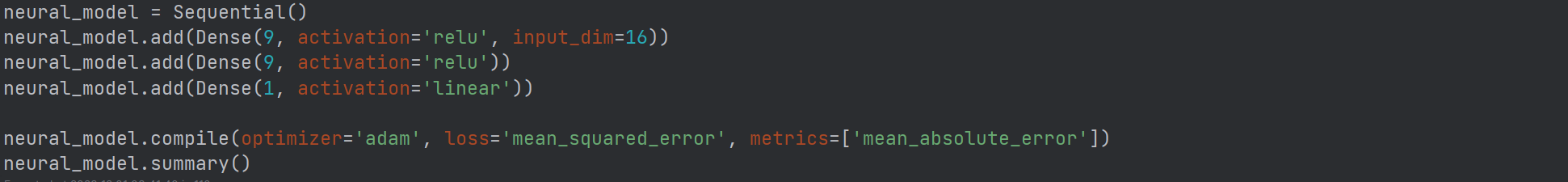
Regarding the construction of neural network models:

Firstly, we use Sequential to create a sequence model.

Then add three layers of Dense (fully connected layers) to the model. The first layer contains 9 neurons and receives inputs from 16 features (the training data after data processing contains 16 features). The last two layers each contain 9 neurons, and the last layer is the output layer with only one neuron (used for regression tasks).

The activation functions are 'relu' (first two layers) and 'linear' (output layer), respectively.

We use the 'adam' optimizer and 'mean' during model compilation\_ Squared\_ The error loss function is a commonly used choice for regression tasks.



We have established 8 training cycles for the model. At the same time, 20% of the training data was designated as validation data. This part of the data is not used for training the model, but to evaluate the performance of the model on unseen data, in order to monitor whether overfitting occurs during the training process. The execution result of model training is assigned to the variable log. This variable contains information from the training process, such as the loss and accuracy of each epoch, as well as the performance on the validation set. This information can be used for subsequent analysis and visualization:



Subsequently, we visualized the losses and mean square error (MSE) during the neural network training process.

We use Matplotlib to draw two subgraphs. The first subgraph shows the variation of training and validation losses with epoch, while the second subgraph shows the variation of training and validation MSE with epoch. Also, we mark the points corresponding to the best performance with scatter points on the loss and MSE curves.



From the provided chart, we can observe the following points:

The trend of loss changes: The training loss (red line) gradually decreases from the first epoch and tends to stabilize from the 2th epoch.The validation loss (green line) fluctuates in the initial few epochs, but starts to decrease and stabilize from the 6th epoch.

The trend of mean square error (MSE) variation: The training MSE (red line) drops sharply after the first epoch and then fluctuates slightly at a lower level. Verify that MSE (green line) reaches a peak after the first epoch, then rapidly decreases, and slightly increases after reaching its lowest point in the sixth epoch.

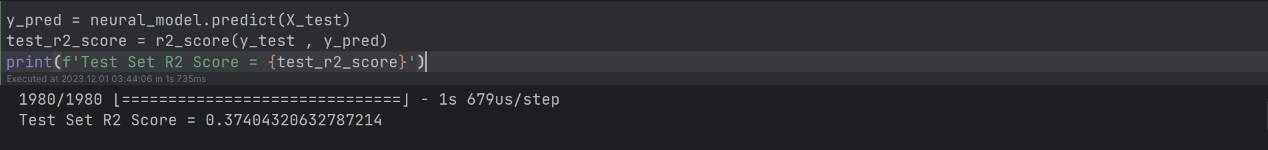
Identification of the best Epoch:

Both graphs are marked with "minimum epoch=6", indicating that the validation loss and validation MSE reached their minimum values at the 6th epoch. This usually indicates that the model has the best generalization ability at that epoch.

Analysis of overfitting and underfitting:

There are no obvious signs of severe overfitting or underfitting. If the validation loss starts to increase while the training loss continues to decrease, it may be a sign of overfitting; On the contrary, if the training loss is significantly higher than the validation loss, it may indicate that the model is underfitting. In the current chart, both training and validation losses maintain a similar trend, which is a good phenomenon. Although the model performs best at the 6th epoch, the overall difference between training and validation metrics is relatively small. This indicates that the model may have been relatively stable.

Finally, we used the model to predict the test set and calculated the r2 score:



The r2 score of the test set is 0.37404320632787214, which is approximately 0.37. This means that your neural network model can explain about 37% of the variability of output variables. Although this is not a very high score, it does mean that the model has a certain predictive ability compared to random prediction.

Based on the r2 score, we believe there is still room for further improvement in model performance, including increasing the amount of data and adding more useful features.

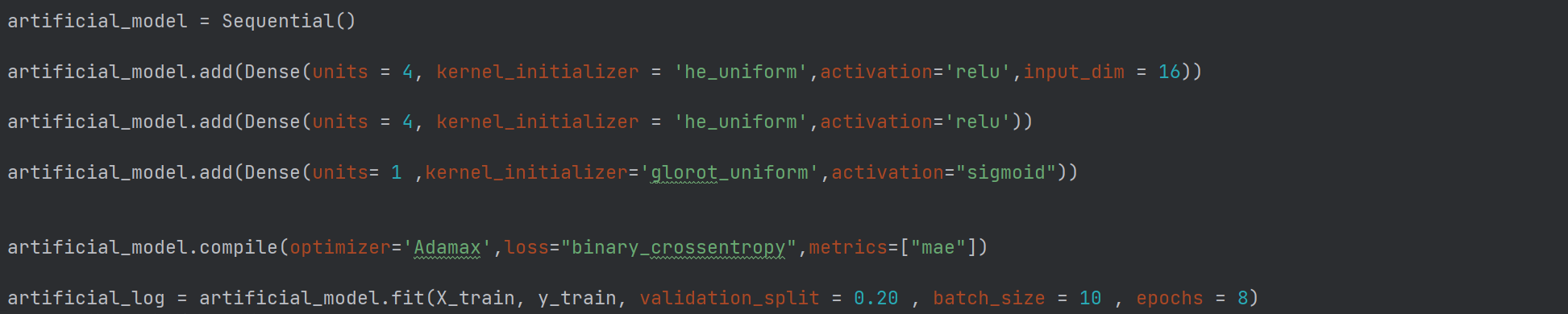
#### 3.2 Artificial neural network

Similar to the previous process, we first initialized a sequence model.

Then, two hidden layers with 4 units were added, using he\_ Uniform initializer and 'relu' activation function. The number of features specified for the input layer is 16.

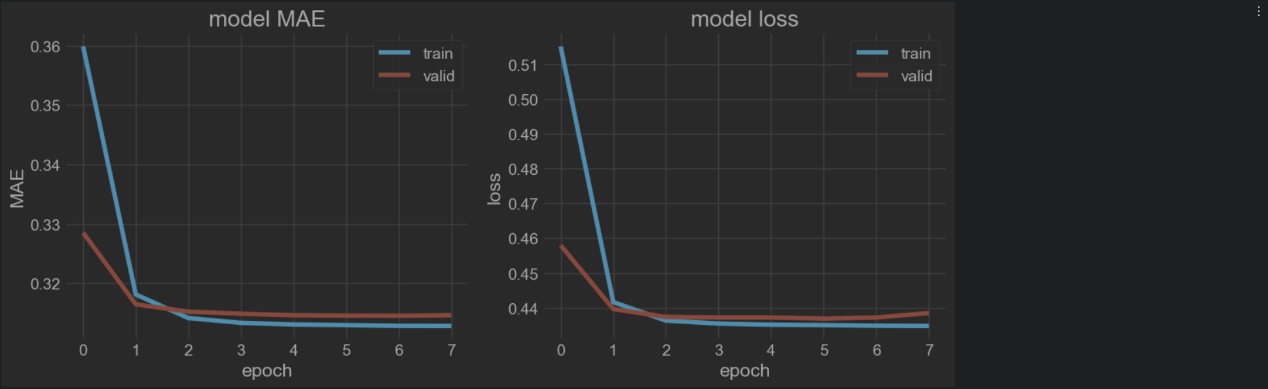
Subsequently, an output layer for a unit was added, using 'glorot\_uniform’ initializer and sigmoid activation function.

When compiling the model, we use the 'adamax' optimizer and 'binary\_crossentropy' as the loss function. Similarly, we trained the model using the fit method, specifying a validation set partition of 20%, a batch size of 10, and a training period of 8:



Next, the graph of the mean absolute error (MAE) and loss (loss) during the training process of the artificial neural network model was drawn. The graphics are arranged in a subgraph of one row and two columns.

The left subgraph shows the MAE on the training and validation sets. MAE is an indicator for evaluating the accuracy of model predictions, which calculates the average absolute difference between the predicted value and the true value. The subgraph on the right shows the losses on the training and validation sets.

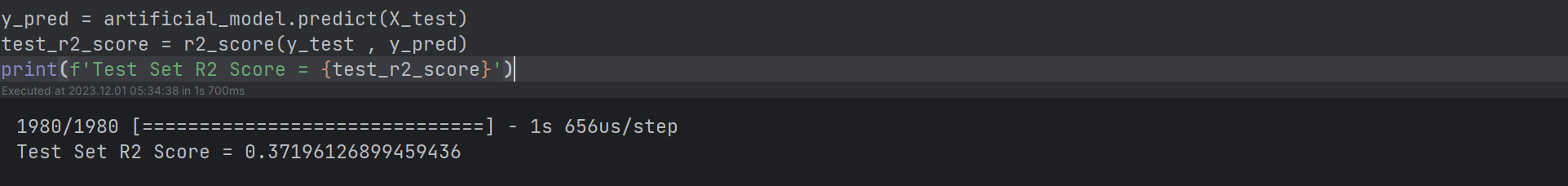


Two graphs show the performance of the model during the training process:

Mean Absolute Error (MAE) plot: The training and validation MAE both sharply decreased after the first epoch, and then gradually stabilized and approached each other, indicating that the model has made progress on both the training and validation sets and performed consistently.

Loss chart: Similar to the MAE graph, the loss rapidly decreases after the first epoch and slowly decreases in subsequent epochs. The training and validation losses tend to stabilize and are very close, which means that the model does not show significant overfitting.

Overall, from the curve perspective, this model performs better than the neural network model, so next, let's take a look at how this model performs on the test set:



Similar to the previous model, the R2 score of the model on the test set is approximately 0.37, indicating that the model can only explain about 37% of the variance of the target variable. This is still not particularly ideal. From the previous curve graph, it can be seen that there is basically no overfitting problem in the model, so it is possible that the input features of the data are not comprehensive enough. If time is enough, we can consider collecting additional data sources or conducting further data processing work next.