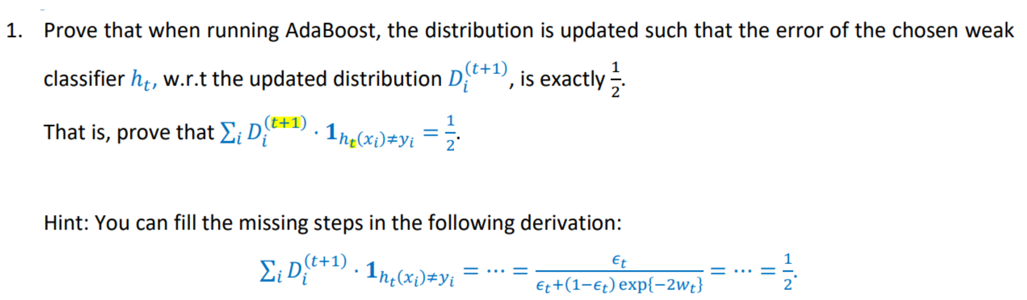
Homework No. 5

ALEXANDER SHENDER 328626114, SNIR HORDAN 205689581

Question 1.



1. Prove that when running AdaBoost, the distribution is updated such that the error of the chosen weak classifier h_t, w.r.t the updated distribution $D_i^{(t+1)}$, is exactly $\frac{1}{2}$.

   That is, prove that $\sum_i D_i^{(t+1)} \cdot \mathbf{1}_{h_t(x_i)\neq y_i} = \frac{1}{2}$.

   Hint: You can fill the missing steps in the following derivation:

$$\sum_i D_i^{(t+1)} \cdot \mathbf{1}_{h_t(x_i)\neq y_i} = \cdots = \frac{\epsilon_t}{\epsilon_t+(1-\epsilon_t)\exp\{-2w_t\}} = \cdots = \frac{1}{2}.$$

We start by indeed from writing the expression for the updated error value with the updated data weights (distribution) for the last chosen weak classifier:

$$\mathrm{E}_{t+1} = \sum_i D_i^{t+1} \cdot 1_{h_t(x_i)\neq y_i}$$

By using:

$$D_i^{t+1} = D_i^t \cdot \frac{\exp(-w_t y_i h_t(x_i))}{\sum_j D_j^t \exp\left(-w_t y_j h_t(x_j)\right)} = D_i^t \cdot \frac{\exp(-w_t y_i h_t(x_i))}{Z_t}$$

Putting back:

$$\mathrm{E}_{t+1} = \sum_i D_i^{t+1} \cdot 1_{h_t(x_i)\neq y_i} = \sum_i \frac{D_i^t \exp(-w_t y_i h_t(x_i))}{Z_t} \cdot 1_{h_t(x_i)\neq y_i}$$

Z_t is a normalization factor, so we can put it outside of the sum:

$$\mathrm{E}_{t+1} = \frac{\sum_i D_i^t \exp(-w_t y_i h_t(x_i))}{Z_t} \cdot 1_{h_t(x_i)\neq y_i}$$

We divide into 2 cases:

$$\begin{cases} E : h_t(x_i) = y_i \\ C : h_t(x_i) \neq y_i \end{cases}$$

For each case,

$$\begin{cases} E : 1_{h_t(x_i)\neq y_i} = 0 \,;\; y_i h_t(x_i) = 1 \\ C : 1_{h_t(x_i)\neq y_i} = 1 \,;\; y_i h_t(x_i) = -1 \end{cases}$$

So we get:

$$\mathrm{E}_{t+1} = \frac{\sum_{i\in E} D_i^t \exp(-w_t y_i h_t(x_i))}{Z_t} \cdot 0 + \frac{\sum_{i\in C} D_i^t \exp(-w_t y_i h_t(x_i))}{Z_t} \cdot 1 = \frac{\sum_{i\in C} D_i^t \exp(w_t)}{Z_t}$$

The numerator contains the expression for the error value, since we have isolated for case $\{E : h_t(x_i) = y_i\}$:

$$\mathrm{E}_t = \sum_{i \in C} D_i^t$$

$$\mathrm{E}_{t+1} = \frac{\mathrm{E}_t \exp(w_t)}{Z_t}$$

Opening the denominator using same 2 cases:

$$Z_t = \sum_{i \in E} D_i^t \exp(-w_t) + \sum_{i \in C} D_i^t \exp(w_t)$$

Putting back:

$$\mathrm{E}_{t+1} = \frac{\mathrm{E}_t \exp(w_t)}{\sum_{i \in E} D_i^t \exp(-w_t) + \sum_{i \in C} D_i^t \exp(w_t)} \quad ; \ / \exp(w_t)$$

$$\mathrm{E}_{t+1} = \frac{\mathrm{E}_t}{\sum_{i \in E} D_i^t \exp(-2w_t) + \sum_{i \in C} D_i^t} = \frac{\mathrm{E}_t}{\sum_{i \in E} D_i^t \exp(-2w_t) + \mathrm{E}_t}$$

Taking $\exp(-2w_t)$ out of the sum:

$$\mathrm{E}_{t+1} = \frac{\mathrm{E}_t}{\exp(-2w_t) \sum_{i \in E} D_i^t + \mathrm{E}_t}$$

The sum over weights of the correct predictions E, is 1-sum over incorrect predictions, since they sum to 1:

$$\sum_{i \in E} D_i^t = 1 - \sum_{i \in C} D_i^t \ = 1 - E_t$$

So we get:

$$\mathrm{E}_{t+1} = \frac{\mathrm{E}_t}{\exp(-2w_t)\,(1 - E_t) + \mathrm{E}_t}$$

Using the expression for the weight of the weak classifier:

$$w_t = \frac{1}{2} \log\left(\frac{1}{\mathrm{E}_t} - 1\right)$$

$$\exp(-2w_t) = \exp\left(-\mathit{log}\left(\frac{1}{\mathrm{E}_t} - 1\right)\right) = \exp\left(\log\left(\left(\frac{1}{\mathrm{E}_t} - 1\right)^{-1}\right)\right) = \left(\frac{1}{\mathrm{E}_t} - 1\right)^{-1} = \left(\frac{1 - \mathrm{E}_t}{\mathrm{E}_t}\right)^{-1}$$
$$= \frac{\mathrm{E}_t}{1 - \mathrm{E}_t}$$

Putting everything back:

$$\mathrm{E}_{t+1} = \frac{\mathrm{E}_t}{\frac{\mathrm{E}_t}{1 - \mathrm{E}_t}(1 - E_t) + \mathrm{E}_t} = \frac{\mathrm{E}_t}{\mathrm{E}_t + \mathrm{E}_t} = \frac{1}{2}$$

Question 2

2.a.

Let $\alpha \in \mathbb{R}_{>0}$

Notice σ is positive-homoegeneous, i.e. $\sigma(\alpha x) = \max\{0, \alpha x\} = \alpha \max\{0, x\} = \alpha\sigma(x)$

Denote $\alpha\theta = \left(\alpha W^{(1)}, ..., \alpha W^{(L)}\right)$.

Let h be defined on θ as in question and $\tilde{h}$ defined on $\alpha\theta$ as in question.

Claim. $\tilde{h}_L(x) = \alpha^L h_L(x)$

Proof. Proof by induction on L.

Base : $\tilde{h}_1(x) = \sigma\left(\alpha {W^{(1)}}^T x\right) \underbrace{=}_{\alpha\, positive\, homogeneous} \alpha\sigma\left({W^{(1)}}^T x\right) = \alpha h_1(x)$

Assume for n<L, $\tilde{h}_n(x) = \alpha^n h_n(x)$

Induction step:

By definition, $\tilde{h_L}(x) = \sigma\left(\alpha {W^{(L)}}^T \tilde{h}_{L-1}(x)\right) \underbrace{=}_{induction\, hypothesis} \sigma\left(\alpha {W^{(L)}}^T \alpha^{L-1} h_{L-1}(x)\right) =$

$\underbrace{=}_{\alpha\, positive-homogeneous} \alpha^L \sigma\left({W^{(L)}}^T h_{L-1}(x)\right) = \alpha^L h_L(x)$ □

By definition, $F_{\alpha\theta}(x) = {W^{(L)}}^T \tilde{h}_{L-1}(x) = {W^{(L)}}^T \alpha^{L-1} h_{L-1}(x) = \alpha^{L-1} F_\theta(x)$

Then $C = \alpha^{L-1}$.

2.b.

$$c_{y_i} = \frac{\exp\left((F_{\alpha\theta}(x))_i\right)}{\sum\limits_{j=1}^{k} \exp\left((F_{\alpha\theta}(x))_j\right)} = \frac{\exp\left(\alpha^L (F_\theta(x))_i\right)}{\sum\limits_{j=1}^{k} \exp\left(\alpha^L (F_\theta(x))_j\right)} \underset{\alpha\to 0}{\to} \frac{1}{k}$$

Then the distribution is uniform.

2.c.

<u>Case 1</u> $(F_\theta(x))_i > (F_\theta(x))_j$, $\forall j \in \{m_1, ..., m_{k-s}\}$ and $(F_\theta(x))_i = (F_\theta(x))_j$, $\forall j \in \{m_{k-s+1}, ..., m_k\}$, $(\{m_1, ..., m_k\} = \{1, ..., k\})$

$$c_{y_i} = \frac{\exp\left((F_{\alpha\theta}(x))_i\right)}{\sum\limits_{j=1}^{k} \exp\left((F_{\alpha\theta}(x))_j\right)} = \frac{\exp\left(\alpha^L (F_\theta(x))_i\right)}{\sum\limits_{j=1}^{k} \exp\left(\alpha^L (F_\theta(x))_j\right)} = \frac{1}{\sum\limits_{j=1}^{k} \exp\left(\alpha^L \left((F_\theta(x))_j - (F_\theta(x))_i\right)\right)} = ...$$

$$... = \frac{1}{s + \underbrace{\sum\limits_{j\in\{m_1,...,m_{k-s}\}} \exp\left(\alpha^L \left((F_\theta(x))_j - (F_\theta(x))_i\right)\right)}_{\underset{\alpha\to\infty}{\to} 0}} \underset{\alpha\to\infty}{\to} \frac{1}{s}$$

($s \geq 1$ because exists at least one maximum value then limit is well-defined)

<u>Case 2</u> $\exists 1 \leq j \leq k \, s.t. \quad (F_\theta(x))_i < (F_\theta(x))_j$

$$c_{y_i} = \frac{1}{1 + \underbrace{\sum\limits_{j\neq i} \exp\left(\alpha^L \left((F_\theta(x))_j - (F_\theta(x))_i\right)\right)}_{\underset{\alpha\to\infty}{\to} \infty}} \underset{\alpha\to\infty}{\to} 0$$

Then distribution is uniform on the $s \leq k$ entries with maximum value and zero otherwise.

i.e. $\left(0, ..., \underbrace{\frac{1}{s}}_{m}, ..., 0\right)$ where $m \in argmax_{1 \leq i \leq k}\left(\left(F_{\theta}\left(x\right)\right)_{i}\right)$.

Wet section

To succeed in the classification task in the best possible way, we first had to go back into feature imputation stage.

Code Review

As before, our code is fully automatic, and performs the model selection automatically.

To the existing model we have added new models, which are AdaBoost, and the MLP.

The same actions on features are later done on the unseen data, and the output is given.

Features imputation

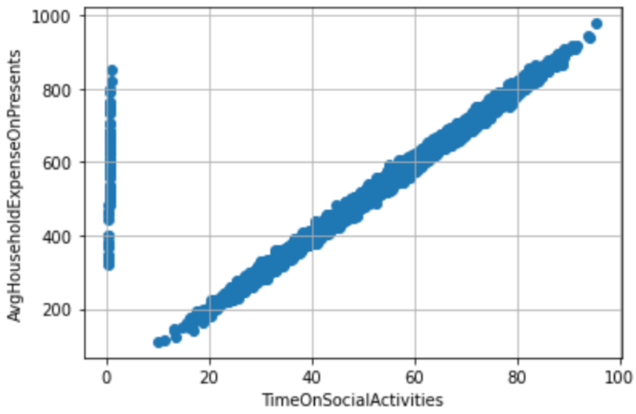
Some feature imputations techniques depend on the dataset, such as replacing missing values with the mean value. For such cases, the mean values found from the operation on the training set were found and saved. And later used on the test set and the unseen data.

1. Discipline score

The discipline score had outliers, but it was noticed that they are simply multiplications of existing discipline scores by 10. Those outliers were divided by 10. The other missing were replaced with a mean.

2. Social Activities

There is a big correlation between the TimeOnSocialActivities and the AvgHouseholdExpenseOnPresents graphs:

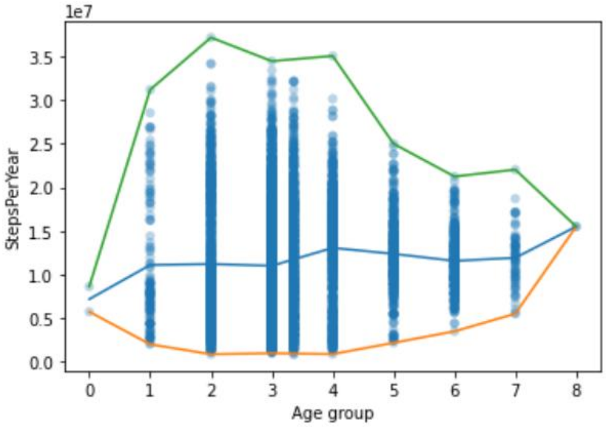


The following imputations are done:

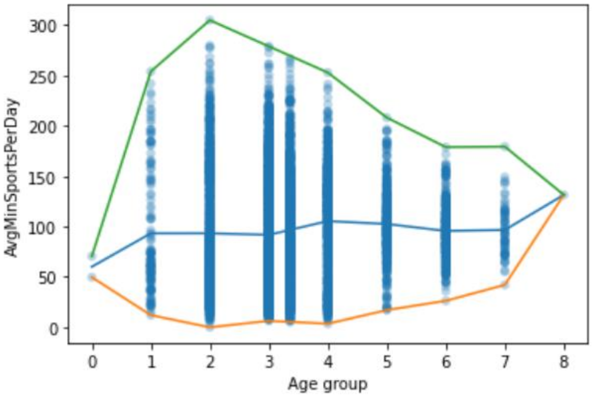
- The outliers (TimeOnSocialActivities < 5) was replaced according to AvgHouseholdExpenseOnPresents
- The non-existing values were also replaced according to AvgHouseholdExpenseOnPresents (if those existed there)
- For all the rest, replacing with the mean value of TimeOnSocialActivities

3. Age Handler

To handle missing age, we have examined the AvgMinSportsPerDay and StepsPerYear parameters, and this is the distribution (with upper and lower boundaries)



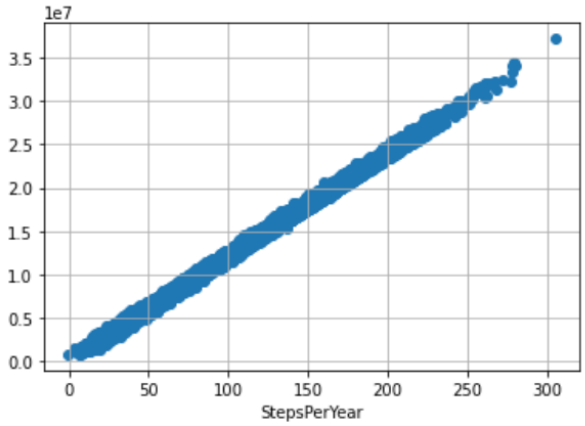
Same for AvgMinSportsPerDay:



We can see some correlation (which is logical).

4. Steps Per Year

There is a strong correlation between AvgMinSportsPerDay and StepsPerYear:



So if the value for StepsPerYear is missing, but is existing in AvgMinSportsPerDay, we calculate this value. Other values are replaced by the mean value.

5. PCR Results

Like with the case with clustering (homework no. 4) we try both the mean and the KNN imputer. For each PCR result, we use the z-value to discover outliers. Then calculate the mean without the outliers. Then we replace the outliers and the missing values by those means or the KNN.

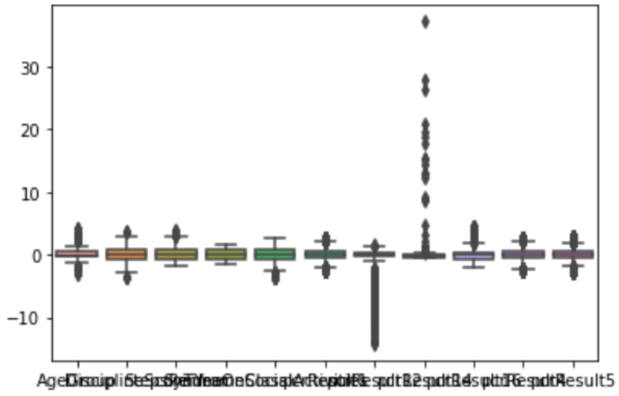
- For training set: we SAVE the mean and STD values that were calculated
- For test/unseen sets: we use the saved mean and std values for each feature to find outliers in the unseen dataset. (we use the same z-score value). We then replace the outliers and missing values with the mean.

6. Syndrome Class

No correlation was found between this feature and any other feature. So we add a new categorical value ‘5’ for ‘unknown syndrome class’.

Features Scaling

StandardScaler was used on all features. The boxplot distribution of the data after the scaling:



Cluster Analysis

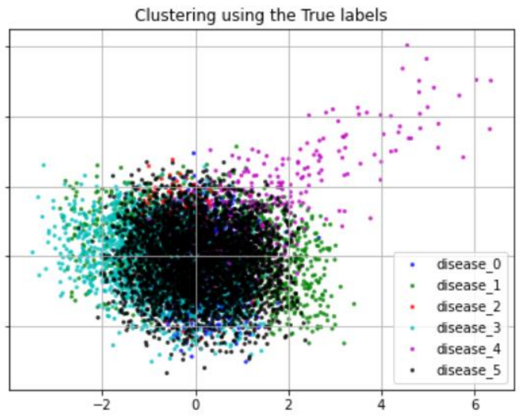
By learning to perform the cluster analysis, we tried to visualize the 6 clusters of the disease type. This would also help us to verify that the feature imputation procedures were indeed effective.

We use the most successful clustering model from the previous homework, which is GaussianMixture Model, since most of the features distributions resemble Gaussian distribution.

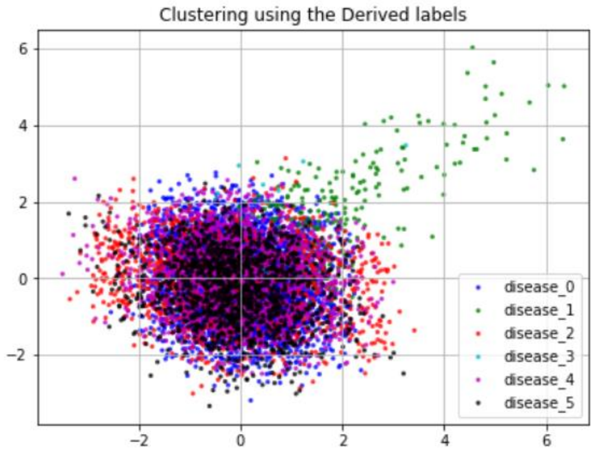
To visualize, we use the 2 PC (3d plot doesn’t add much more information).

We used the whole training set of 9000 samples.

Using the TRUE samples (from the dataset):



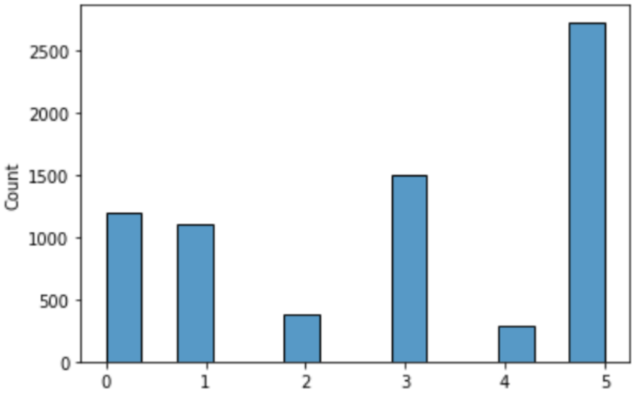
And using the Derived cluster indexes from the GMM:



We can see that GMM succeeded to discover some clusters (like the easy one which spans through big x and y values), but in general it doesn’t correspond too much. We will later use this visualization to assert the precision of the classifier.

Class Imbalance

For the disease type, the distribution of the classes in original dataset is not balanced:



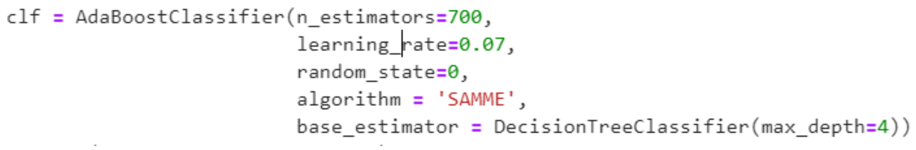
We handle the imbalanced data using SMOTE (Synthetic Minority Oversampling Technique), which created new samples for the minority classes, equating them with the majority class. The class imbalance handling was also added to the automatic training process.

Classification

We have implemented new gained knowledge to tackle the classification problem with more success than previously. We first concentrated on the Disease classification, since it was the most challenging task from the HW3.

1. Adaboost

We have used the AdaBoost classifier with various parameters, experimenting with the algorithm (sklearn offers SAMME, SAMME.R), number of estimators, base estimator and its parameters. The most successful parameters were the following:



```
clf = AdaBoostClassifier(n_estimators=700,
                         learning_rate=0.07,
                         random_state=0,
                         algorithm = 'SAMME',
                         base_estimator = DecisionTreeClassifier(max_depth=4))
```

Which gave around 86% on the test set for the Disease classification.

Pros

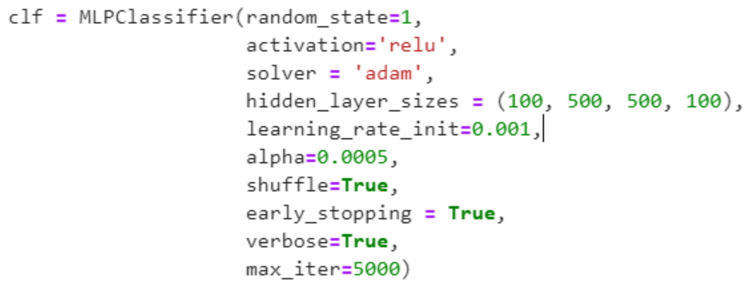
- Freedom of choosing base estimators.
- Easy to implement.

Cons

- Weak base estimators can lead to low margins overfitting
- Uniform noise can distort results

2. MLP

The sklearn library provides the MLP classifiers. We have as well examined different parameters variations; we found the following to be relatively fast converging and accurate:



```
clf = MLPClassifier(random_state=1,
                    activation='relu',
                    solver = 'adam',
                    hidden_layer_sizes = (100, 500, 500, 100),
                    learning_rate_init=0.001,
                    alpha=0.0005,
                    shuffle=True,
                    early_stopping = True,
                    verbose=True,
                    max_iter=5000)
```

In reality it would converge already after 50 epochs (iterations), with the accuracy of 89% on the test set, thus proving superior to the Adaboost.

Pros

- Flexibility of optimizing classification and regression problems.
- Free to choose any number of layers, thereby being able to solve complex pattern recognition.
- Improves with more data.

Cons

- No clear guideline on how to choose hyperparameters (layer sizes, amount of layers, learning rate, etc.)
- Computationally expensive.
- Requires large amounts of data to converge

3. MLP + Transfer Learning

In transfer learning, we used the following process:

a. First, train the binary classification on the Disease (0 = No disease, 1 = Any other disease). Use the SMOTE to fix class imbalance between classes (No disease, Disease)
b. Build a new model based on the trained model in (a.). Add new layers (only those layers will be trained)
c. Train the new model.

Using this process, we could achieve about 55% of accuracy on the Disease classification, which didn't improve our earlier tries.

Our analysis is that transfer learning did not improve results, because binary classification of disease/no disease did not generalize to differentiating between diseases, which led to low accuracy. This occurred even though the no disease class is largest, which is why we tried this method out in first place.

Transfer learning is usually used for NLP and Image Recognition and trained on much larger datasets. This is another reason for the low accuracy derived from the model.

(The code is available in transfer.py in the submitted work.)

4. Ensemble Learning

Ensemble learning is learning by a collection of estimators where we take a hard vote, regarding the best classification for each instance agreed on by most models, or a soft vote, summing over the probability assigned to each class and then picking the class with the maximum sum of probabilities assigned to it.

If the underlying collection of models is robust then ensemble learning provides a mechanism to increase certainty in the classification task.

Results

After running the automatic model selection, those are the models that were chosen by the process and their metrics on the test dataset.

Risk detection

The model chosen was an ensemble of four neural networks and AdaBoost with a decision tree base classifier with hard vote.

Chosen model : (AdaBoostClassifier(), {'n_estimators': [50, 100], 'base_estimator': [DecisionTreeClassifier(max_depth=1)]})

Metrics : f1_score

Value : 0.877 on validation set

F1 score is 0.77959 on test set.

Metric Selection

F1 score is ideal because we need a balance of low false negatives and high accuracy, because don't want to tell healthy people they are at risk but also don't want to make someone falsely certain he will have no consequences rom the virus

Spreader detection

Chosen model : (DecisionTreeClassifier(), {'max_depth': [5, 10, 15]}, True)

Metrics : recall_score

Value : 0.954 on valdation dataset

Recall on test dataset of 0.9466.

Metric Selection

Recall is the optimal metric to assess risk of spread, because a False Negative should be penalized harshly. A false negative can lead to massive spread so we would rather stay on the side of safety.

Disease detection

The model chosen was an ensemble of four neural networks and AdaBoost with a decision tree base classifier with hard vote.

```
    Chosen model : (VotingClassifier(estimators=[('mlp1',
                      MLPClassifier(hidden_layer_sizes=(100, 150, 150,
                                                        100, 10),
                                    max_iter=150)),
                     ('mlp2',
                      MLPClassifier(hidden_layer_sizes=(50, 100, 200,
                                                        200, 100),
                                    max_iter=150)),
                     ('mlp3',
                      MLPClassifier(hidden_layer_sizes=(50, 100, 100,
                                                        200, 100),
                                    max_iter=150)),
                     ('ada',
                      AdaBoostClassifier(base_estimator=DecisionTreeClassifier(max_depth=4),
                                         n_estimators=100))]), {}, True)
    Metrics : accuracy_score
    Value : 0.916 on validation set
```

Accuracy on test set: 0.69555

Metric Selection

Different diseases require varied interventions, therefore accuracy is vital.