1. **Problem description:**

Create a classifier using a dataset of 47180 patients to train the model.

Each patient will have 27 attributes:

* Gender: Gender of the patient (e.g., Male or Female).
* Age: Age of the patient.
* LOSdays: Length of Stay in days.
* admit\_type: Type of admission (e.g., EMERGENCY, ELECTIVE).
* admit\_location: Location from which the patient was admitted.
* AdmitDiagnosis: Diagnosis upon admission.
* Insurance: Type of insurance the patient has.
* Religion: Patient's religion.
* marital\_status: Marital status of the patient.
* ethnicity Ethnic background of the patient.
* NumCallouts: Number of callouts.
* NumDiagnosis: Number of diagnoses.
* NumProcs: Number of procedures.
* AdmitProcedure: Procedure upon admission.
* NumCPTevents: Number of CPT events.
* NumInput: Number of inputs.In a hospital or medical context, "input" typically refers to the amount of fluids, medications, or nutrients given to a patient.
* NumLabs: Number of lab tests.
* NumMicroLabs: Number of microbiology labs.
* NumNotes: Number of notes.
* NumOutput: Number of outputs.
* NumRx: Number of prescriptions.
* NumProcEvents: Number of procedure events.
* NumTransfers: Number of transfers.
* NumChartEvents: Number of chart events.
* ExpiredHospital: Indicates if the patient expired in the hospital (0 for No, 1 for Yes).
* TotalNumInteract: Total number of interactions.
* LOSgroupNum: Group number based on the length of stay.

The classifiers will then be given another dataset of patients with the “ExpiredHospital” column being unknown. Using the other 26 attributes of the patient, the classifiers will have to predict whether this patient has expired in hospital or not.

1. **Data preprocessing and transformations:**

Here are all of the data preprocessing I did for the classifiers:

* Dropping the *‘LOSgroupNum’* column:

Because 42447 out of 47180 values in this column are missing, the importance of this feature in the training process is minimal. Furthermore, because of how little information we have on this feature, using it during the training process can make the classifiers overfit. Therefore, I thought it would be better to drop it altogether.

This line of code below is used to remove the feature from the dataset.



* Filling missing values in the dataset:

Because the dataset has two types of data: numerical and categorical, and both types have missing values. Therefore, I had to employ two approaches:

* + For numerical features: The missing numbers in a feature are filled with the average (the mean) value of that feature.

The code:



Where x1 are the numerical features.

* + For categorical features: The missing data in a feature are filled with the most frequent class (the mode) of that feature.

The code:



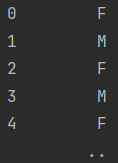
Where x2 is are the categorical features.

* Encoding categorical features:

Because classifiers can’t interpret human language, classes that are written in words/letters will have to be encoded into numbers. Where each number will represent a class of that feature. This can be done using the .factorize() function from the *pandas* package.



A black background with white numbers

Description automatically generatedBefore and after encoding:

* Normalization for numerical features:

Some models like Support Vector Classifier (SVC) or Neural Network (NN) would perform better when the data in the numerical features are normalized. Here are some reasons why:

* + Because the scale of the numbers varies a lot depending on the feature. Normalization will scale all of the numbers to the same range, so the model can’t be biased towards features with larger numbers.
  + Because this dataset has a lot of outliers, normalization reduces the emphasis placed on those outliers and prevents overfitting.
  + I used the *StandardScaler()* from the *sklearn.preprocessing* package for the normalization.



A black background with white text

Description automatically generated

* Lastly, I used an 80-20 split to generate the train and test set for the classifiers.



1. **Solving the problem:**

I experimented with a lot of classifiers, all of the ones that were taught in this course and some more outside of this course. Then I selected the best performing classifier and start training many models of that classifier, the model with the highest f1\_macro score in a training session will be submitted to Kaggle. To get the best performing model when training a classifier, I used some techniques:

* Changing the weight of the labels: As the dataset is unbalanced (the number of patients that expired is 8.9 times those who didn’t. Therefore, the number of negative labels will be much larger than the number of positive ones. So I had to use this line code to calculate the scale.



And then increase the weight of the positive values by using the ‘scale\_pos\_weight’ parameter, this worked for all of the classifiers except for NN.



For NN, I had to pass the scaling into the loss function.



* Hyperparameters tuning:

For the classifiers that performed well by default, I started tuning the hyperparameters of those classifiers. These classifiers include: Random Forest, XGBoost and LightGBM. And all 3 of those were tuned using k-folds cross-validation, more specifically:

* + Random Forest and XGBoost were tuned using **RandomizedSearchCV** from the *sklearn.model\_selection* package.



* + LightGBM were tuned using **BayesSearchCV** from the *skopt* package.



* + In k-fold cross-validation, the dataset is divided into k equal-sized folds. The model is trained on k-1 folds and evaluated on the remaining fold. This process is repeated k times, with each fold being used as the validation set once. The average performance across all k folds is then calculated to give an estimate of how well the model will generalize to new data.
    - For RandomSearchCV, we can specify a range of values that each hyperparameter can take. Then the algorithm will randomly choose a set of hyperparameters in that range and starts training and evaluating on the folds. The number of times this process happens is specified by the user. After the search is finished, the model with the best-performing hyperparameters will be returned.
    - For BayesSearchCV, the process is almost the same as RandomSearchCV. With the only difference is that parameters are not randomly chosen, but rather picked using **Gaussian Process**. This helps to find the optimal set of hyperparameters faster.
    - Details on what parameters were tuned in each classifier will be given in part 4.

1. **Classification techniques used and results:**

Here’s a list of all of the classifiers that I used: Support Vector Classifier (LinearSVC from *sklearn.svm*), Random Forest (RandomForestClassifier from *sklearn.ensemble*), K-nearest-neighbors (KNeighborsClassifier from *sklearn.neighbors*), Neural Network (*pytorch*) and Gradient Boosted Tree (*XGBoost* and *LightGBM*).

Now I’ll go through the code and results of each classifier:

I’ll start with the classifiers that didn’t perform too well first (based on F1 macro score).

* Neural Network:

This was my third worst performing classifier, mainly because I’m quite inexperienced with neural networks, and I don’t think neural networks can perform too well on this dataset. Though, it did yield some results.

* + First, I want to talk about the structure of the NN. Here’s the code and a graph detailing the structure of my NN:

A screen shot of a computer program

Description automatically generatedA diagram of a software process

Description automatically generated

* + - The activation function for the linear layers is the **Rectified Linear Unit (ReLU)** function.
    - The input is split into categorical and numerical.
    - Because the NN can’t interpret categorical data even when encoded, I had to run the categorical features through a process called “embedding”. Embedding is essentially turning different classes in a feature into vectors of numbers, making it possible for the NN to interpret. The embeddings will then go through a linear layer.
    - For the numeric features, because the NN is being trained in batches, it’s best to normalize the values in that batch to not create any bias.
    - After the NN has finished preprocessing both types of features, the 2 inputs will be combined into a single input.
    - The concatenated input will go through a linear layer and a dropout layer. The dropout layer randomly drops a certain percentage of the input (in this case, 20% of the input). This is to prevent the NN from overfitting.
    - Lastly, after dropping out, the input will go though another linear layer, and then finally output.
  + Secondly, I want to go through the training process of the NN:

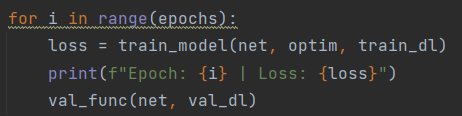
I used the **Adam** optimizer in the *torch* package for my NN.



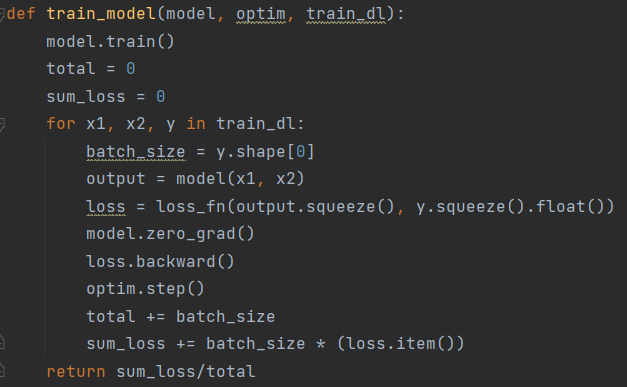
The learning rate starts at 0.015 but will decay as the training goes on.

The NN was trained through 200 epochs.





The train function:



Simply put, this function performs the forward and backpropagation phase of the NN and returns the loss.

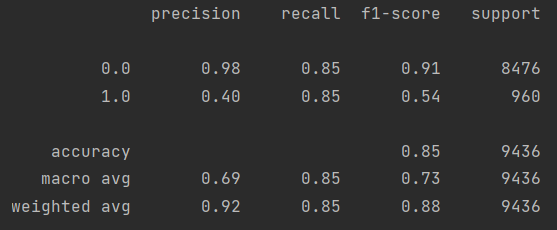
The validation function:

A screen shot of a computer program

Description automatically generated

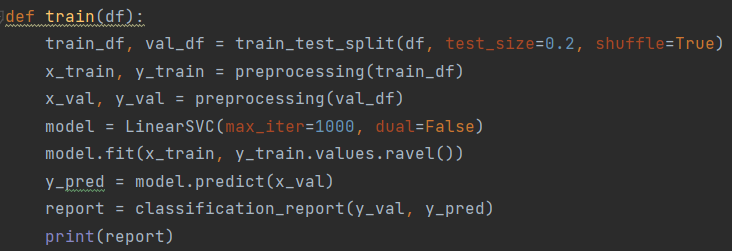
This function reports the results of the model after each epoch.

* + Results:

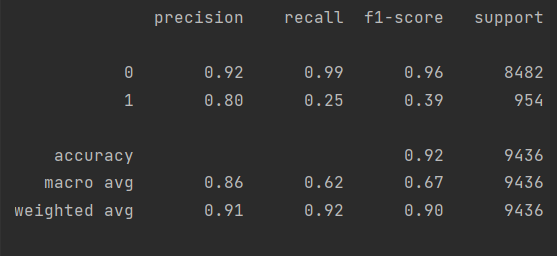


We can see that the NN didn’t do too well on the F1 macro score, but it at least performed better than the baseline. This could be because I didn’t spend too much time on the NN as I didn’t think it could perform as well as other classifiers.

* LinearSVC:
  + The code:



* + The results:

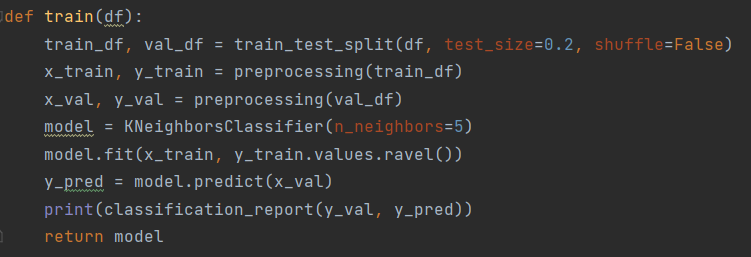


As you can see from the results, this was my worst classifier with only 0.67 on the f1 macro score, this could be because I didn’t bother optimizing the classifier, mainly because the default model didn’t do as well as the ensemble classifiers. So I thought my time would be better spent on experimenting with those.

* K-nearest-neighbors:

This is the last of the bad performing classifiers. There didn’t seem to be any parameters that I could tune aside from choosing the number of neighbors. But after some trial and error, the optimal number of neighbors still performed quite poorly.

* + The code:



* + The results:

A screenshot of a computer

Description automatically generated

The model scored only 0.7 on f1 macro.

Now we move on to the classifiers that actually performed well.

* Random Forest:

Process of making the Random Forest Classifier:

1. After preprocessing the data, I initialized the **RandomForestClassifier** with a class\_weight parameter to deal with the unbalanced dataset.

A black background with white text

Description automatically generated

1. Hyperparameters tuning:

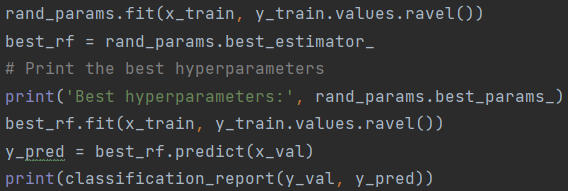
A computer screen shot of a code

Description automatically generated

I used **RandomizedSearchCV** to find the most optimal parameters for the Random Forest.

* + For this classifier, I only tuned two parameters: *n\_estimators* (number of trees to be used in the model) and *max\_depth* (the maximum depth a tree can reach).
  + The parameters distribution specifies the range of the parameter that the randomized search can choose. E.g:

*‘n\_estimators’: randint(50, 200)* means the search can choose a random value between 50 and 200 to be the number of trees used.



* + After training and evaluating 50 randomly chosen sets of parameters on 5 folds. The **RandomizedSearchCV** will return the model with the best-performing hyperparameters.

A screen shot of a computer code

Description automatically generated

* + We can then fit the best

1. **The best classifier and how it works:**
2. **Kaggle Submission Score:**