-Priyanka Boral (MT19127)-Reecha Kumari Giri(MT19134)- Akanksha Dewangan(MT19049)

MLBA ASSIGNMENT-2

(Group-9)

★ ***Readme*:**

We have to give command like:

!python3 /content/group_9_priyankaboral_mt19127.py "kaggle_train.csv" "kaggle_test.csv" "output.txt"

A. To run the code named Priyanka_Boral_MT19127.py file with best MCC

(i) first we have to go at Scripts folder in our desktop where python is installed along with two test and train folder kaggle_train.csv, kaggle_test.csv and Priyanka_Boral_MT19127_first.py

(ii) Path where python is present:

C:\Users\akanksha\AppData\Local\Programs\Python\Python37\Scripts

(iii) **>ipython3 Priyanka_Boral_MT19127.py kaggle_train.csv kaggle_test.csv latest.csv metric_values.txt**

(iv)You get the output file **latest1.csv** where all prediction are present

(v) metrics were written in **metric_values.txt file**

Output file submitted using this code: latest1.csv

B. To run the code named Priyanka_Boral_MT19127.py file with 2nd best MCC

(i) first we have to go at Scripts folder in our desktop where python is installed along with two test and train folder kaggle_train.csv, kaggle_test.csv and Priyanka_Boral_MT19127_second.py

(ii) Path where python is present:

C:\Users\akanksha\AppData\Local\Programs\Python\Python37\Scripts

(iii) **>ipython3 Priyanka_Boral_MT19127.py kaggle_train.csv kaggle_test.csv latest2.csv metric_values.txt**

(iv)You get the output file **latest2.csv** where all prediction are present

(v) metrics were written in **metric_values.txt file**

Output file submitted using this code: latest2.csv

C. To run the code named Priyanka_Boral_MT19127.py file with 3rd best MCC

(i) first we have to go at Scripts folder in our desktop where python is installed along with two test and train folder kaggle_train.csv, kaggle_test.csv and Priyanka_Boral_MT19127_first.py

(ii) Path where python is present:

C:\Users\akanksha\AppData\Local\Programs\Python\Python37\Scripts

(iii) **>ipython3 Priyanka_Boral_MT19127_first.py kaggle_train.csv kaggle_test.csv latest3.csv metric_values.txt**

(iv)You get the output file **latest3.csv** where all prediction are present

(v) metrics were written in **metric_values.txt file**

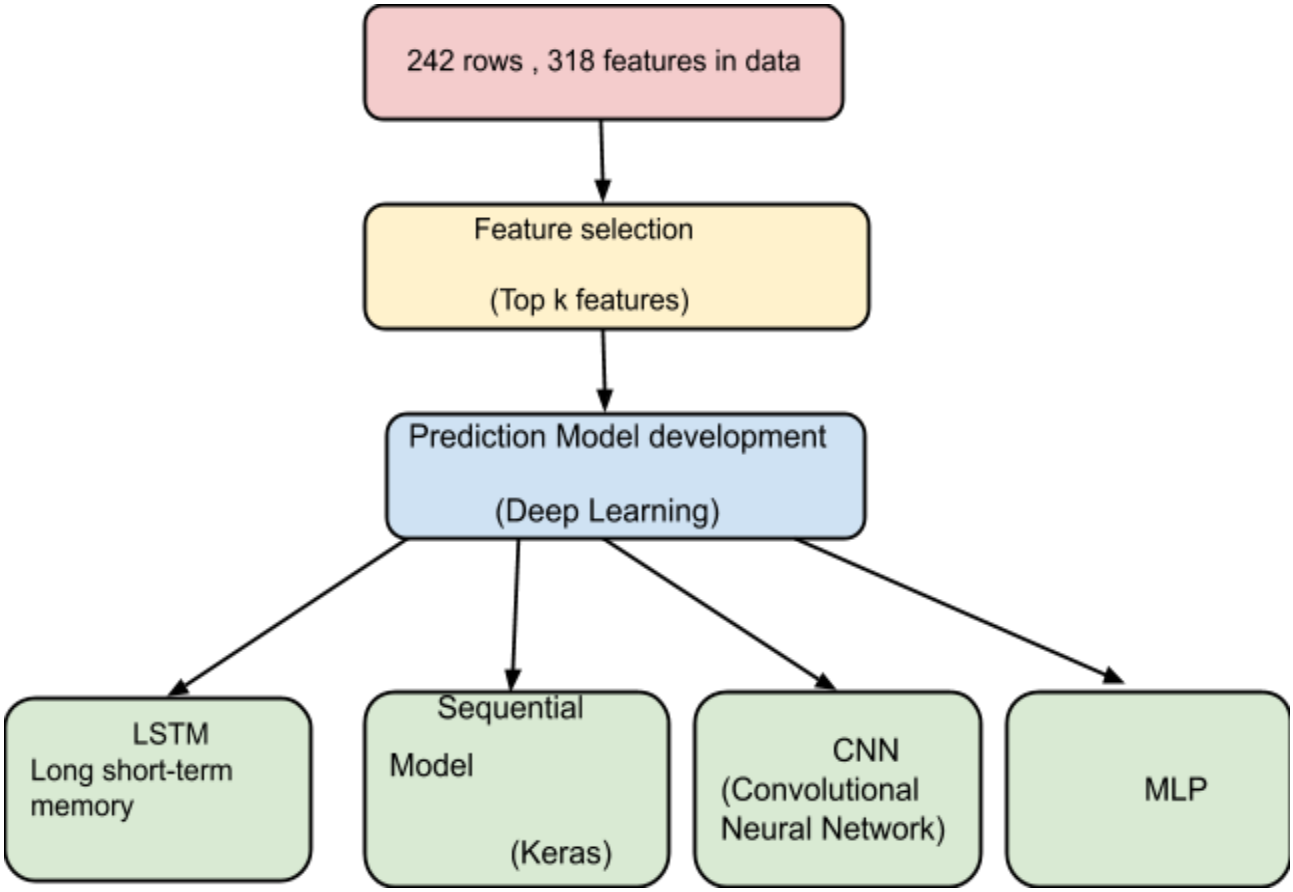
Output file submitted using this code: latest3.csv

Report contains information regarding all the techniques used for this assignment and analysis.

★ **Aim**: To develop deep learning models to classify high and low risk cancer patients.

★ Preprocessing and Methodology:

Complete description through flow chart



Preprocessing and Feature Selection:

1. Feature Selection:

(i) SelectKBest RFE(Recursive feature elimination) for 100 feature extraction

SelectKBest RFE(Recursive feature elimination) for 82 feature extraction

SelectKBest RFE(Recursive feature elimination) for 64 feature extraction

(ii) SelectKBest, f_classif sel_f = SelectKBest(score_func=f_classif, k=100)

SelectKBest, f_classif sel_f = SelectKBest(score_func=f_classif, k=82)

SelectKBest, f_classif sel_f = SelectKBest(score_func=f_classif, k=64)

2. Various models are developed for prediction of high and low risk cancer patients.

3. Finally after obtaining predictions from the developed model, their id's and labels are stored in .csv file.

Models developed using deep learning techniques:

1. **LSTM (Long short-term memory)**
2. **Sequential model (Keras)**
3. **CNN (Convolutional neural network)**
4. **MLP (Multi-layered perceptron)**

All the above models are giving different results on using various parameters and on various runs. Best validation scores which are obtained on the parameters are mentioned in the below table.

For validation MCC (Mathews correlation coefficient) is used.

Classifiers	**Validation Accuracy**	**Validation MCC**	**Leaderboard**

-Priyanka Boral (MT19127)-Reecha Kumari Giri(MT19134)- Akanksha Dewangan(MT19049)

1. **LSTM (** `model.add(LSTM((200),input_shape=(1,318),activation='relu',return_sequences=True))`	55.2	22	63.025
2. **Sequential Model (**without feature selection) `= keras.Sequential([keras.layers.Dense(120,activation='relu',input_shape=(318,))`	66	23	63.985
3. **Sequential Model** with feature selection= `keras.Sequential([keras.layers.Dense(120,activation='relu',input_shape=(318,))`	54.6	27.32	58.644
4. **CNN (Convolutional Neural Network using conv 1D)** Without feature selection `model.add(Conv1D(64, kernel_size=5, activation='relu', input_shape=(318,1)))`	53.06	17.6	50.120

5. CNN (Convolutional Neural Network using conv 1D) With feature selection `model.add(Conv1D(64, kernel_size=5, activation='relu', input_shape=(318,1)))`	52.3	16.4	54.355
MLPClassifier(solver='adam', alpha=1e-5, **hidden_layer_sizes=(1,318), random_state=0)**	51	2.4	-----

- For validation accuracy, cross validation K-fold (5-fold) has been used.

- In the above table besult results of each model are shown but are checked through different parameters like by changing various layers, dropouts, batches, epochs etc.

CONCLUSION:

- From above observation, it is concluded that in our case, **Sequential model (`keras.Sequential([keras.layers.Dense(120,activation='relu',input_shape=(318,)) with validation MCC=35.2.`**
- 2nd best result is obtained, using **Sequential model (`keras.Sequential([keras.layers.Dense(60,activation='relu',input_shape=(318,)) with validation MCC= 27.32.`**
- 3nd best result is obtained using CNN `model.add(Conv1D(64, kernel_size=5, activation='relu', input_shape=(318,1)))` with validation MCC=27.

Note: The results may vary on different parameters and on various runs.

-Priyanka Boral (MT19127)-Reecha Kumari Giri(MT19134)- Akanksha Dewangan(MT19049)