EC8020 – COMPUTER ENGINEERING DESIGN PROFECIENCY

DESIGN TASK 04 - MACHINE LEARNING

Group details: 2018/E/099, 2018/E/108, 2018/E/123

Problem statement

Breast cancer is the second most common cancer globally, particularly among women. Detecting it accurately is crucial for effective treatment. In this research, the goal was to classify two major types of breast cancer: benign cancer and malign cancer. Two machine learning methods, Naïve Bayes (NB) and K-Nearest Neighbors (KNN) were used for this binary classification and got the accuracies respectively 0.961932 and 0.975109.

Dataset

For the study, Wisconsin breast cancer datasets was used. It consists of 699 clinical cases, with 11 attributes. There we 16 missing data were identified. So, the dataset was limited to 683 samples.

Attributes:

i. ID
ii. Clump thickness
iii. Uniformity of cell size
iv. Uniformity of cell shape
viii. Marginal adhesion
ix. Single epithelial cell size
x. Bare nuclei
xi. Bland chromatin
v. Normal nucleoli
vi. Mitoses
vii. Cancer class (Benign or Malign)

Procedure

01. Data collection and analysis

i. First, the dataset was downloaded from the repository.

ii. Checked the attributes and shape of the dataset. There were 699 samples, and 11 attributes included this dataset. The shape of the dataset was (699,11)

iii. Checked duplicated records and null values. There were 8 duplicated samples and 16 null values.

02. Data preprocessing

i. Removed ID column from the dataset.

ii. Removed null values.

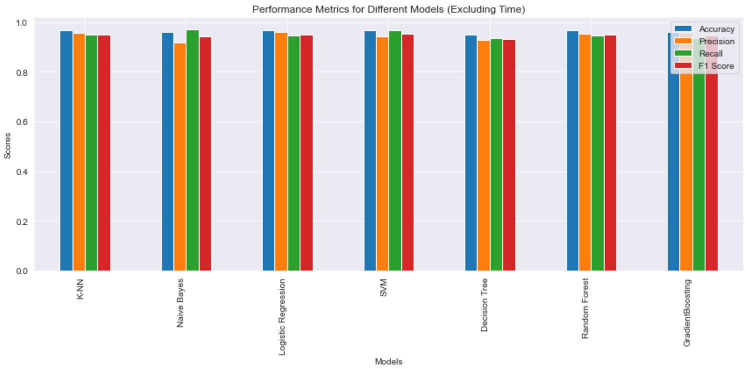
iii. Map target value (2 -> 0 and 4-> 1)

03. Model implementation

i. For the model implementation we have chosen 5 additional machine learning models with NB and KNN. K – fold cross validation was used along with all these classification models. (K=5)

Used machine learning models:

- Logistic Regression (LR)
- Support Vector Classifier (SVC)
- Decision Trees (DT)
- Random Forest (RF)
- Gradient Boosting (GB)



04. Hyperparameter Optimization

LR : 'C': 10, 'penalty': 'l1', 'solver': 'liblinear'
SVC: 'C': 0.1, 'gamma': 'scale', 'kernel': 'rbf'
RF: 'max_depth': 30, 'min_samples_leaf': 2, 'min_samples_split': 10, 'n_estimators': 200
DT: 'criterion': 'gini', 'max_depth': 20, 'min_samples_leaf': 4, 'min_samples_split': 10
GB: 'learning_rate': 0.2, 'max_depth': 5, 'min_samples_split': 2, 'n_estimators': 150

05. Performance comparison

ML models	Accuracy	Precision	Recall	F1-score	Processing Time (s)
NB	0.959049 ±0.017577	0.919158 ±0.04212	0.970656 ±0.010427	0.943665 ±0.022804	0.171842
KNN	0.966402 ±0.025515	0.955835 ±0.042263	0.949911 ±0.058289	0.951497 ±0.038155	0.345099
LR	0.96637 ±0.0204	0.959318 ±0.040028	0.945656 ±0.036265	0.951735 ±0.029125	0.203113
SVC	0.96784 ±0.018769	0.944107 ±0.036586	0.966578 ±0.024956	0.954864 ±0.026096	0.347583
RF	0.96638 ±0.020914	0.952336 ±0.039493	0.945656 ±0.048562	0.949216 ±0.020981	4.420627
DT	0.951717 ±0.016988	0.927205 ±0.036019	0.937323 ±0.026214	0.931733 ±0.023412	0.153057
GB	0.96199 ±0.024561	0.955559 ±0.043025	0.937323 ±0.046336	0.947414 ±0.036736	1.343661

Conclusion

- According to the results obtained from above mentioned machine learning models, SVC has the highest accuracy of 0.96784 ± 0.018769.