INNOVACCER HACKER CAMP -19

ASSIGNMENT :

MADE BY: SAMIDHA VERMA

DATASET – 1(TRIP ADVISOR)

PREPROCESSING:

- Changing the value of months and weeks to labelled data
- Changing minimum value of Member years from -1806 to 0
- Correcting 3,5 and 4,5 values Hotel Stars to 3.5 and 4.5 respectively.
- One hot encoding the string values

MODELS USED:

None of the models could give an accuracy of greater than 53.4% with this dataset. This may be due to the fact that the **dataset is very small** and **highly unbalanced** with 11 ,30 ,72 ,164, 227 out of 504 being the distribution of Score 1, 2, 3, 4 and 5 hotels respectively.

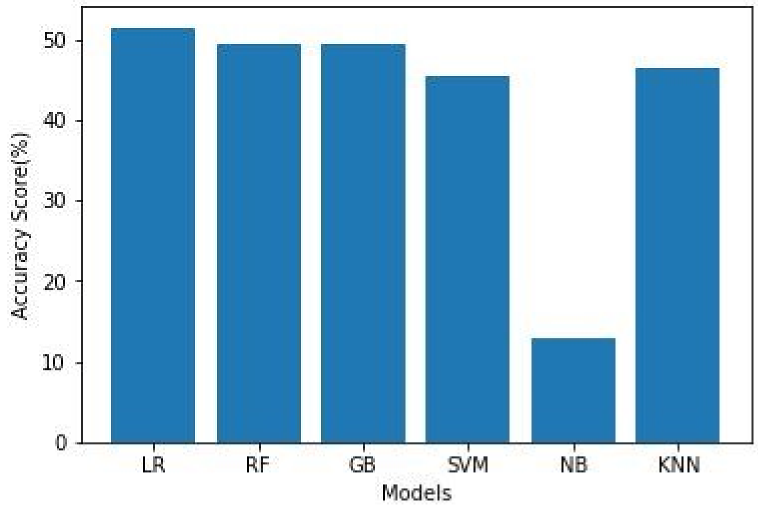
I tested the dataset with **Logistic Regression** because it helps in giving a fair idea of the non linear decision boundaries and is the first and the simplest of all classifiers. It outperformed even random forests in certain cases.

Random Forests and Gradient Boosting performed quite well in comparison to SVM and Naive Bayes with an **accuracy_score ranging between 43.5% to 53.4% on tuning**.

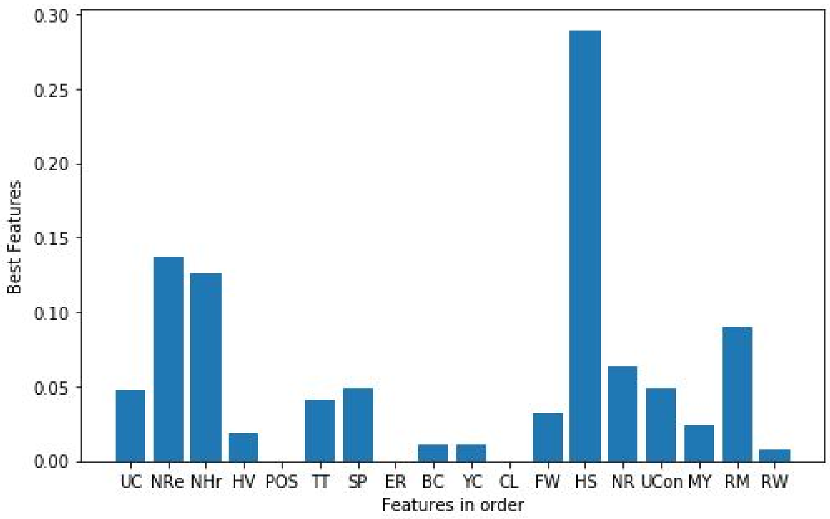
KNN being able to form clusters of data points lying close together has been able to perform with an **accuracy_score of 42% to 49% on average**.

I deliberately used, **Naive Bayes** on this dataset to show that the **features are dependant on each other unlike what Naive Bayes assumes**, and hence, it performed very bad with an **accuracy_score of about 11.2% to 13.1%** on average

MODEL COMPARISON



FEATURE IMPORTANCE GRAPH



FEATURE KEYS

USER COUNTRY - UC
NR. REVIEWS - NRE
NR. HOTEL REVIEWS - NHR
HELPFUL VOTES - HV
PERIOD OF STAY - POS
TRAVELER TYPE - TT
SWIMMING POOL - SP
EXERCISE ROOM - ER
BASKETBALL COURT - BC
YOGA CLASSES - YC
CLUB - CL
FREE WIFI - FW
HOTEL STARS - HS
NR. ROOMS - NR
USER CONTINENT - UCON
MEMBER YEARS - MY
REVIEW MONTH - RM
REVIEW WEEKDAY - RW

DATASET – 2 (BREAST CANCER)

KEY FOCUS: To pick the model that gives best recall value, since in the case of breast cancer classification **reducing false negatives should be our main motive**. For this in the first half I have rated SVM the best, as it gave the highest recall score.

SOME OBSERVATIONS:

- **Logistic Regression** – Logistic Regression being the most basic classifier came to my mind first as I wanted to keep my model as simple as possible and this algorithm gives a fair idea of the decision boundary as well. It performed well with an **accuracy of 97%** and a **recall of about 96.1%.**

- **Keras** – Gives excellent **accuracy of 96.5%,** but it's **recall value is 2.7%** which is very bad. This may be because of **lack of data** and the data being **unbalanced** with 458 out of 699 patients having a benign tumor leading to **underfitting** of data.

- **Random Forests and Gradient Boosting** – These, after SVM were the best performers, Gradient Boosting being slightly better than Random Forests and had **a recall value of 92.3% and 96.2% respectively** on average.

- **SVM** – It is dependent on the data points rather than the features and hence performs specially well in cases of binary classification. It had an average case **accuracy of 96.4%** and **recall of greater than 97%** which was the best among all models in almost all cases.

- **Naive Bayes -** Although Naive Bayes have the assumption that features are independent of each other, upon cross validating many times practically as well as on reading certain research papers, I concluded that it has also performed fairly well with this dataset giving a **recall value of about 96%** on average.

TUNING RANDOM FOREST TO INCREASE SENSITIVITY:

In the second half of the assignment, I mainly focused **increasing the sensitivity of the model** because missing out on the fact that a person has a malignant tumor can be disastrous therefore should be the key focus of our model

I used Random Forests here particularly because its features can be tuned easily. I have tried to reduce the number of false negatives down to 1 on a testing data of size 140 by using **GridSearchCV** and tuning the parameters as follows (based on the result when keeping **recall_score** as the scorer):

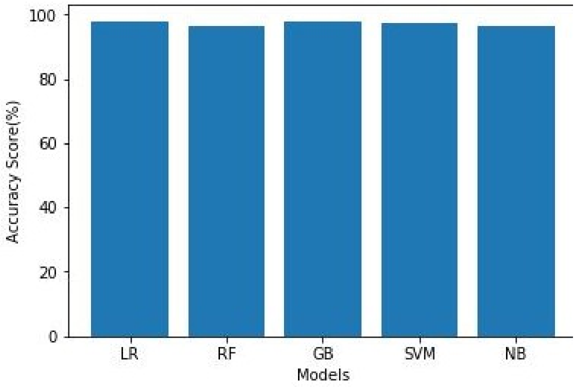
- max_depth: 15
- max_features: 3
- min_samples_split: 3
- n_estimators: 100

However, this result was the same even when the scorer was **precision_score**.

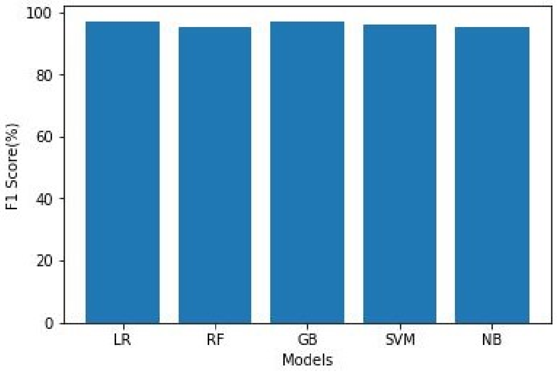
Hence, to know the value of the operation point I used **predict_proba()** on the random forest classifier in order to further use the result to plot the precision_recall_graph.

I found out that the **recall would be 1.0** for this tuned random forest classifier when the **threshold is 0.28333333**, but the **precision would fall down to 0.96296296.** The **balance** between the recall and precision was reached for **threshold value 0.33083333** when **both the scores were 0.98076923**. ***So, I concluded that taking <u>threshold as 0.33083333</u> is the better option.***

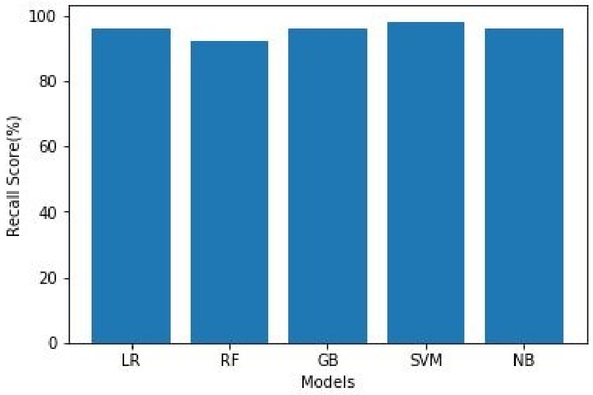
MODEL COMPARISON



ACCURACY SCORE



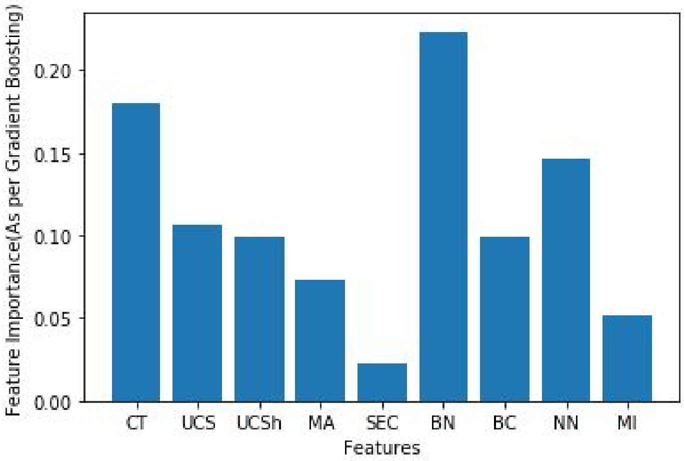
F1 SCORE



RECALL SCORE

FEATURE IMPORTANCE GRAPH

CT - CLUMP THICKNESS
UCS - UNIFORMITY OF CELL SIZE
UCSH - UNIFORMITY OF CELL SHAPE
MA - MARGINAL ADHESION
SEC - SINGLE EPITHELIAL CELL SIZE
BN - BARE NUCLEI
BC - BLAND CHROMATIN
NN - NORMAL NUCLEOLI
MI - MITOSES



VALUES OF FALSE POSITIVES AND FALSE NEGAIVES

MODEL	FALSE POSITIVES	FALSE NEGATIVES
LOGISTIC REGRESSION	2	5
RANDOM FOREST CLASSIFIER	2	6
GRADIENT BOOSTING	2	4
SVM	2	2
NAIVE BAYES	2	3
RANDOM FOREST (Tuned with GridSearchCV and predict_proba)	2	0