HOPE AI ASSIGNMENT - 5
Classification

GitHub Link for dataset:
https://github.com/JayachandraPrabha/Assignment-5-Classification/blob/main/CKD.csv

Problem Statement / Requirement:

A requirement from the Hospital Management asked us to create a predictive model which will predict Chronic Kidney Disease (CKD) based on several parameters. The Client has provided the dataset of the same.

1.) Identify your problem statement

2.) Tell basic info about the dataset (Total number of rows, columns)

3.) Mention the pre-processing method if you're doing any (like converting string to number – nominal data)

4.) Develop a good model with a good evaluation metric. You can use any machine learning algorithm; you can create many models. Finally, you have to come up with a final model.

5.) All the research values of each algorithm should be documented. (You can make a tabulation or screenshot of the results.)

6.) Mention your final model, justify why you have chosen the same. .

---

1. Identify the Problem statement:

As the Hospital management wants to **predict Chronic Kidney Disease (CKD)** with the provided dataset.

A model has to be developed which will predict Chronic Kidney Disease (CKD).

Approach:

i) Stage-I → ML (Dataset in Numerical format)

ii) Stage-II → Supervised Learning (requirement is clear)

iii) Stage-III → Classification (Output is in textual format (Yes/No type))

2. Basic Information about the dataset:

Total number of rows: 399

Total number of columns: 28

Shape of the dataset: (399, 28)

3. Pre-processing method:

converting string to number→ nominal data

As the dataset contains various (yes/no, present/not present, normal/abnormal) types of data. Inorder to convert the incomparable categorical/string data to numerical data by using **one hot encoding method.**

4. Good model with good evaluation metrics (roc_auc_score) [Reverse operating characteristics area under curve] score:

Using machine learning algorithms, finally coming up with a good model.

NaiveBayes Classifier	**GaussianNB**	MultinomialNB	ComplementNB	**BernoulliNB**
roc_auc_value	**1.0**	0.8776	0.8776	**1.0**

1. Support Vector Machine (SVC)
   → roc_auc_score value: 0.8631863171770662

2. Decision Tree Classifier (DTC)
   → roc_auc_score value: 0.9733333333333334

3. KNN Classifier
   → roc_auc_score value: 0.8542222222222222

4. K-means clustering
   → roc_auc_score value: 0.

5. **Random Forest (RC)**
   → roc_auc_score value: 1.0

6. Logistic Regression
   → roc_auc_score: 0.9986979166666666

ML algorithms: **Random Forest, Gaussian NB, Bernoulli NB** whose roc_auc_score values are selected, hence declared as the final model.

5. Research values (Classification_report & roc_auc_score values of the models):
Screen snip Link:
https://docs.google.com/document/d/13nNRpfYoaVd-t9Z5Rodirky4k559e3QnwEqHB2clIUQ/edit?usp=sharing

6. Mention your final models, justify why you have chosen the same?

The finalized models are **Gaussian Naive Bayes(NB), Bernoulli NB and Random Forest.** After analyzing with various algorithms and tuning its hyper/tuning parameters whose roc_auc_score values were as follows:

S.No	Name of the Algorithm	roc_auc_score value	Model output
1	**Gaussian Naive Bayes (NB)**	**1.0**	**Good**
2	Multinomial NB	0.8776	Poor
3	Complement NB	0.8776	Poor
4	**Bernoulli NB**	**1.0**	**Good**
5	Support Vector Machine (SVC)	0.8631863171770662	Moderate
6	Decision Tree Classifier (DTC)	0.9733333333333334	Moderate
7	**Random Forest (RF)**	**1.0**	**Good**
8	KNN Classifier	0.8542222222222222	Moderate
9	Logistic Regression	0.9986979166666666	Moderate

Conclusion:

In conclusion, the best models obtained are

1. **Gaussian Naive Bayes(NB)**
2. **Bernoulli NB and**
3. **Random Forest**.

The **Gaussian Naive Bayes(NB), Bernoulli NB and Random Forest** algorithms provided the roc_auc_score values are 1.0 (nearly 100% of the accuracy).
Hence **Gaussian Naive Bayes(NB), Bernoulli NB and Random Forest** machine learning classification algorithms were finalized as the best models.
