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

1.) Identify your problem statement

Domain Selection:
The dataset consists of numerical input features, with only a few being categorical. However, during preprocessing, we can convert the categorical features into numerical ones. Hence, we categorize this domain as Machine Learning.

Learning Selection:

In predicting Chronic Kidney disease, our dataset comprises multiple input features and a single target output of classification type (Yes/No). We can use supervised learning, bcoz we have clearly defined input and output requirements. and type as classification

Machine Learning (Supervised Learning - Classification).

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

In Chronic Kidney Disease dataset we have **399 rows and 28 columns**

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

4 input feature has categorical nominal data in machine we cant able use categorical value directly so we convert to number by using one hot encoding. Also we remove the dummy section for this 4 input feature

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

In supervised learning classification within machine learning, we experimented with various models and metrics to identify the optimal one. Below, I will list the models and evaluation metrics used in this process:

Models:

Logistics Regression
Decision Tree Classifier
Random Forest Classifier

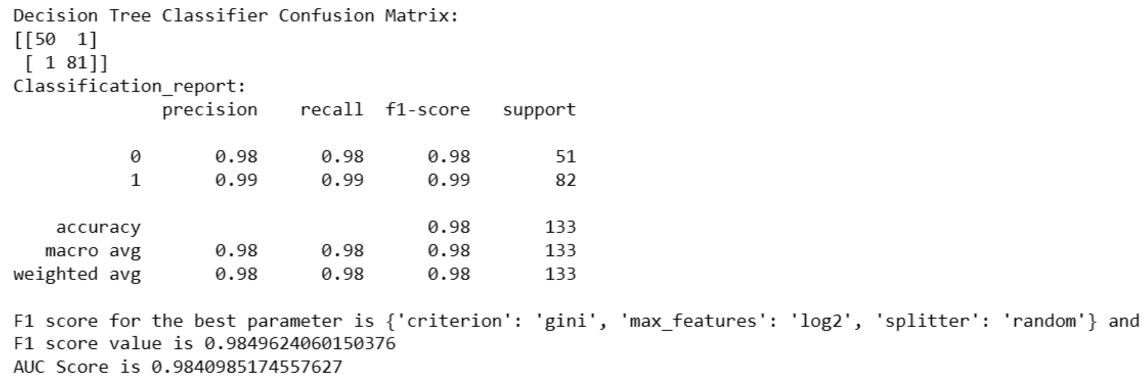
Naive Bayes (GaussianNB, Multinomial NB, ComplementNB, BernoulliNB, Categorical NB)
KNN
SVMC

Metrics:

Confusion Matrix
Classification Report
F1 Score
ROC AUC Score

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

1.Decision Tree Classifier



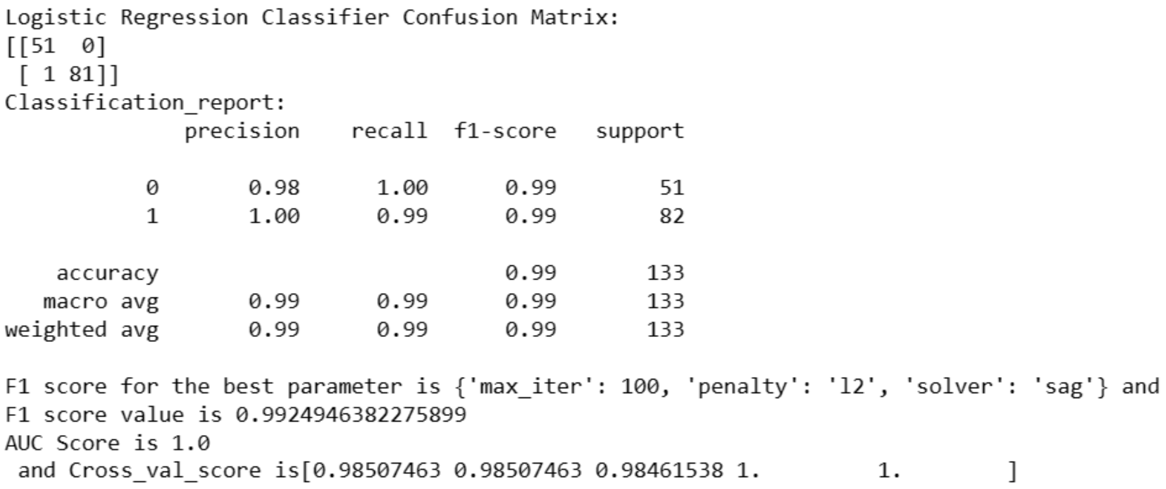
```
Decision Tree Classifier Confusion Matrix:
[[50  1]
 [ 1 81]]
Classification_report:
              precision    recall  f1-score   support

           0       0.98      0.98      0.98        51
           1       0.99      0.99      0.99        82

    accuracy                           0.98       133
   macro avg       0.98      0.98      0.98       133
weighted avg       0.98      0.98      0.98       133

F1 score for the best parameter is {'criterion': 'gini', 'max_features': 'log2', 'splitter': 'random'} and
F1 score value is 0.9849624060150376
AUC Score is 0.9840985174557627
```

2.Logistics Regression



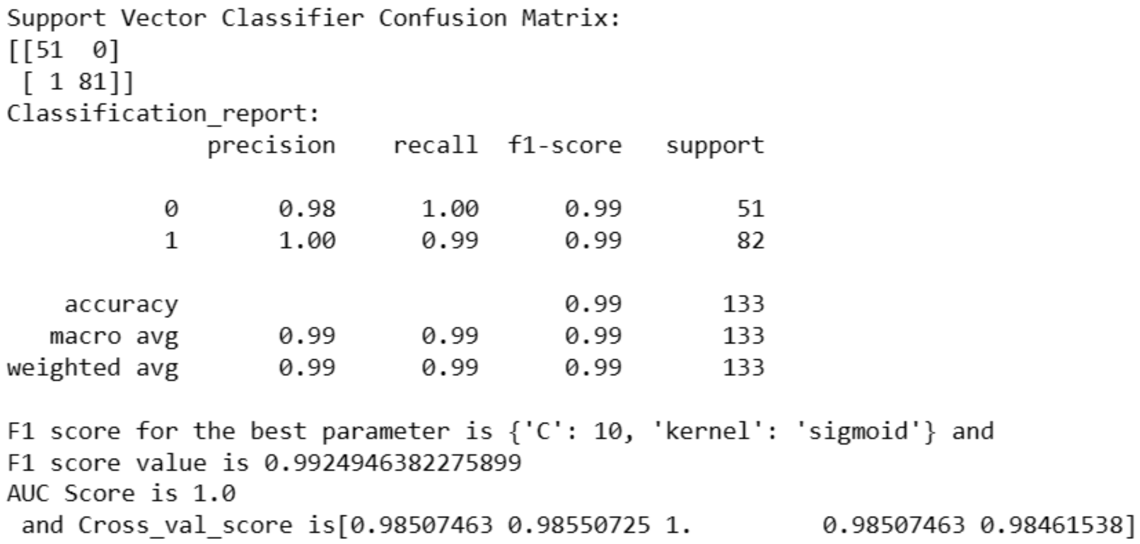
```
Logistic Regression Classifier Confusion Matrix:
[[51  0]
 [ 1 81]]
Classification_report:
              precision    recall  f1-score   support

           0       0.98      1.00      0.99        51
           1       1.00      0.99      0.99        82

    accuracy                           0.99       133
   macro avg       0.99      0.99      0.99       133
weighted avg       0.99      0.99      0.99       133

F1 score for the best parameter is {'max_iter': 100, 'penalty': 'l2', 'solver': 'sag'} and
F1 score value is 0.9924946382275899
AUC Score is 1.0
 and Cross_val_score is[0.98507463 0.98507463 0.98461538 1.         1.        ]
```

3. Support Vector Classifier



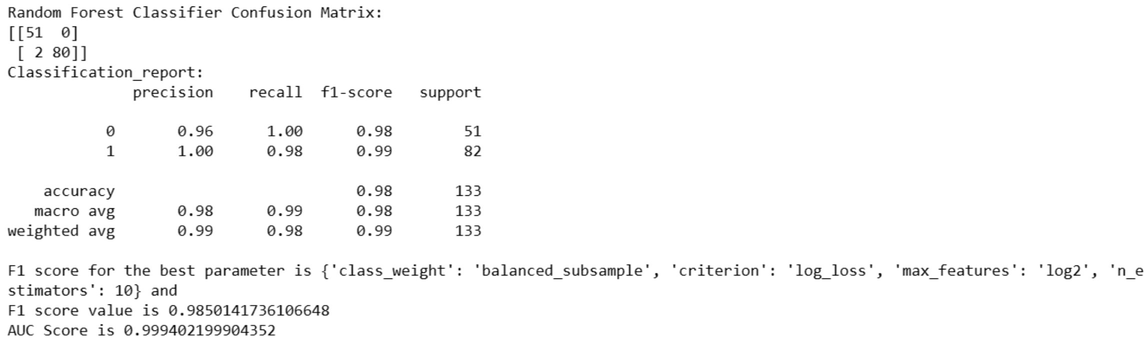
```
Support Vector Classifier Confusion Matrix:
[[51  0]
 [ 1 81]]
Classification_report:
              precision    recall  f1-score   support

           0       0.98      1.00      0.99        51
           1       1.00      0.99      0.99        82

    accuracy                           0.99       133
   macro avg       0.99      0.99      0.99       133
weighted avg       0.99      0.99      0.99       133

F1 score for the best parameter is {'C': 10, 'kernel': 'sigmoid'} and
F1 score value is 0.9924946382275899
AUC Score is 1.0
 and Cross_val_score is[0.98507463 0.98550725 1.         0.98507463 0.98461538]
```

4.Random Forest Classifier



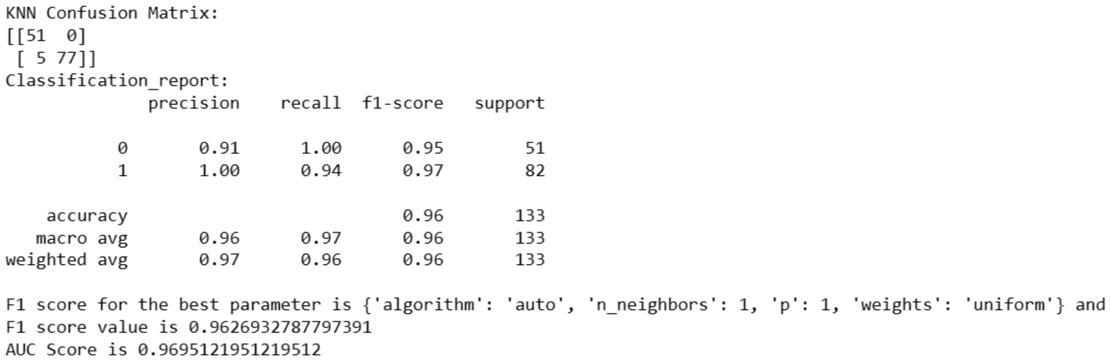
```
Random Forest Classifier Confusion Matrix:
[[51  0]
 [ 2 80]]
Classification_report:
              precision    recall  f1-score   support

           0       0.96      1.00      0.98        51
           1       1.00      0.98      0.99        82

    accuracy                           0.98       133
   macro avg       0.98      0.99      0.98       133
weighted avg       0.99      0.98      0.99       133

F1 score for the best parameter is {'class_weight': 'balanced_subsample', 'criterion': 'log_loss', 'max_features': 'log2', 'n_e
stimators': 10} and
F1 score value is 0.9850141736106648
AUC Score is 0.999402199904352
```

5.KNN



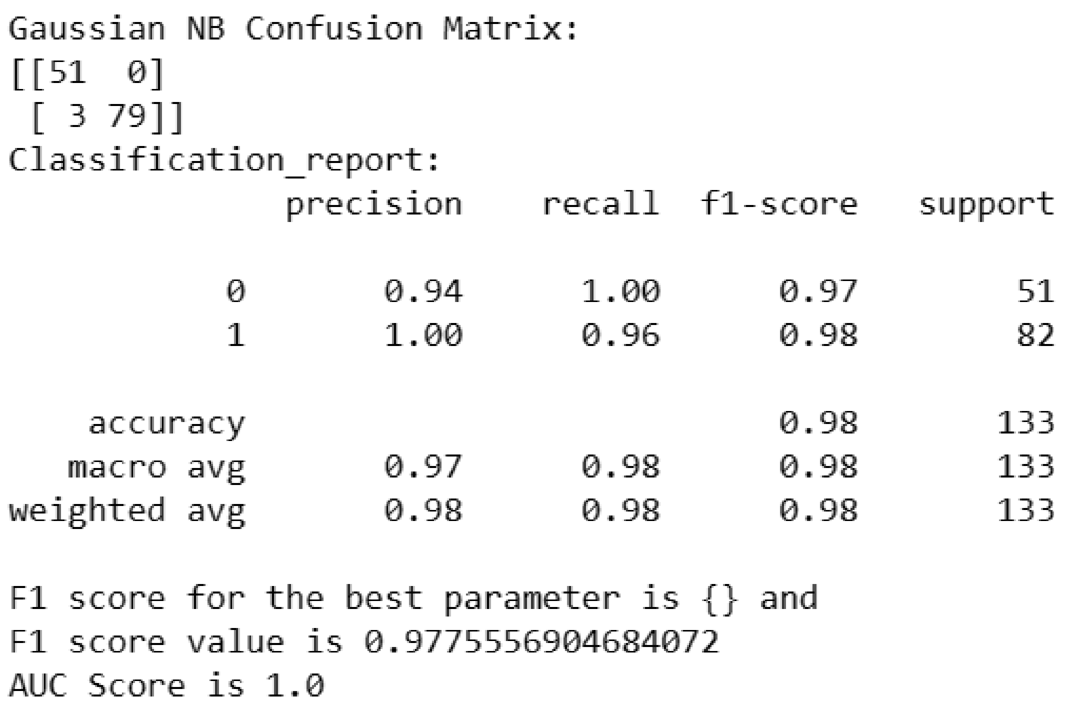
```
KNN Confusion Matrix:
[[51  0]
 [ 5 77]]
Classification_report:
              precision    recall  f1-score   support

           0       0.91      1.00      0.95        51
           1       1.00      0.94      0.97        82

    accuracy                           0.96       133
   macro avg       0.96      0.97      0.96       133
weighted avg       0.97      0.96      0.96       133

F1 score for the best parameter is {'algorithm': 'auto', 'n_neighbors': 1, 'p': 1, 'weights': 'uniform'} and
F1 score value is 0.9626932787797391
AUC Score is 0.9695121951219512
```

6. Naive Bayes

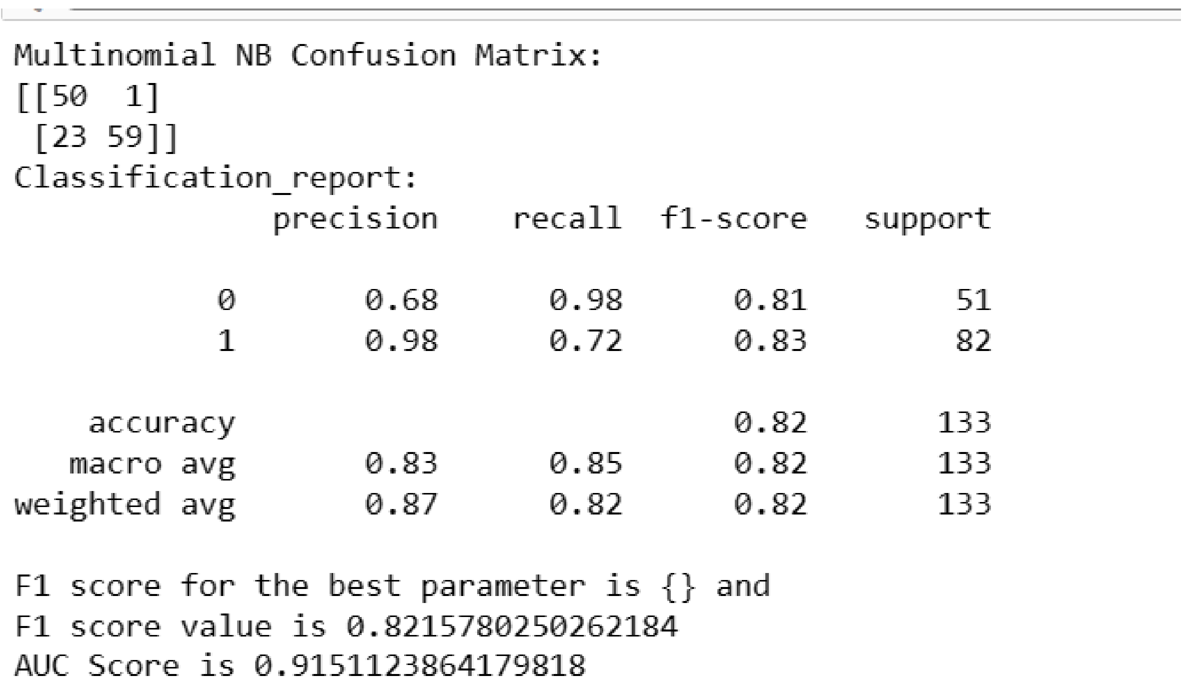


```
Gaussian NB Confusion Matrix:
[[51  0]
 [ 3 79]]
Classification_report:
              precision    recall  f1-score   support

           0       0.94      1.00      0.97        51
           1       1.00      0.96      0.98        82

    accuracy                           0.98       133
   macro avg       0.97      0.98      0.98       133
weighted avg       0.98      0.98      0.98       133

F1 score for the best parameter is {} and
F1 score value is 0.9775556904684072
AUC Score is 1.0
```

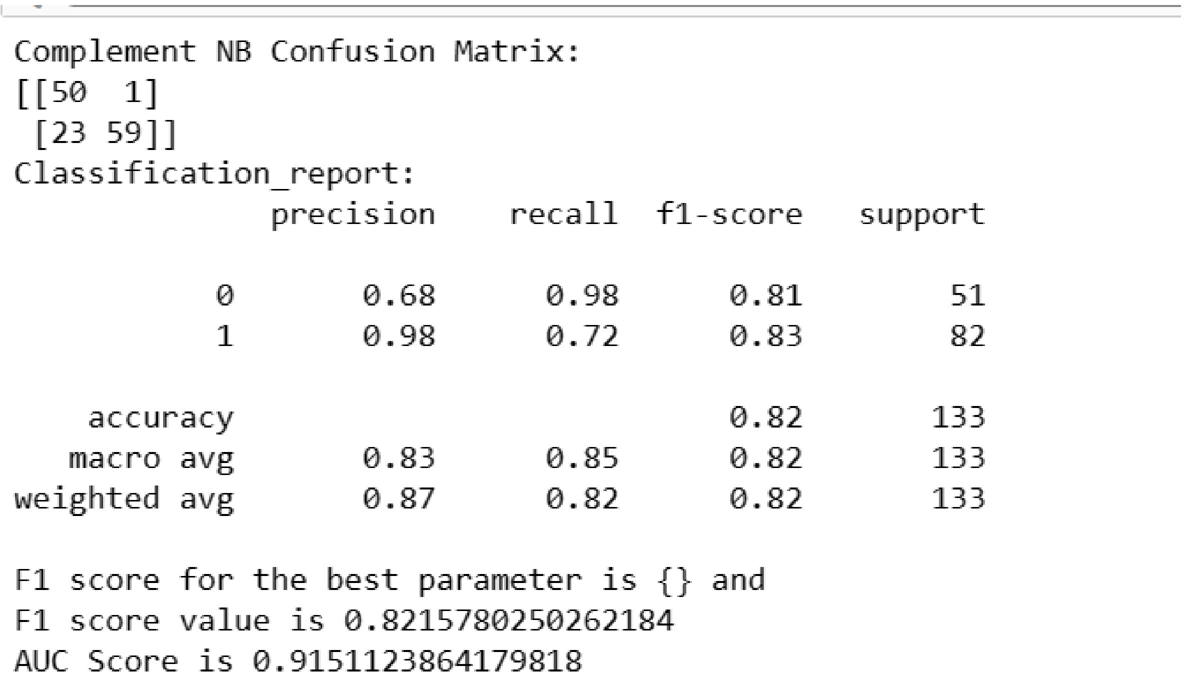


```
Multinomial NB Confusion Matrix:
[[50  1]
 [23 59]]
Classification_report:
              precision    recall  f1-score   support

           0       0.68      0.98      0.81        51
           1       0.98      0.72      0.83        82

    accuracy                           0.82       133
   macro avg       0.83      0.85      0.82       133
weighted avg       0.87      0.82      0.82       133

F1 score for the best parameter is {} and
F1 score value is 0.8215780250262184
AUC Score is 0.9151123864179818
```

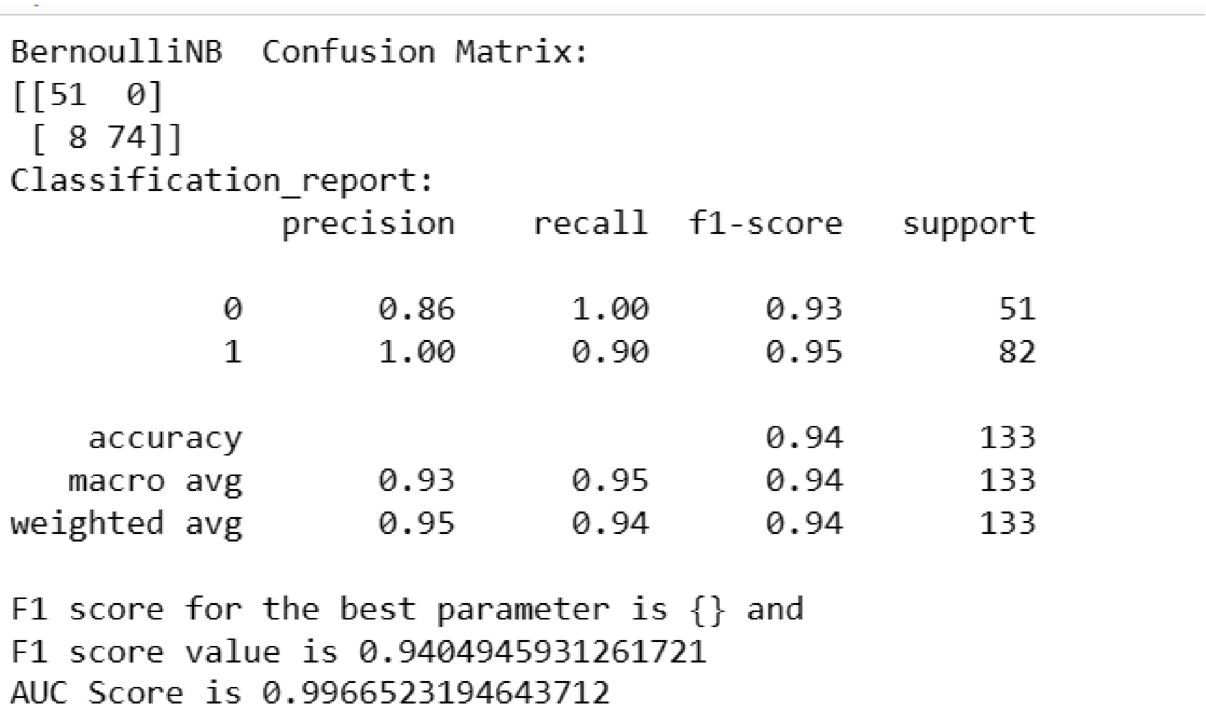


```
Complement NB Confusion Matrix:
[[50  1]
 [23 59]]
Classification_report:
              precision    recall  f1-score   support

           0       0.68      0.98      0.81        51
           1       0.98      0.72      0.83        82

    accuracy                           0.82       133
   macro avg       0.83      0.85      0.82       133
weighted avg       0.87      0.82      0.82       133

F1 score for the best parameter is {} and
F1 score value is 0.8215780250262184
AUC Score is 0.9151123864179818
```



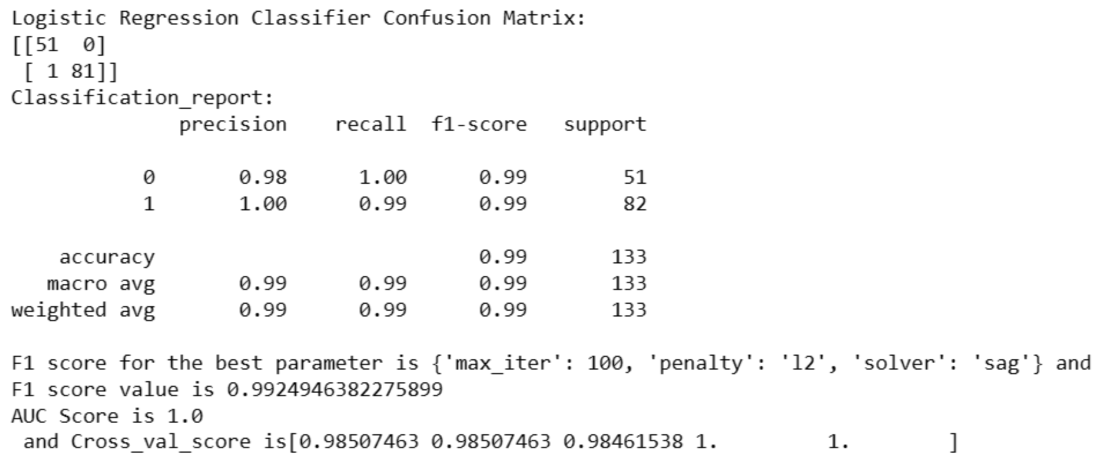
```
BernoulliNB  Confusion Matrix:
[[51  0]
 [ 8 74]]
Classification_report:
              precision    recall  f1-score   support

           0       0.86      1.00      0.93        51
           1       1.00      0.90      0.95        82

    accuracy                           0.94       133
   macro avg       0.93      0.95      0.94       133
weighted avg       0.95      0.94      0.94       133

F1 score for the best parameter is {} and
F1 score value is 0.9404945931261721
AUC Score is 0.9966523194643712
```

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

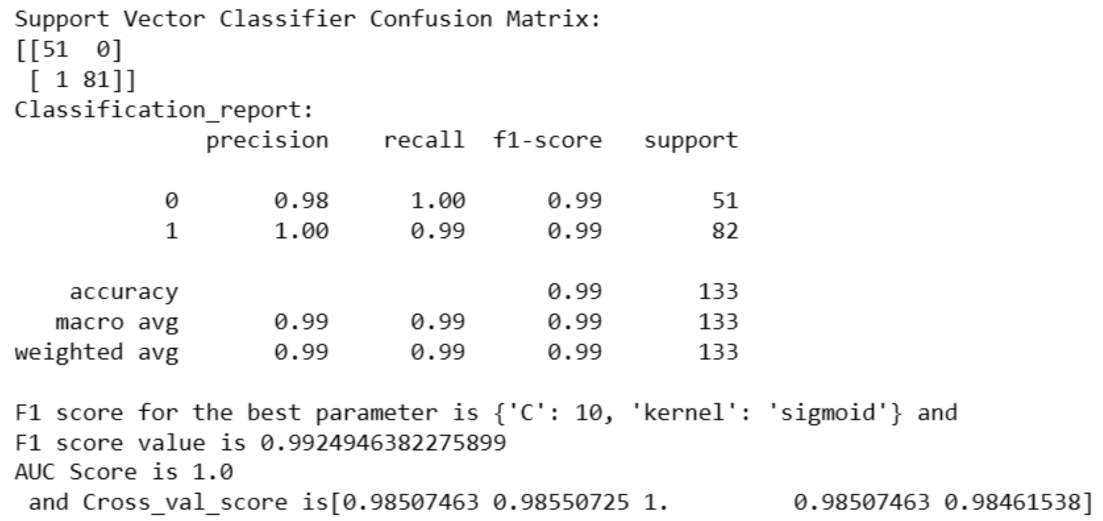


```
Logistic Regression Classifier Confusion Matrix:
[[51  0]
 [ 1 81]]
Classification_report:
              precision    recall  f1-score   support

           0       0.98      1.00      0.99        51
           1       1.00      0.99      0.99        82

    accuracy                           0.99       133
   macro avg       0.99      0.99      0.99       133
weighted avg       0.99      0.99      0.99       133

F1 score for the best parameter is {'max_iter': 100, 'penalty': 'l2', 'solver': 'sag'} and
F1 score value is 0.9924946382275899
AUC Score is 1.0
 and Cross_val_score is[0.98507463 0.98507463 0.98461538 1.         1.        ]
```



```
Support Vector Classifier Confusion Matrix:
[[51  0]
 [ 1 81]]
Classification_report:
              precision    recall  f1-score   support

           0       0.98      1.00      0.99        51
           1       1.00      0.99      0.99        82

    accuracy                           0.99       133
   macro avg       0.99      0.99      0.99       133
weighted avg       0.99      0.99      0.99       133

F1 score for the best parameter is {'C': 10, 'kernel': 'sigmoid'} and
F1 score value is 0.9924946382275899
AUC Score is 1.0
 and Cross_val_score is[0.98507463 0.98550725 1.         0.98507463 0.98461538]
```

Both the Logistic Regression and Support Vector Classifier (SVC) models seem to perform well based on this dataset. Here are some considerations:

Logistic Regression:

- High precision, recall, and F1-score for both classes (0 and 1).
- Perfect AUC score (1.0).
- Cross-validation scores are consistently high, indicating good model performance

Support Vector Classifier (SVC):

- Similar performance to Logistic Regression with high precision, recall, and F1-score for both classes.
- Perfect AUC score (1.0).
- Cross-validation scores are consistently high, suggesting good model performance

Most of the hospital system computation efficient and detailed explanation of information is very important to predict the Chronic Kidney Disease we must need to know about diagnosis very clearly without any misinformation in that case Logistic Regression is good model compared to Support Vector classifier