Objective:

The main purpose of this document is to identify the best model to predict the risk of Chronic Kidney Disease for an individual based on the given criteria in the dataset.

Requirements & Decisions:

Project Name	CKD Predictor
Dataset Filename	CKD.csv
Goal	To predict risk of Chronic Kidney Disease of an individual.

Problem Identification	
Stage1 (Domain Selection)	Machine Learning
Stage2 (Learning Method)	Supervised Learning
Stage3 (Data Type)	Classification

Dataset Info.	
No. of column	25
No. of rows	399

Research Values:

Based on the request, the dataset was imported and the models are created using different algorithms in machine learning. The accuracy of the models are captured and documented below.

Pre-processing	
Data Type	Nominal
Method	One Hot Encoding
Purpose	Converting the string to numerical data

Algorithms and Confusion Matrix Values:

Algorithms used to predict	1.Support Vector Classifier 2.Decision Tree Classifier 3.Random Forest Classifier 4.Logistic Regression 5.KNegibors Classifier 6.Gaussian Navies Bayes 7.Multinomial Navies Bayes 8.Bernoulli Navies Bayes 9.Categorical Navies Bayes 10.Complement Navies Bayes
Evaluation Metrics	Confusion matrix

The below tables shows the algorithms used to create the model and its respective evaluation metrics values.

CLASSIFICATION	Decision Tree		Random Forest		Support Vector		Logistic Regression		KNN	
	Risk of CKD		Risk of CKD		Risk of CKD		Risk of CKD		Risk of CKD	
	NO	YES	NO	YES	NO	YES	NO	YES	NO	YES
What is the % of correctly and wrongly classified class **(Precision)**	0.81	0.98	0.98	0.99	0	0.62	0.88	0.96	0.61	0.92
What is the % of correctly classified class **(Recall)**	0.98	0.87	0.98	0.99	0	1	0.93	0.92	0.91	0.65
What is the performance of the model when precision is high and recall is low and vice versa **(F1-Measure)**	0.89	0.92	0.98	0.99	0	0.77	0.9	0.94	0.73	0.77
What is the overall performance **(Accuracy)**	0.91		0.98		0.62		0.93		0.75	
What is the average % of correctly and wrongly classified class **(Macro Average-Precision)**	0.9		0.98		0.31		0.92		0.77	
What is the average % of correctly classified class **(Macro Average-Recall)**	0.92		0.98		0.5		0.93		0.78	
What is the average performance of the model when precision is high and recall is low and vice versa **(Macro Average-F1-Measure)**	0.91		0.98		0.38		0.92		0.75	
What is the proportion rate of correctly and wrongly classified class **(Weighted Average-Precision)**	0.92		0.98		0.39		0.93		0.81	
What is the proportion rate of the correctly classified class **(Weighted Average-Recall)**	0.91		0.98		0.62		0.93		0.75	
What is the proportion rate of the model when precision is high and recall is low and vice versa **(Macro Average-F1-Measure)**	0.91		0.98		0.48		0.93		0.75	

CLASSIFICATION	GaussianNB		MultinomialNB		BernoulliNB		CompletemntNB	
	Risk of CKD		Risk of CKD		Risk of CKD		Risk of CKD	
	NO	YES	NO	YES	NO	YES	NO	YES
What is the % of correctly and wrongly classified class **(Precision)**	0.96	1	0.67	0.98	0.85	1	0.67	0.98
What is the % of correctly classified class **(Recall)**	1	0.97	0.98	0.71	1	0.89	0.98	0.71
What is the performance of the model when precision is high and recall is low and vice versa **(F1-Measure)**	0.98	0.99	0.79	0.82	0.82	0.94	0.79	0.82
What is the overall performance **(Accuracy)**	0.98		0.81		0.93		0.81	
What is the average % of correctly and wrongly classified class **(Macro Average-Precision)**	0.98		0.82		0.92		0.82	
What is the average % of correctly classified class **(Macro Average-Recall)**	0.99		0.84		0.95		0.82	
What is the average performance of the model when precision is high and recall is low and vice versa **(Macro Average-F1-Measure)**	0.98		0.81		0.93		0.81	
What is the proportion rate of correctly and wrongly classified class **(Weighted Average-Precision)**	0.98		0.86		0.94		0.86	
What is the proportion rate of the correctly classified class **(Weighted Average-Recall)**	0.98		0.81		0.93		0.81	
What is the proportion rate of the model when precision is high and recall is low and vice versa **(Macro Average-F1-Measure)**	0.98		0.81		0.93		0.81	

Research Observation:

From the below classification report we absorbed that "**Random Forest**" and "**GaussianNB**" gave the similar values. In order to identify the best model the input was pre-processed using "**Standard Scaler**" then from the output below it is clear the **RANDOM FOREST CLASSIFIER** has the better accuracy.

After Preprocession the input with StandardScaler

RandomForestClassifier:

	precision	recall	f1-score	support
0	0.98	1.00	0.99	45
1	1.00	0.99	0.99	75
accuracy			0.99	120
macro avg	0.99	0.99	0.99	120
weighted avg	0.99	0.99	0.99	120

GaussianNB:

	precision	recall	f1-score	support
0	0.94	1.00	0.97	45
1	1.00	0.96	0.98	75
accuracy			0.97	120
macro avg	0.97	0.98	0.97	120
weighted avg	0.98	0.97	0.98	120

Without Preprocession the input with StandardScaler

RandomForestClassifier:

	precision	recall	f1-score	support
0	0.98	0.98	0.98	45
1	0.99	0.99	0.99	75
accuracy			0.98	120
macro avg	0.98	0.98	0.98	120
weighted avg	0.98	0.98	0.98	120

GaussianNB:

	precision	recall	f1-score	support
0	0.96	1.00	0.98	45
1	1.00	0.97	0.99	75
accuracy			0.98	120
macro avg	0.98	0.99	0.98	120
weighted avg	0.98	0.98	0.98	120

Conclusion:

Based on the above research table the model created using the algorithm **Random Forest Classifier** using the Hyper Factor parameter **'bootstrap': False, 'criterion': 'gini', 'max_features': 'sqrt', 'n_estimators': 500, 'warm_start': True** is having the below values and should be used for this project.