AML-Minor Assignment 1

Dataset: Wine data. Chosen dataset is good for classification models.
Link: http://archive.ics.uci.edu/ml/datasets/wine

Dataset description:
Abstract: Using chemical analysis determine the origin of wines
Number of attributes:13
Number of instances:178
Dataset characteristics: multivariate.
Missing values: None.
Class distribution: number of instances of class1: 59, class2: 71, and class3: 48.
Output:
C:\Users\bkokkula\Anaconda3\python.exe C:/Users/bkokkula/PycharmProjects/ML/BBBB1.py

Dataset head first 5 rows:

	alcohol	malic_acid	ash	...	hue	od280/od315_of_diluted_wines	proline
0	14.23	1.71	2.43	...	1.04	3.92	1065.0
1	13.20	1.78	2.14	...	1.05	3.40	1050.0
2	13.16	2.36	2.67	...	1.03	3.17	1185.0
3	14.37	1.95	2.50	...	0.86	3.45	1480.0
4	13.24	2.59	2.87	...	1.04	2.93	735.0

[5 rows x 13 columns]

Correlation within attributes:

alcohol	1.000000
proline	0.643720
color_intensity	0.546364
total_phenols	0.289101
magnesium	0.270798
flavanoids	0.236815
ash	0.211545
proanthocyanins	0.136698
malic_acid	0.094397
od280/od315_of_diluted_wines	0.072343
hue	-0.071747

nonflavanoid_phenols   -0.155929

alcalinity_of_ash   -0.310235

Name: alcohol, dtype: float64

Results:

rmse = 0.27261920869236034

r2 = 0.89282841263788

Linear Reg score: 0.89282841263788

Linear Reg Cross Validation: 0.8324810622343758

KNN Reg score: 0.6104441776710685

KNN Reg Cross Validation: 0.4250654942948383

KNN Model Confusion matrix:

   0 1 2

0 23 1 1

1 1 13 3

2 2 8 7

KNN Model score: 0.7288135593220338

KNN Model Cross Validation: 0.6810606060606061

SVC Model Confusion matrix:

   0 1 2

0 25 0 0

1 2 13 2

2 0 1 16

SVC Model score: 0.9152542372881356

SVC ModelCross Validation: 0.95

Ensemble RF Model Confusion matrix:

```
   0  1  2
0 25  0  0
1  0 16  1
2  0  0 17
```

Ensemble RF Model score: 0.9830508474576272

Ensemble RF Model Cross Validation: 0.9833333333333332

Neural Network Model Confusion matrix:

```
   0  1 2
0 22  3 0
1  0 17 0
2  1 16 0
```

Neural Network Model score: 0.6610169491525424

Neural Network Model Cross Validation: 0.7916666666666667

The accuracy of each model using cross validation and a conclusion which outline **Ensemble RF model** is the best model for the chosen dataset due to best accuracy of 98%.

 See below the accuracy of each model and ranks.

	Model	CV_value	Accuracy
4	Ensemble RF Model:	0.983333	0.983051
3	SVC Model:	0.950000	0.915254
0	Linear Reg Model:	0.832481	0.892828
2	KNN Model:	0.681061	0.728814
5	Neural Network Model:	0.791667	0.661017
1	KNN Reg Model:	0.425065	0.610444

Process finished with exit code 0

Conclusions: By seeing above accuracy of each model and rank that suggests the ensemble random forest performed the best with this type of dataset. Random forest applies an ensemble algorithm called bagging method to the decision trees, which helps reduce variances and overfitting. This technique strengthen our prediction accuracy. As a result, by using different classifications models the wine dataset have been classified successfully with best accuracy.