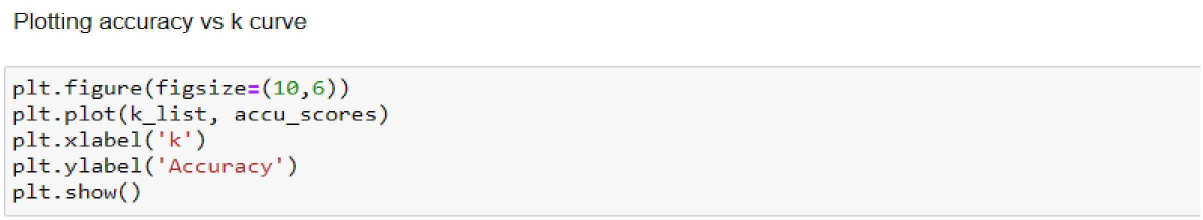
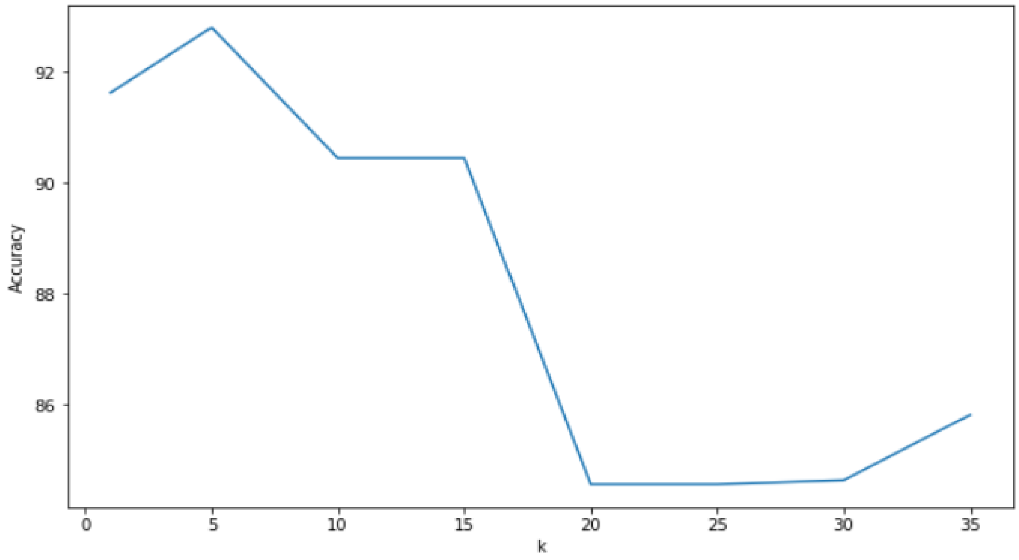
[CM5] i) **For Iris**, Accuracy vs K curve is plotted,

5-fold Cross Validation

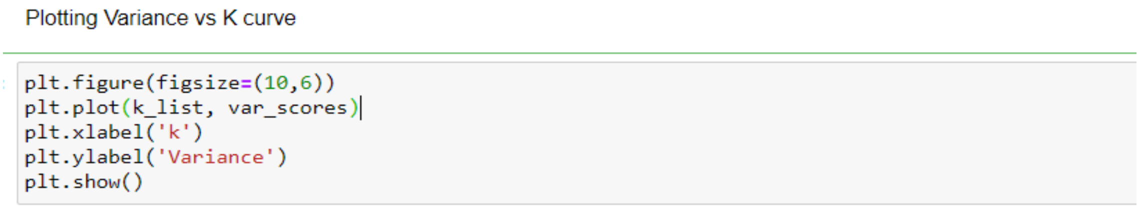


Plotting accuracy vs k curve

```
plt.figure(figsize=(10,6))
plt.plot(k_list, accu_scores)
plt.xlabel('k')
plt.ylabel('Accuracy')
plt.show()
```

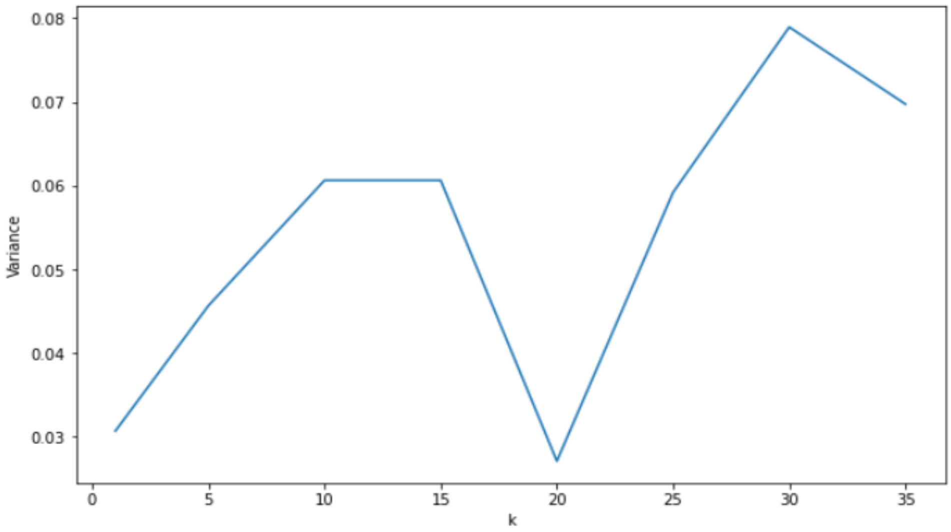


For 5 fold cross validation, the Variance vs k plot is as below:



Plotting Variance vs K curve

```
plt.figure(figsize=(10,6))
plt.plot(k_list, var_scores)
plt.xlabel('k')
plt.ylabel('Variance')
plt.show()
```

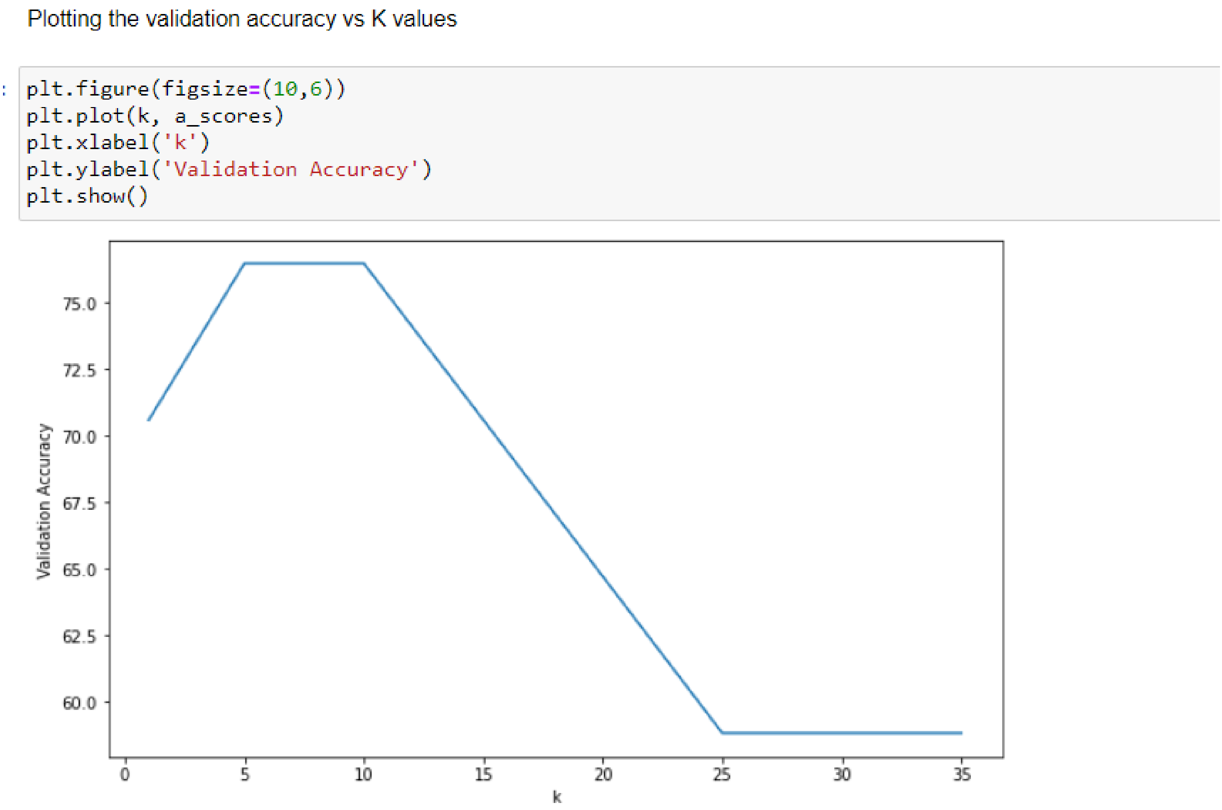


From the accuracy vs k plot, it can be observed that **K=5** has the highest classification accuracy at **92.8% .**

From variance vs k plot, it can be observed that the chosen k value i.e. **K= 5** exhibits low variance indicating that the accuracy values obtained in different folds of cross validation do not have much deviation from the mean accuracy values across the folds. As k increases, we can observe that the deviation is significantly high except at k=20 where it’s the minimum.

Therefore, **K=5** can be chosen as the best k value for the Iris dataset in terms of accuracy and variance.

[CM5] **For Heart Disease, Validation accuracy vs K** curve is plotted



Plotting the validation accuracy vs K values

```
plt.figure(figsize=(10,6))
plt.plot(k, a_scores)
plt.xlabel('k')
plt.ylabel('Validation Accuracy')
plt.show()
```

From the above plot it can be observed that K= [5, 10] have the maximum accuracy of 76.47 %.

K= 15 also has a high accuracy of 70.58%.

In KNN, Lower k values make the model susceptible to noise and higher k values increase computation costs and in turn affect the performance of the model when working on unobserved data. Also, Lower k value causes higher variance and overfitting and Higher k causes higher bias and under fitting. So we should always try to balance the bias and variance while choosing the right k value.

Hence, as a balance, K = 15 is chosen as the best K value.

Do you find any advantage to one form of validation over the other?

- In k fold cross validation, we can verify how accurate our model is on multiple folds and different subsets of data. It makes sure that the population of our data is correctly represented. Therefore, we can ensure that it **generalizes** well to the unknown data in the future and improves the accuracy of the model.

- It can be very helpful in cases where the dataset is small (Ex. Iris) and splitting it into train-validate-set is not feasible.

  For the above reasons, k fold cross validation can be a better validation technique than just splitting the training set into validation set by 90%-10% split which may lead to unbalanced partitioning and affect the performance of the model.