**Preprocessing:**

The dataset contains numeric data for all the columns. As mentioned in the description the target class or prediction variable is ‘class’. So as per my observation the processing needed for the numeric data is standardization. This is done by using the sklearn library which has inbuilt preprocessing. The scale part considers the Nan data as missing which will not be exclude from any calculation. It is also best to apply on the train data rather than on the entire dataset. Generally, we calculate the mean and variance of the dataset set value around them and ignore shape distribution which will not give appropriate solution. So, we must standardize the data that will lead to uniform distribution.

**Classifier:**

The parameter of the Knn is the metrics that are used to determine the distance. The distance algorithm used are Minkowski and Eucledian. We can also decide the number of the neighbors that are to be considered. The neighbor is the most deciding factor. The weight function is used for prediction. There is two parameters uniform and distance. All points are equally weighted, and in distance inverse the distance closer the neighbor the more influence.

The validation is part of the process of determining bets of K values. The significance is it can calculate based on unknow data and predict he performance of the data. This method is useful because is gives less biased data. We take the mean of the scores that are obtained by the cross validation and then obtain maximum value of mean to find the optimal value of K.

We have also implemented elbow method for obtaining the optimal K. In this we can plot the K-value against error rate graph and determine the place where error rate is not increasing and get stable error rate.

Chart, scatter chart

Description automatically generated

**Explain what your criteria was for selecting the three attributes:**

The first step is to select the appropriate attributes because working on the entire dataset is difficult and time consuming. We have used correlation matrix to drop the correlated columns. The first step we can watch the matrix and consider one attribute from the most correlated attributes. As per the requirement we must find 3 best attributes, so we must remove 5 attributes out of 8. The three attributes should be less correlated. So, we have considered threshold value as 0.17. The three attributes according to our calculation preg, plas, pres.

Chart, treemap chart

Description automatically generated

**Visualization of classifier in 2D projection:**

Shape, rectangle

Description automatically generated

Above figure shows data at value ‘0’ is considered as most accurate. As we know the dataset

comes with some inaccuracy, thus the values that lie on -1 or 1 are considered as impurities in

the dataset. Predictions made using KNN classifiers show that most of the predictions of this

models show accuracy of 75% when k value is considered 1.

Shape, rectangle

Description automatically generated with medium confidence

This model shows the accuracy of 77% when K value is 6.

Shape, rectangle

Description automatically generated

This model is showing the accuracy of 79% when K value is 79%.

According to observation the maximum accuracy is obtained when value for K is considered as 31.As per the visualization in calculating optimal K the error rate at 31 was less.

**Interpret and compare results:**

The confusion matrix is used to calculatet thre precision ,recall ,F1score and accuracy of the test data set.The importance is to calculate performance of the test data set using N X N matrix.It also summarizes the performance of the dataset.The value of K with minimum error rate is choosen to give maximum accuracy

**Chart

Description automatically generated with medium confidence**

**Summary:**

Calendar

Description automatically generated with medium confidence

In the above we can see all the summary element of the data set.The accuracy for test dataset when k = 31 is 79 percent. This accuracy is better than the most of the other K we have considered.

**Chart, treemap chart

Description automatically generated**

The accuracy for the test data set is 77 percent when K is 6

**Chart, treemap chart

Description automatically generated**

The accuracy for the test data set is 75 percent when K is 1

**References:**

* [**https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.scale.html**](https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.scale.html)
* [**https://www.youtube.com/watch?v=j-uHsi-KjOs**](https://www.youtube.com/watch?v=j-uHsi-KjOs)
* [**https://www.datacamp.com/community/tutorials/k-nearest-neighbor-classification-scikit-learn**](https://www.datacamp.com/community/tutorials/k-nearest-neighbor-classification-scikit-learn)
* [**https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.scale.html**](https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.scale.html)
* [**https://medium.com/analytics-vidhya/how-to-determine-the-optimal-k-for-k-means-708505d204eb**](https://medium.com/analytics-vidhya/how-to-determine-the-optimal-k-for-k-means-708505d204eb)
* [**https://machinelearningmastery.com/feature-selection-with-real-and-categorical-data/**](https://machinelearningmastery.com/feature-selection-with-real-and-categorical-data/)