**KNN CLASSIFICATION ALGORITHM - RADIUS BASED vs CLASSWISE WEIGHTED DISTANCE**

**In this section, by using the husband and wife individual salaries (labeled data), we try to predict the mental stress in the couple. We will develop 2 different KNN models from scratch for this task and try to improve our prediction. Please refer the code file for the program.**

**We know that K-Nearest Neighbor algorithm is one of the most important classification algorithms to begin with on any dataset with fairly a smaller number of attributes. Its classification is very effective as its decision boundary could be non-linear like the actual real-world data would be.**

**1.RADIUS BASED KNN:**

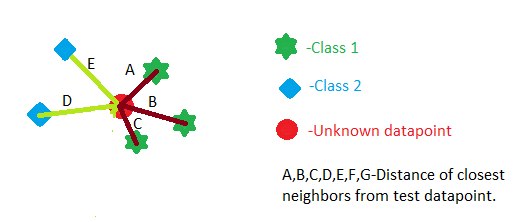
* Out of the many different variations of this algorithm, using radius as a parameter instead of N number of neighbors is one. This could be understood by the following diagram.

|  |
| --- |
| 1.KNN with different radius |
|  |
| Source: bigishere.wordpress.com |

* In the above diagram, the unknown datapoint will be classified as class B if radius of k=3 is assumed. This is because 2 datapoints in the k=3 radius are red and one is blue. So, instead of checking the N nearest neighbors, only points in the proximity are considered. The same datapoint will be classified as class A if radius is considered as k=6 as the number of blue datapoints in the outer circle are greater than red.
* We classify all datapoints in the training and validation set and compare with the original values (labels) to check different performance metrics like Accuracy, Recall, Sensitivity, PPV etc. We iterate this above process by increasing the radius in each iteration and note the performance metrics. Then we select the radius with the best combination of output parameters and classify the unknown data.
* From Fig.1,2,3,4,5, we could see that most of the performance metrics are high for the radius=0.8 after implementing this algorithm so we choose radius= 0.8 for classifying the unknown datapoints in the future.

**2.KNN CLASSWISE WIEGHTED DISTANCE.**

* Same data strategy followed in this as above, just instead of checking with all the neighbors in a certain radius, we are checking with a certain prefixed number of closest neighbors.
* Sometimes, only instead of checking the maximum number of neighbors of a class from N nearest neighbors, their distances from the unknown point is also considered.



* We could see this by an example as follows. Let’s say if we are considering 5 neighbors. Red circle as our datapoint has 5 neighbors A, B, C, D, E out of which A, B, C belong to class 1 and D, E belong to class 2. If we calculate the distance as (dist(A)+dist(B)+dist(C)) and (dist(D)+dist(E)), probability is that we will get the later value as smaller than the prior even though if A, B, C are individually small, and we might incorrectly classify this datapoint as class 2.
* Thus, we consider taking the mean value ((dist(D)+dist(E))/2) and ((dist(A)+dist(B)+dist(C))/3) while making our conclusion about the classification.
* Now while we have selected our augmentation, we select the optimal neighbors as 3 in our hypothesis because, the performance metrics in Fig. 6,7,8,9 and 10 show us that these are best for neighbors=3. Though neighbors =2 seems to be working perfectly, it tends to cause some overfitting issues and our prediction would not be generalized. Moreover,

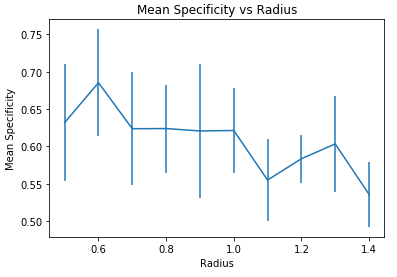
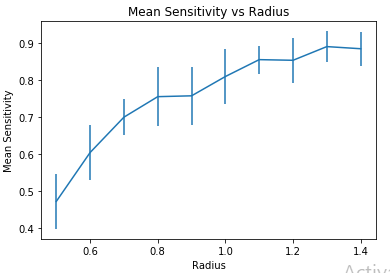
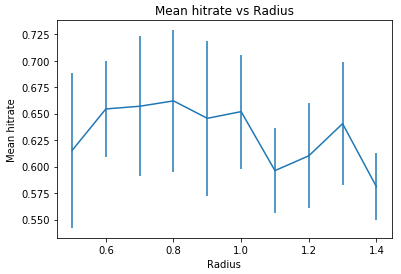
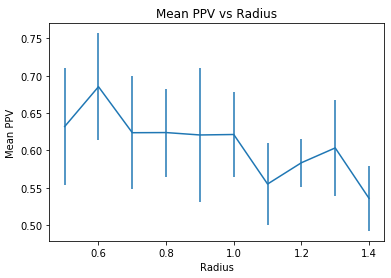
**REVIEW BETWEEN ABOVE 2 ALGORITHMS:**

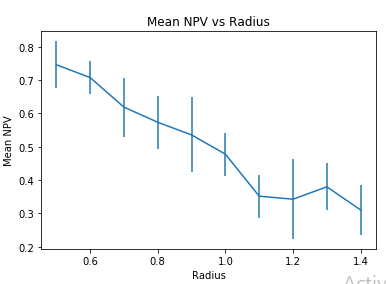
* First thing to note here from decision boundary is that, the areas of the meshgrid which are far away from datapoints (white space in fig.11), can be well classified by the KNN but the radius based KNN fails in such cases and depends hugely on approximation.
* Moreover, all the performance metrics for KNN are better than KNN radius based for optimal value of neighbors=3 and radius =0.8 respectively.
* The KNN radius algorithm will consider that the data is in circular pattern around the datapoint which is to strict a condition and usually not the case in real datasets. So, it’s safe to say that in this case, KNN with class wise weighted distance will perform better.

**MEAN PERFORMANCE PLOTS-RADIUS BASED**

**Fig.1 ,2,3,4,5-Represents performance for different radius.**

**Vertical line at each point represents the variance of metric over 10 repetition with the same radius and reshuffled training and testing.**

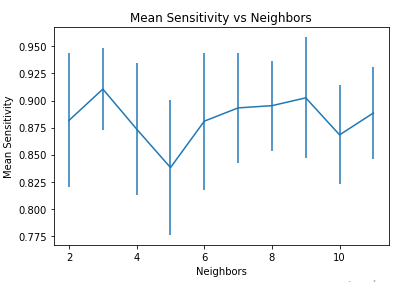
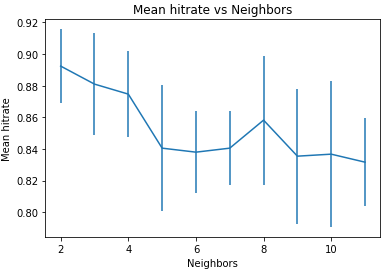


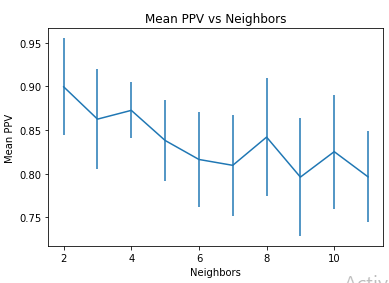
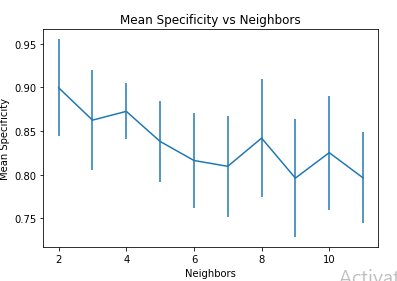


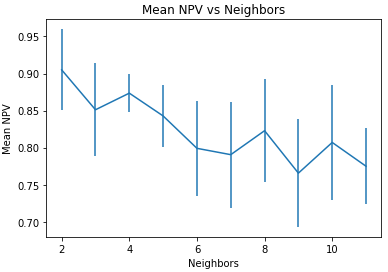
**MEAN PERFORMANCE PLOTS- CLASSWISE WEIGHTED DISTANCE**

**Fig.6,7,8,9,10- Represents performance for different number of neighbors.**

**Vertical line at each point represents the variance of metric over 10 repetition with the same number of neighbors and reshuffled training and testing.**

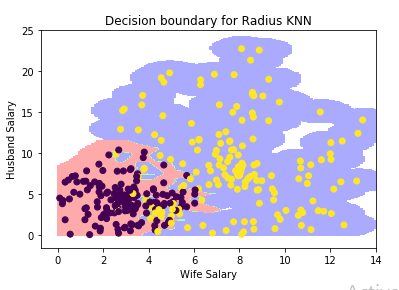






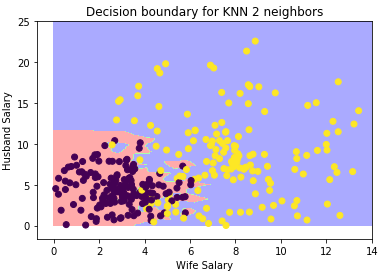
**Fig.11 DECISION BOUNDARY FOR KNN-RADIUS BASED (R=0.8)**

**Red=Non-stressed couples, Blue= Stressed couples, White =Cannot be classified by the classifier and too much approximation to give a default value.**



**Fig.12. DECISION BOUNDARY FOR KNN-CLASSWISE WEIGHTED DISTANCE**

**Red=Non-stressed couples, Blue= Stressed couples**



One thing is clear from the entire data that couples with both high husband salary and high wife salary will suffer from mental stress.