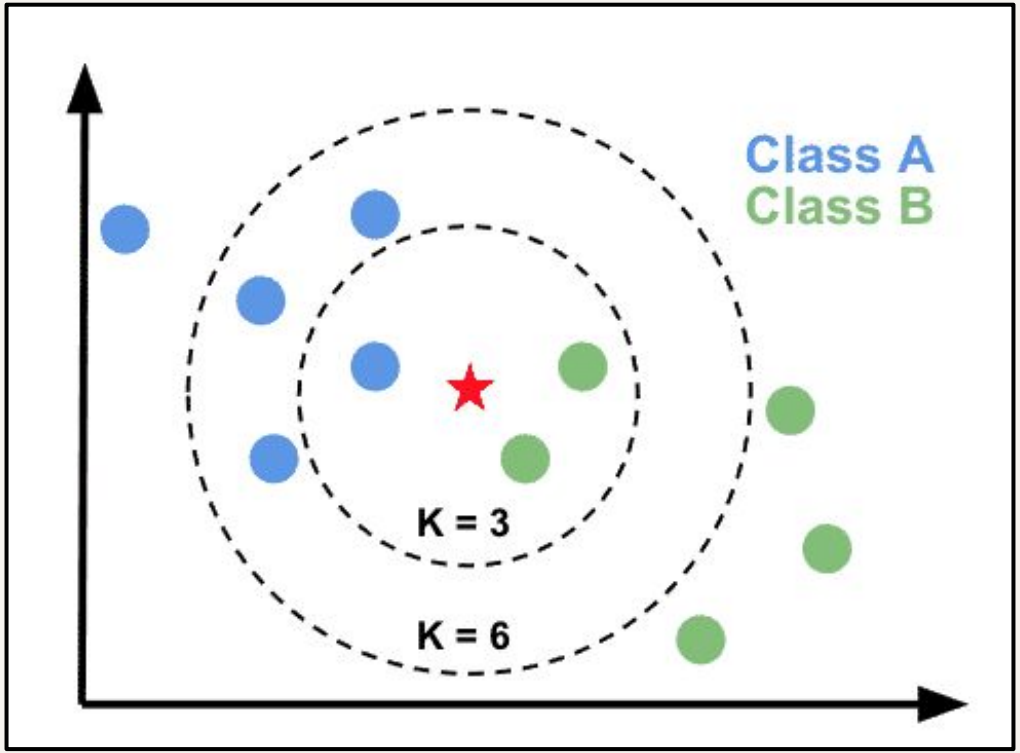
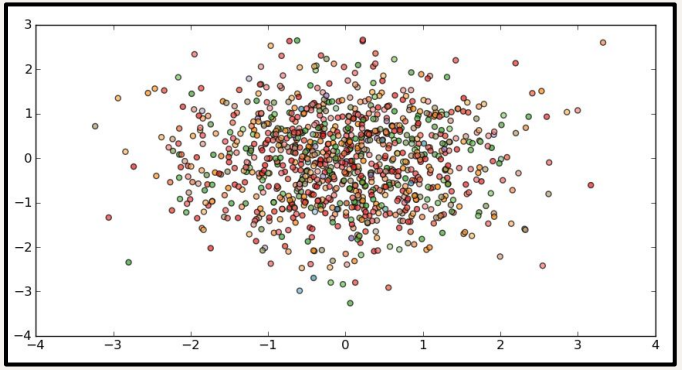
K-Closest Clusters (KCC): A Novel Cluster-Based Alternative to KNN

Julia Chen and Yuan Yin

KNN



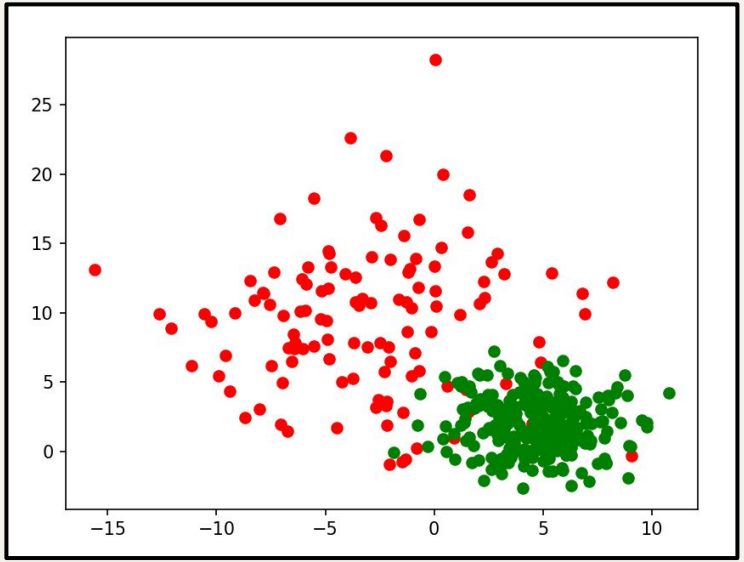
Problems

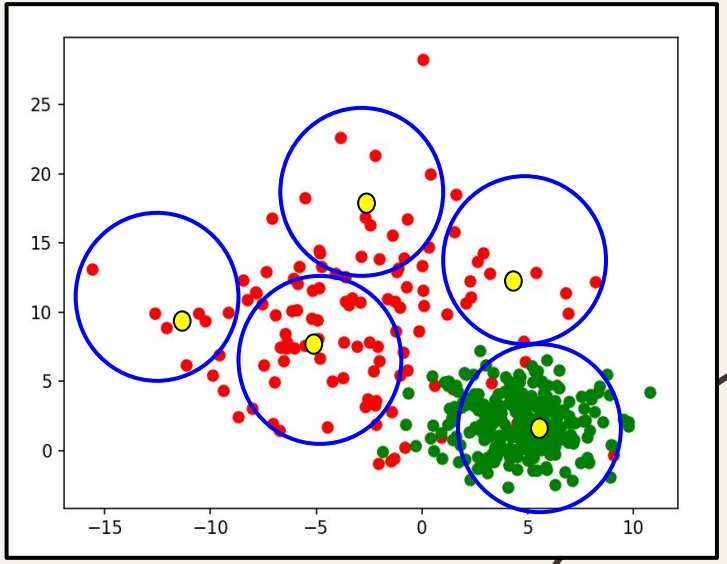


Limitations of KNN

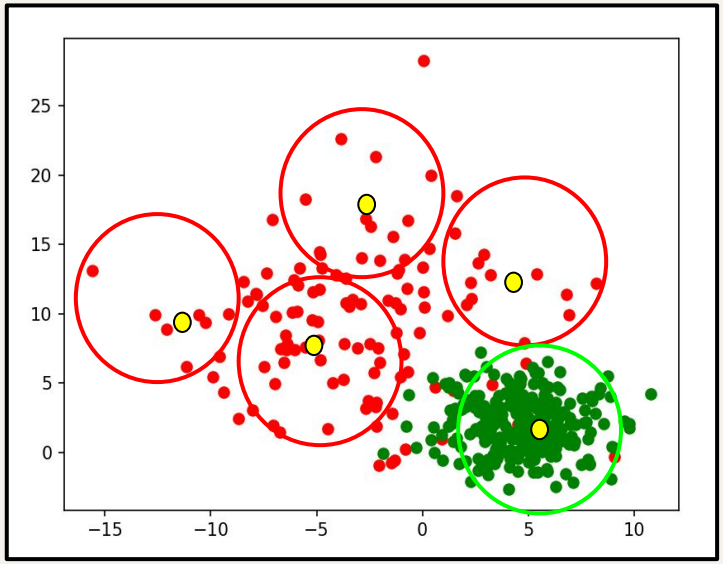
- Time Complexity
  - Calculates distance to every data point
  - Sorts all distances
- Memory
  - Lazy learner → does not store model

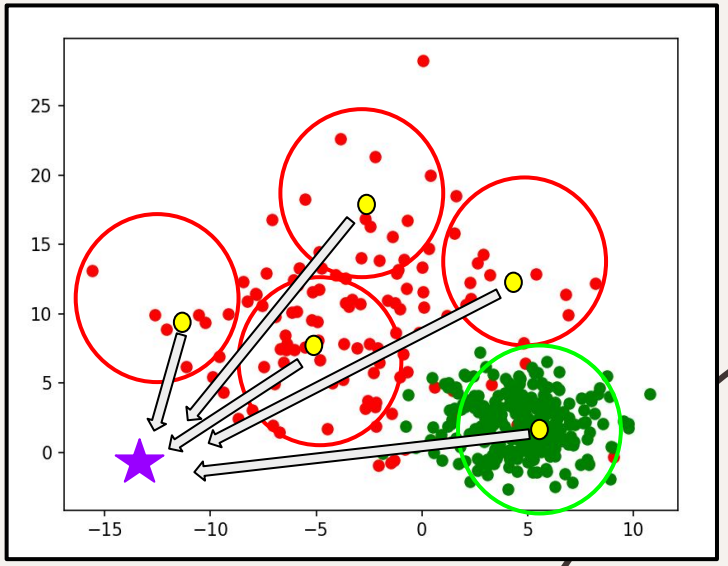
Model



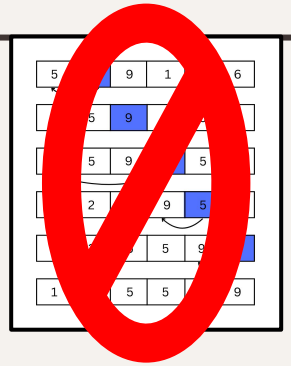


Model





Improvements



Time complexity issue

- KCC does not have to sort thousands of distances (not a lazy learner)

Memory issue

- KCC stores a model
  - Locations of centroids and their classifications

Determine K-value through validation

Research

“A New Method for Improving the Performance of K Nearest Neighbor using Clustering Technique”

- Similar technique
- Used arbitrary k-values
- Ensemble Learning

Table 2. Results of Nearest Cluster method in comparison with the traditional *K*NN method over SAHeart data set

				K=3	*K*=5	*K*=7
K-Nearest Neighbor				94.67±5.13	97.83±5.46	100.50±4.8
Nearest Cluster	Single Clustering	Single Classifier	M=1	94.83±3.74	98.67±6.65	100.17±4.7
		Ensemble	M=3	95.50±4.37	100.00±5.0	100.67±2.1
			M=5	99.33±2.25	**101.00±4.6**	**100.83±1.8**
	Clustering Ensemble	Single Classifier	M=1	96.83±4.22	94.50±6.92	96.83±5.60
		Ensemble	M=3	**99.67±4.13**	95.67±5.01	98.17±6.18
			M=5	99.17±2.93	98.83±3.76	100.17±7.6

Dataset

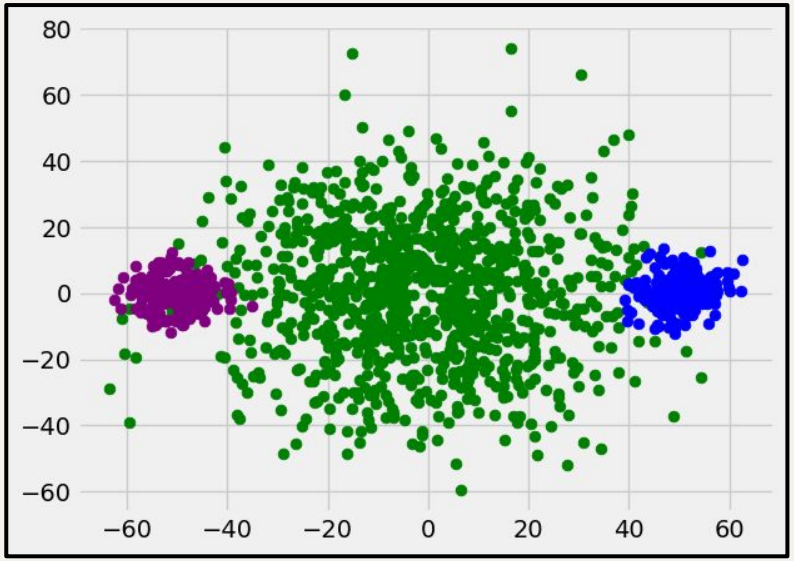
Generated dataset → Scikit-Learn (make_blobs)

KCC: (70-15-15 split)

- 980 Training instances
- 210 Validation instances
- 210 Testing instances

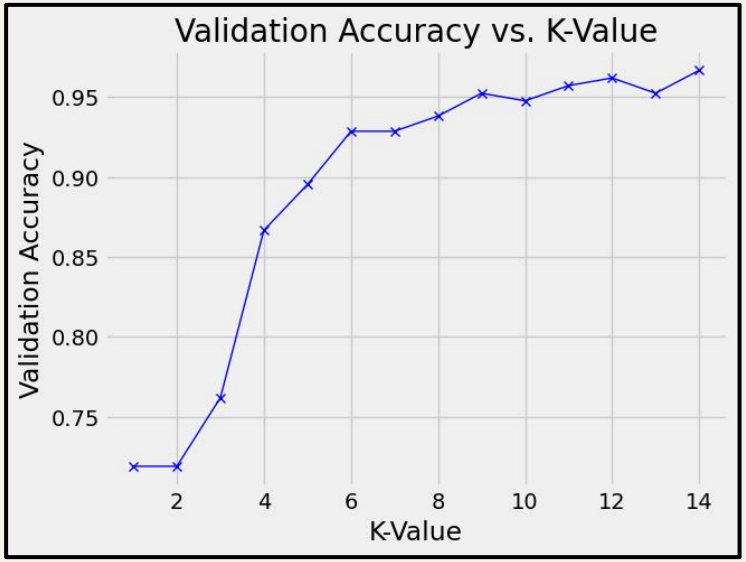
KNN:

- 980 Training instances
- 210 Testing instances



Methods: Pseudocode

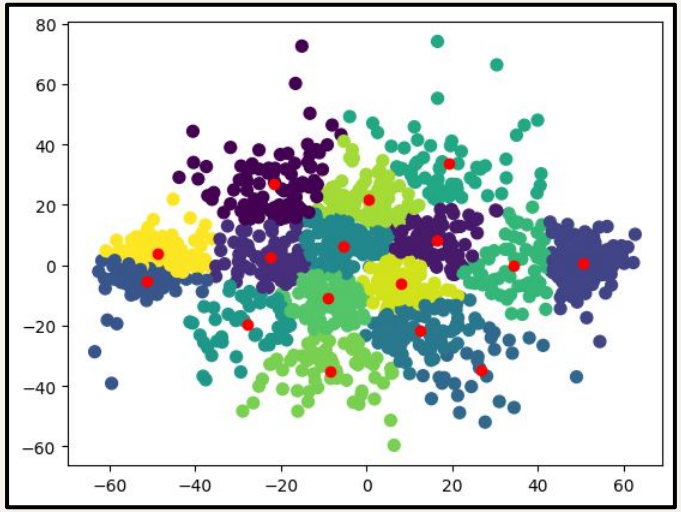
1. Use K-Means to make clusters
2. Assign a class to each of the clusters
3. Test model with k clusters on validation dataset
4. Plot accuracy vs. k value
5. Choose optimal k value



Methods: Pseudocode

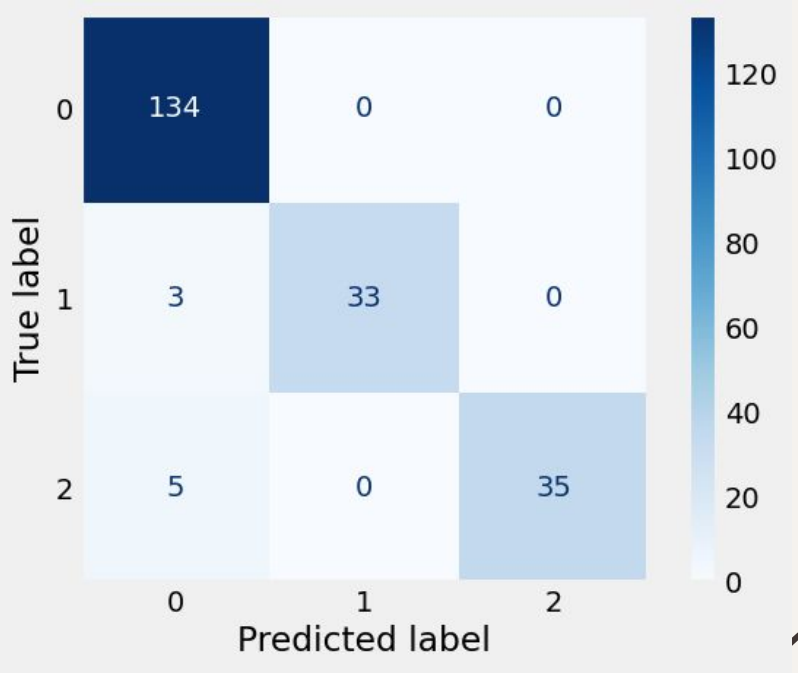
6. Use choose the model that has k clusters

7. Classify testing data set using KNN
   - Centroids as neighbors
   - k=1



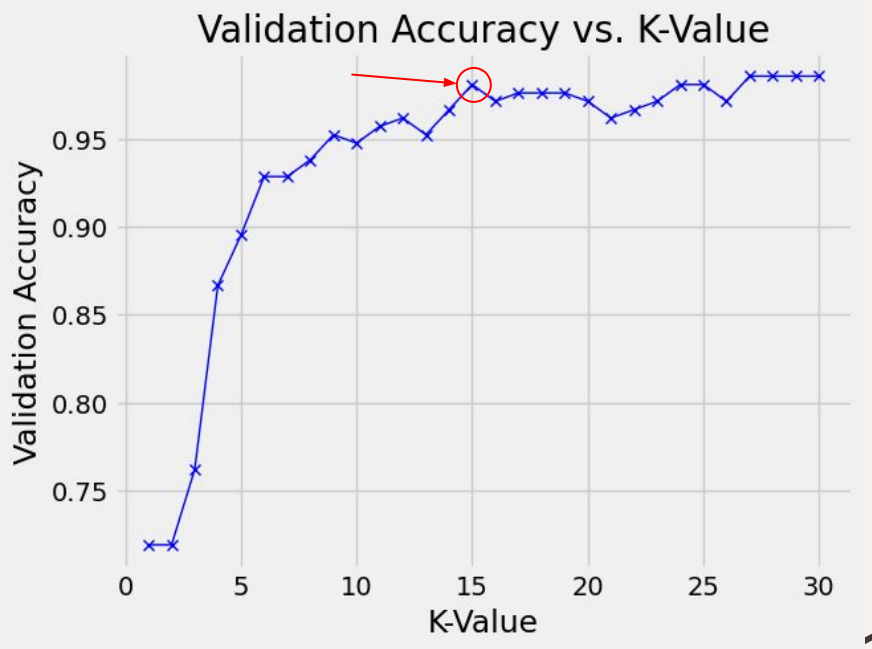
Results: Performance Metrics

- Accuracy: 96.2%
- Macro Average Precision: 0.931
- Macro Average Recall: 0.981

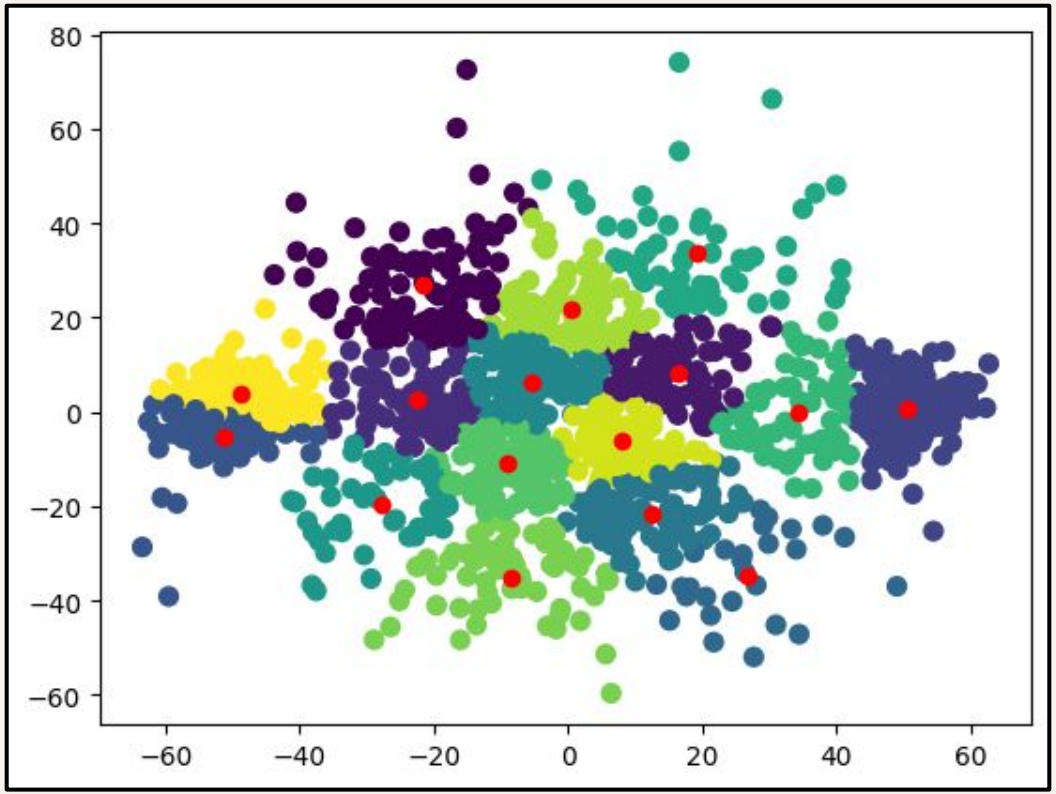


Results: Hyperparameters

- Chose a k-value that balances accuracy and efficiency
- K = 15



Results: Clusters



Discussion: Comparing KNN and KCC

Metric	KNN	KCC
Accuracy	95.7%	96.2%
Precision	0.922	0.931
Recall	0.979	0.981
Classification Time	0.229s	0.007s

Future Work

- Automate process for determining optimal k-value
- Research equations to better determine upper threshold of k-value calculations to avoid validation → model build time significantly decreases