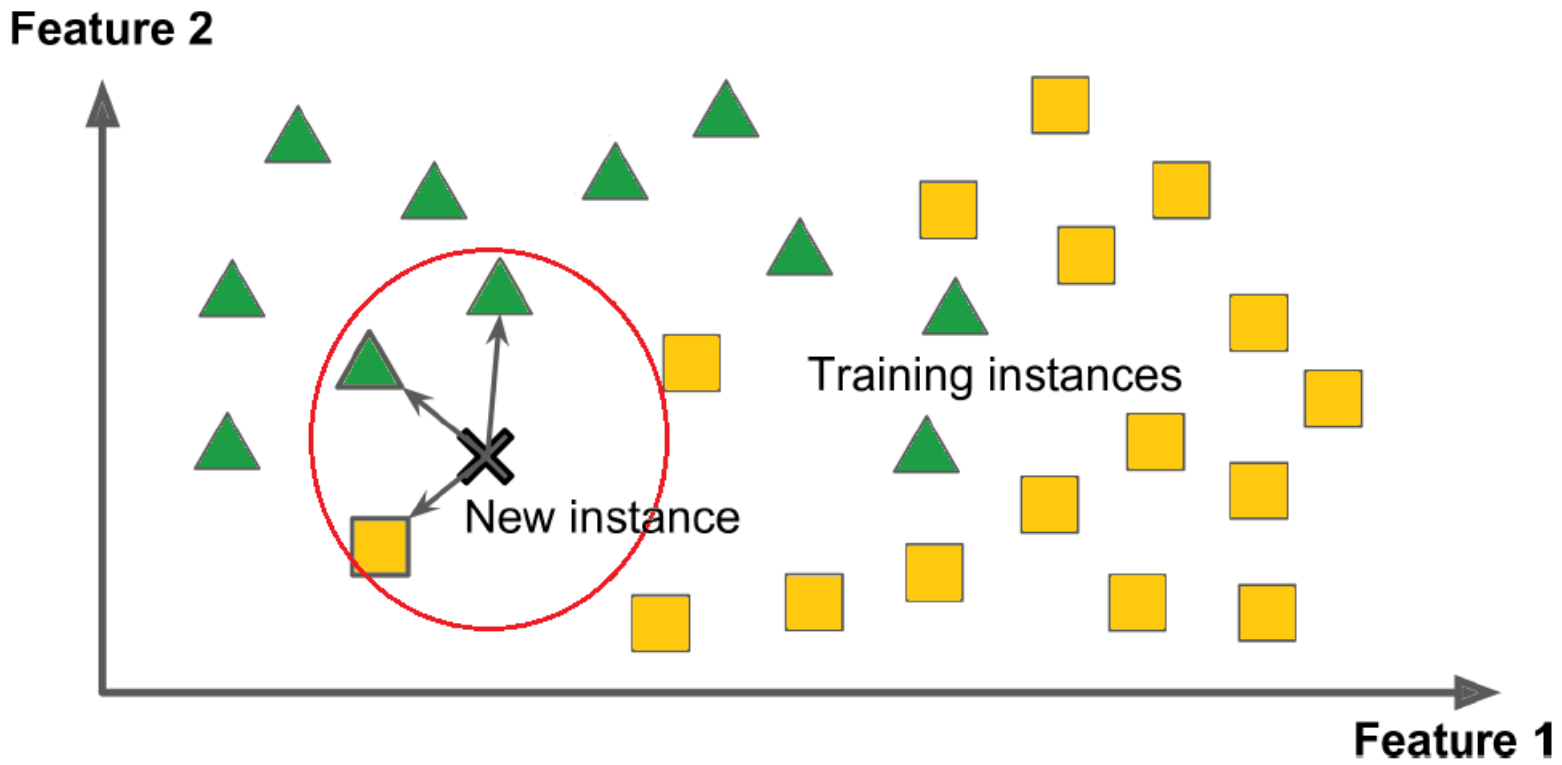
The impact of distance metrics on the performance of KNN algorithm

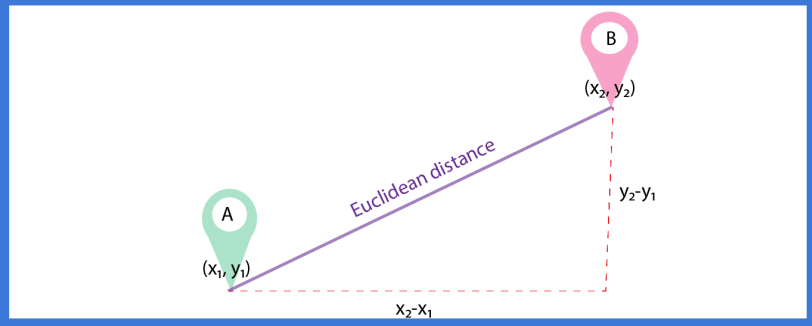
Angel Emanuel Rodríguez Román

The objective of this work is to evaluate the impact of different distance metrics (such as Euclidean, Manhattan, city block, etc) on the performance of the KNN algorithm with different values for K, using both balanced and imbalanced datasets.

KNN Algorithm



Measure vs Metric



What is a distance metric?

It’s a function that defines a way to measure the distance between two points in an space.

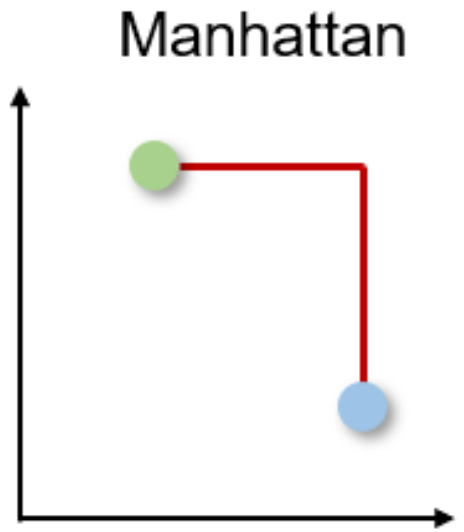
It **must** satisfy 3 conditions:

- Non-negativity: d(x, y) >= 0
- Identity: d(x, y) = 0 if and only if x == y
- Symmetry: d(x, y) = d(y, x)
- Triangle Inequality: d(x, y) + d(y, z) >= d(x, z)

Fiveable, Distance Metric, https://fiveable.me/key-terms/discrete-geometry/distance-metric

Manhattan distance
(Citi block or taxi cab)

$$d(x,y) = \sum_{i=1}^{n} |xi - yi|$$



E.G.

x = <-3,3> y = <1,-5>

$$d(x,y) = |-3 - 1| + |3 - (-5)|$$

$$d(x,y) = |-4| + |8|$$

$$d(x,y) = 12$$

Math.Net, Distance Metrics, https://numerics.mathdotnet.com/Distance#Distance-Metrics

Euclidean distance

$$d(x,y) = \left(\sum_{1}^{n} |xi - yi|^2\right)^{1/2}$$

E.G.

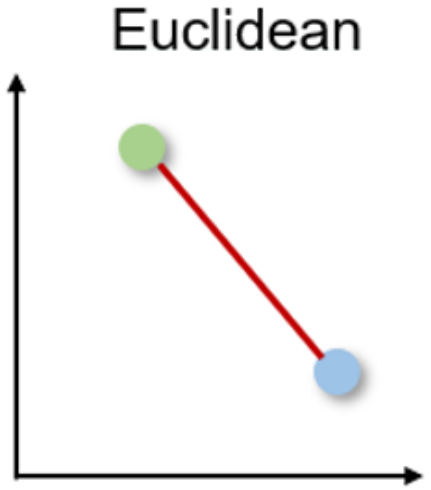
x = <-3,3> y = <1,-5>

$$d(x,y) = [(-3-1)^2 + (3-(-5))^2]^{1/2}$$

$$d(x,y) = [(-4)^2 + (8)^2]^{1/2}$$

$$d(x,y) = (80)^{1/2}$$

$$d(x,y) = 8.94$$



Math.Net, Distance Metrics, https://numerics.mathdotnet.com/Distance#Distance-Metrics

Minkowski

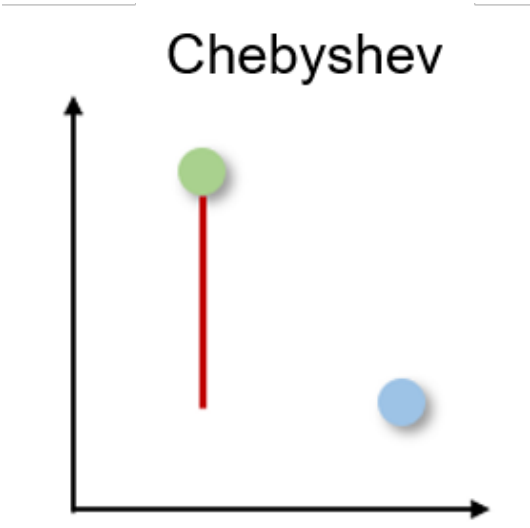
$$d(x,y) = \left(\sum_{1}^{n} |xi - yi|^{p}\right)^{1/p}$$

- Manhattan (City block p=1)
- Euclidean (p=2)
- Chebyshev (p→∞)

Math.Net, Distance Metrics, https://numerics.mathdotnet.com/Distance#Distance-Metrics

Chebychev distance
(Chessboard)

$$d(x,y) = \sum_{i=1}^{n} \max\ |xi\ - yi|$$



E.G.

x = <-3,3>

y = <1,-5>

$$d(x,y) = \max\ (|-4|, |8|)$$

$$d(x,y) = \max\ (4\,, 8)$$

$$d(x,y) = 8$$

Math.Net, Distance Metrics, https://numerics.mathdotnet.com/Distance#Distance-Metrics

Canberra distance

$$d(x,y) = \sum_{i=1}^{n} \frac{|xi - yi|}{|xi| + |yi|}$$

The Canberra distance is a weighted version of the Manhattan distance, introduced and refined 1967 by Lance, Williams and Adkins. It is often used for data scattered around an origin, as it is biased for measures around the origin and very sensitive for values close to zero.

Math.Net, Distance Metrics, https://numerics.mathdotnet.com/Distance#Distance-Metrics

Braycurtis distance (Sorensen)

$$d(x,y) = \frac{\sum|xi - yi|}{\sum|xi + yi|}$$

The Braycurtis distance views the space as grid, similar to the city block distance. If both points are in zero coordinates, the Braycurtis distance is undefined. It's commonly used in botany, ecology and in general, in environmental science fields.

Teknomo, Kardi (2015) Similarity Measurement. http:\people.revoledu.comkardi tutorialSimilarity

Cosine distance

$$1 - \frac{u \cdot v}{\|u\|_2 \|v\|_2}$$

The cosine distance contains the dot product scaled by the product of the Euclidean distances from the origin. It represents the angular distance of two vectors while ignoring their scale.

This metric is widely used in NLP. In applications such as comparing similarity between words.

Math.Net, Distance Metrics, https://numerics.mathdotnet.com/Distance#Distance-Metrics

Other distance metrics

- Hamming Distance (for **integers** and boolean)
- Dice (boolean)
- Jaccard (boolean)
- Rogerstanimoto (boolean)
- Russellrao (boolean)
- Sokalmichener (boolean)
- Jensenshannon(boolean)
- Mahalanobis $\sqrt{(u-v)V^{-1}(u-v)^T}$

SciPy, Documentation(Distance computations), https://docs.scipy.org/doc/scipy/reference/spatial.distance.html

Data

Dataset	Cardinality	Attributes	IR	About	Balanced
Nutt	28	1070	1	This dataset contains attributes related to brain tumors. The goal is to predict between classic glioblastomas and non classic.	Yes
Apendicitis	106	8	4.04	The data represents 7 medical measures taken over 106 patients on which the class label represents if the patient has appendicitis (class label 1) or not (class label 0).	No
Sonar	208	61	1.14	This dataset contains information about rocks and mines (Energy within a particular frequency band).	Yes
Heart	270	14	1.25	This dataset contains valuable information to detect the absence (1) or presence (2) of heart disease.	Yes
Haberman	306	4	2.77	This data set contains cases from a study that was conducted between 1958 and 1970 at the University of Chicago's Billings Hospital on the survival of patients who had undergone surgery for breast cancer. The task is to determine if the patient survived 5 years or longer (positive) or if the patient died within 5 year (negative).	No

Dataset	Cardinality	Attributes	IR	About	Balanced
Ionosphere	351	34	1.78	This dataset provides data from a system that consists of a phased array of 16 high-frequency antennas with a total transmitted power on the order of 6.4 kilowatts. The task is to determine if a given signal is Good (g) or Bad (b).	No
Monk-2	432	7	1.11	The MONK's problems are a collection of three binary artificial classification problems (MONK-1, MONK-2 and MONK-3) over a six-attribute discrete domain. Each problem involves learning a binary function defined over this domain, from a sample of training examples that belong to class 0 or class 1.	Yes
WDBC (Breast Cancer Wisconsin Diagnostic)	569	31	1.68	This database contains 30 features computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image. The task is to determine if a found tumor is benign or malignant (M = malignant, B = benign).	No
Electricity	2400	14	1.47	This data was collected from the Australian New South Wales Electricity Market. The objective is to predict the change of the price (UP or DOWN) .	Yes
Spam	4597	58	1.53	This database contains information about 4597 e-mail messages. The task is to determine whether a given email is spam (class 1) or not (class 2).	No

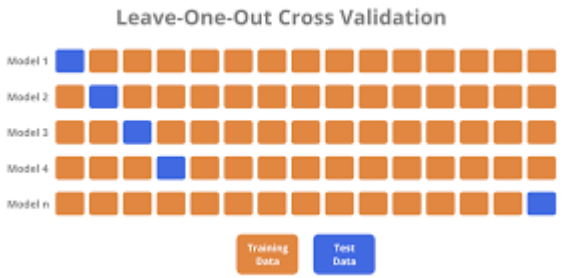
Experimental Framework

Distances : {“euclidean”, “manhattan”, “chebyshev”, “canberra”, “braycurtis”, “cosine”}

K : {1, 3, 5}

Algorithms: $\{Distances \times K\}$

Validation Method: Leave One Out Cross Validation (LOO)

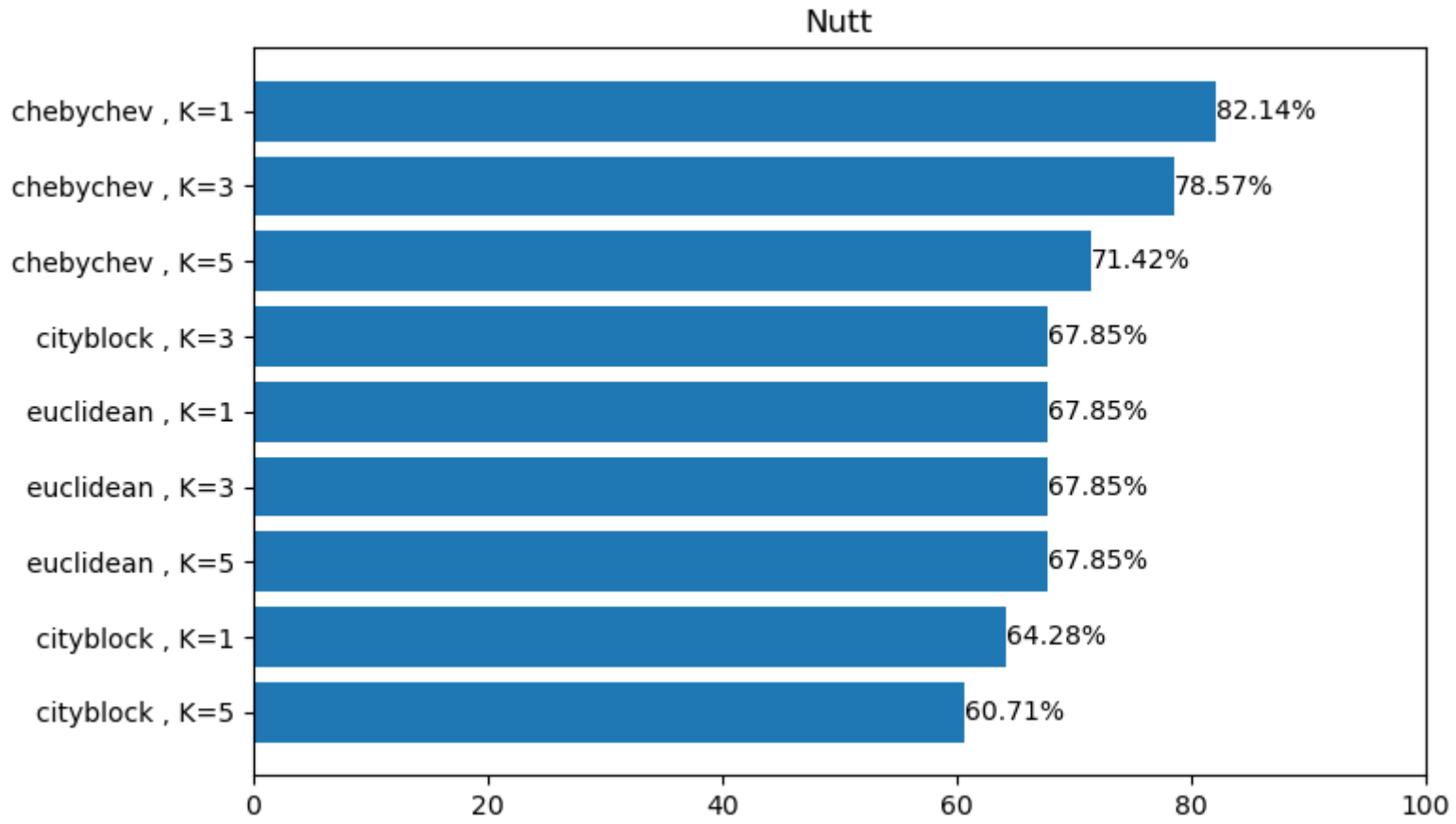


Measure: Balanced Accuracy (BA)

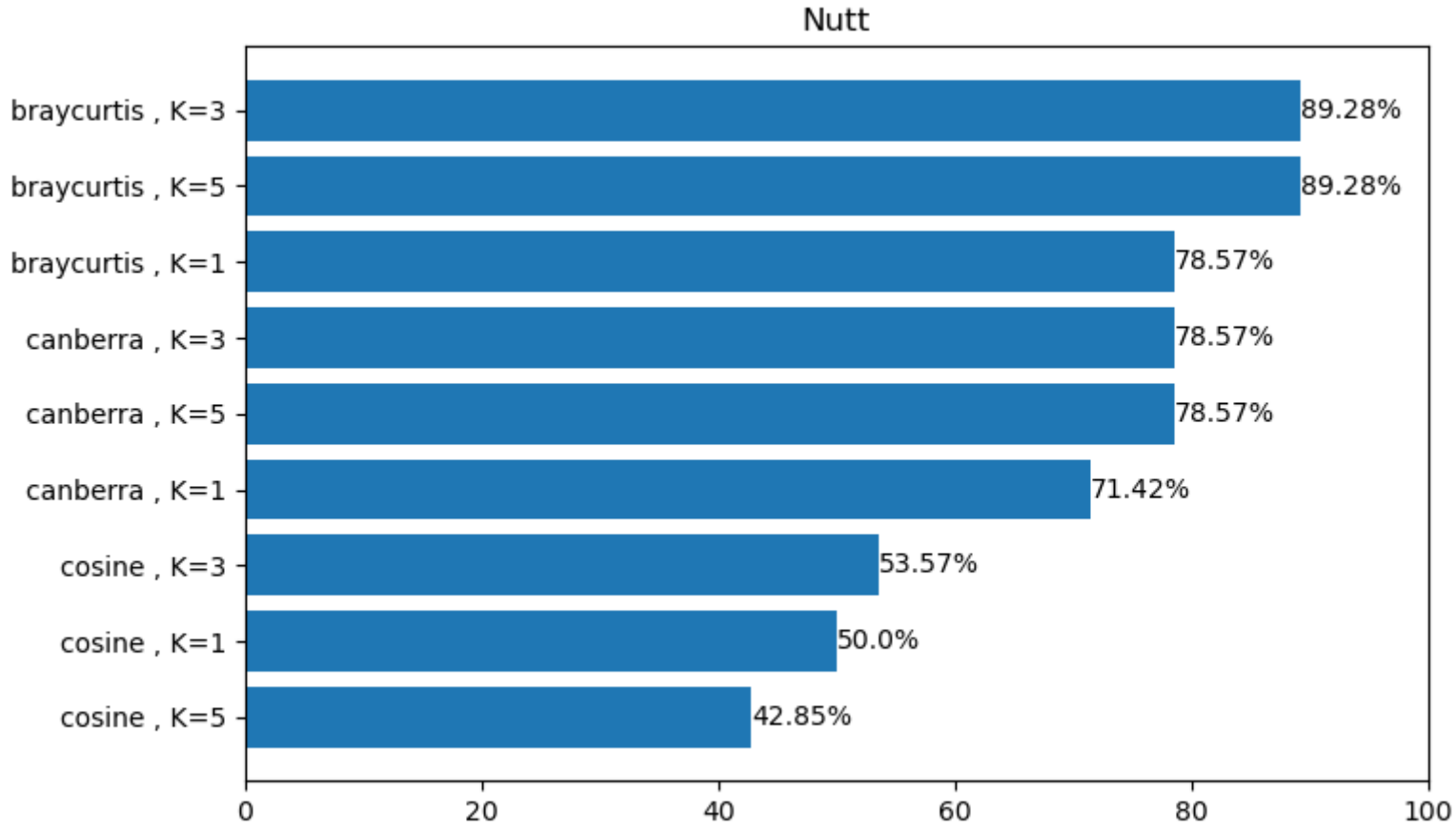
$$BA = \frac{Sensitivity + Specificity}{2}$$

Results

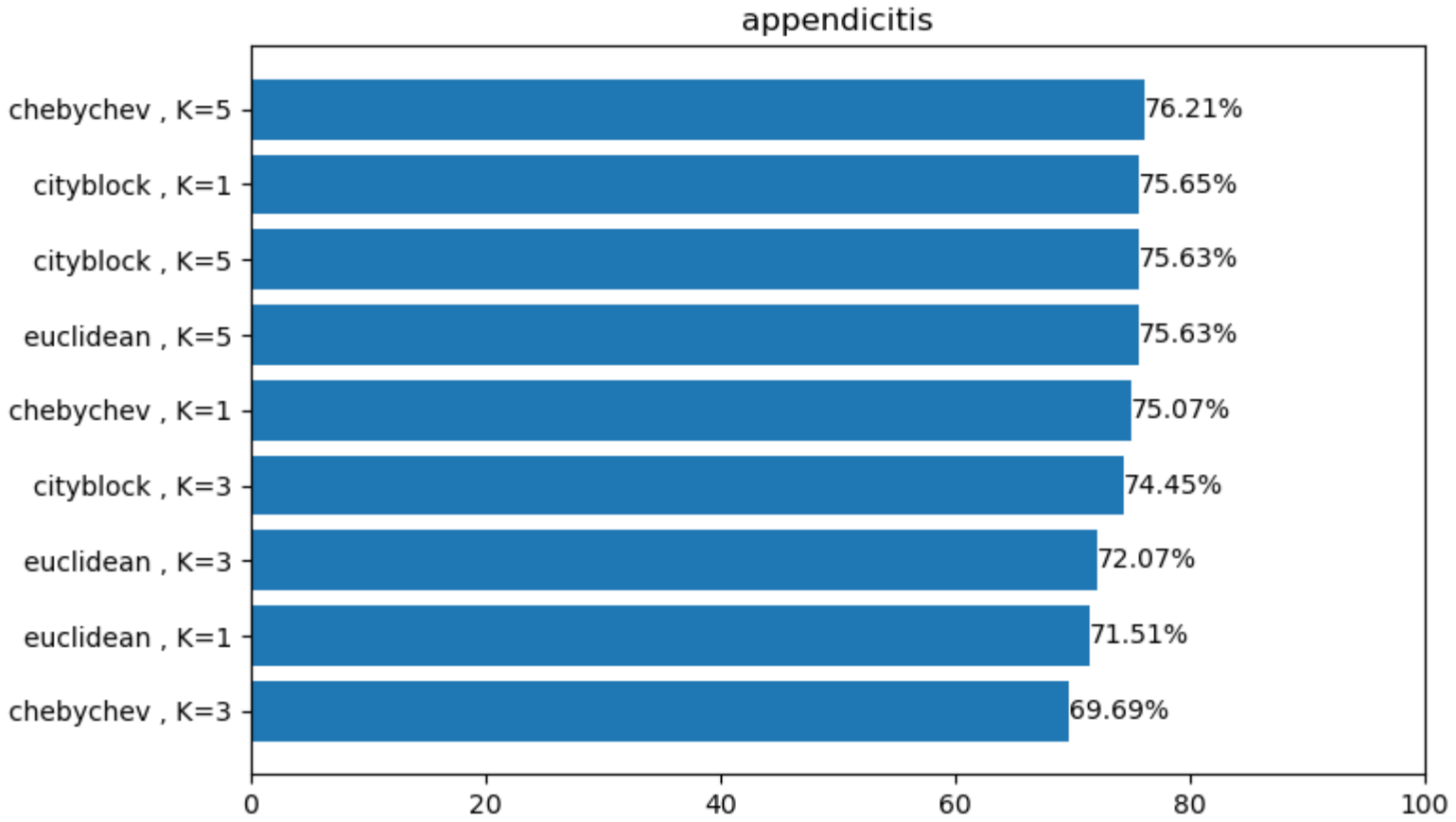
Nutt


Nutt
chebychev , K=1
chebychev , K=3
chebychev , K=5
cityblock , K=3
euclidean , K=1
euclidean , K=3
euclidean , K=5
cityblock , K=1
cityblock , K=5
82.14%
78.57%
71.42%
67.85%
67.85%
67.85%
67.85%
64.28%
60.71%
0
20
40
60
80
100

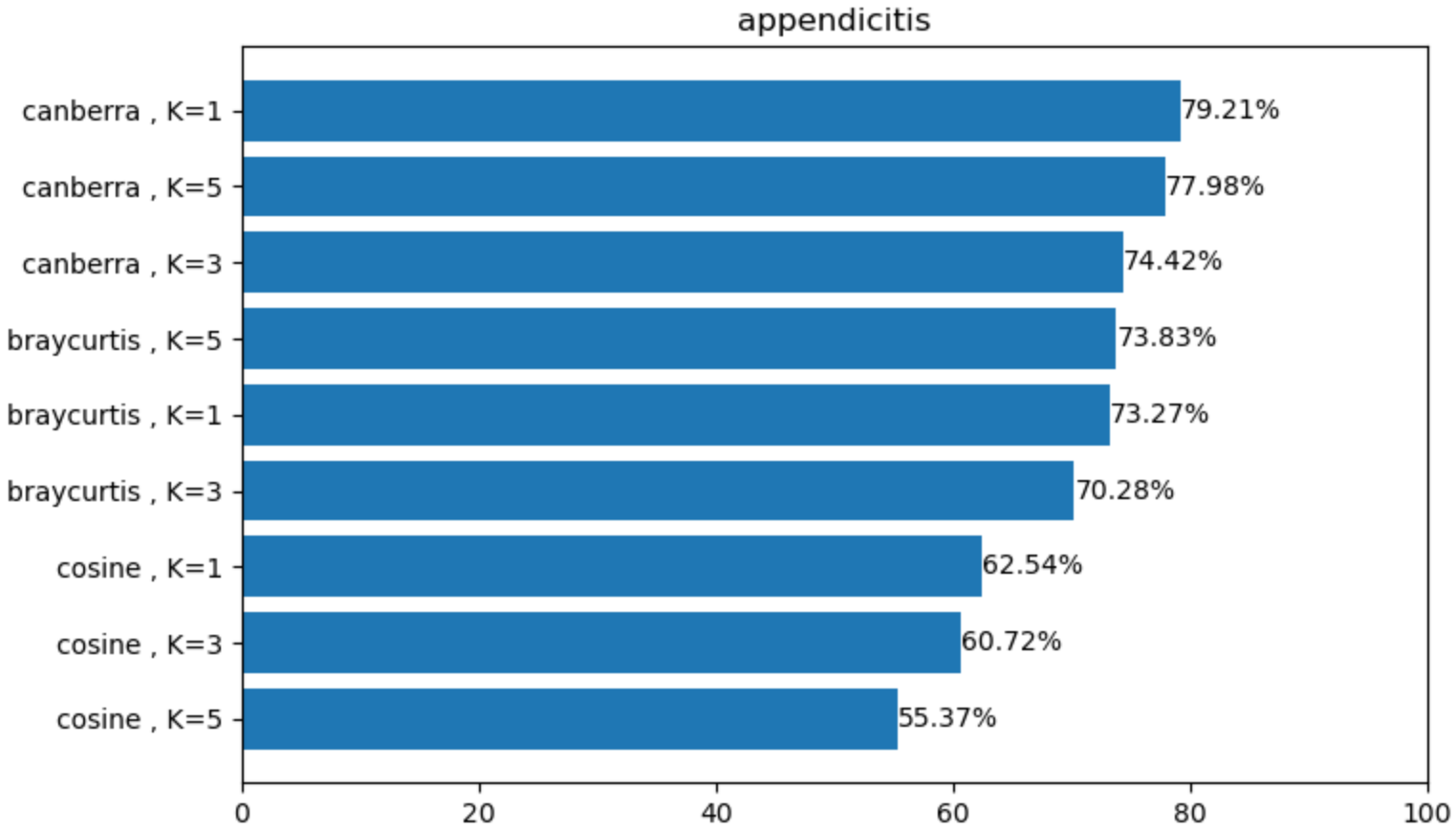
Nutt


Nutt
braycurtis , K=3
braycurtis , K=5
braycurtis , K=1
canberra , K=3
canberra , K=5
canberra , K=1
cosine , K=3
cosine , K=1
cosine , K=5
89.28%
89.28%
78.57%
78.57%
78.57%
71.42%
53.57%
50.0%
42.85%
0
20
40
60
80
100

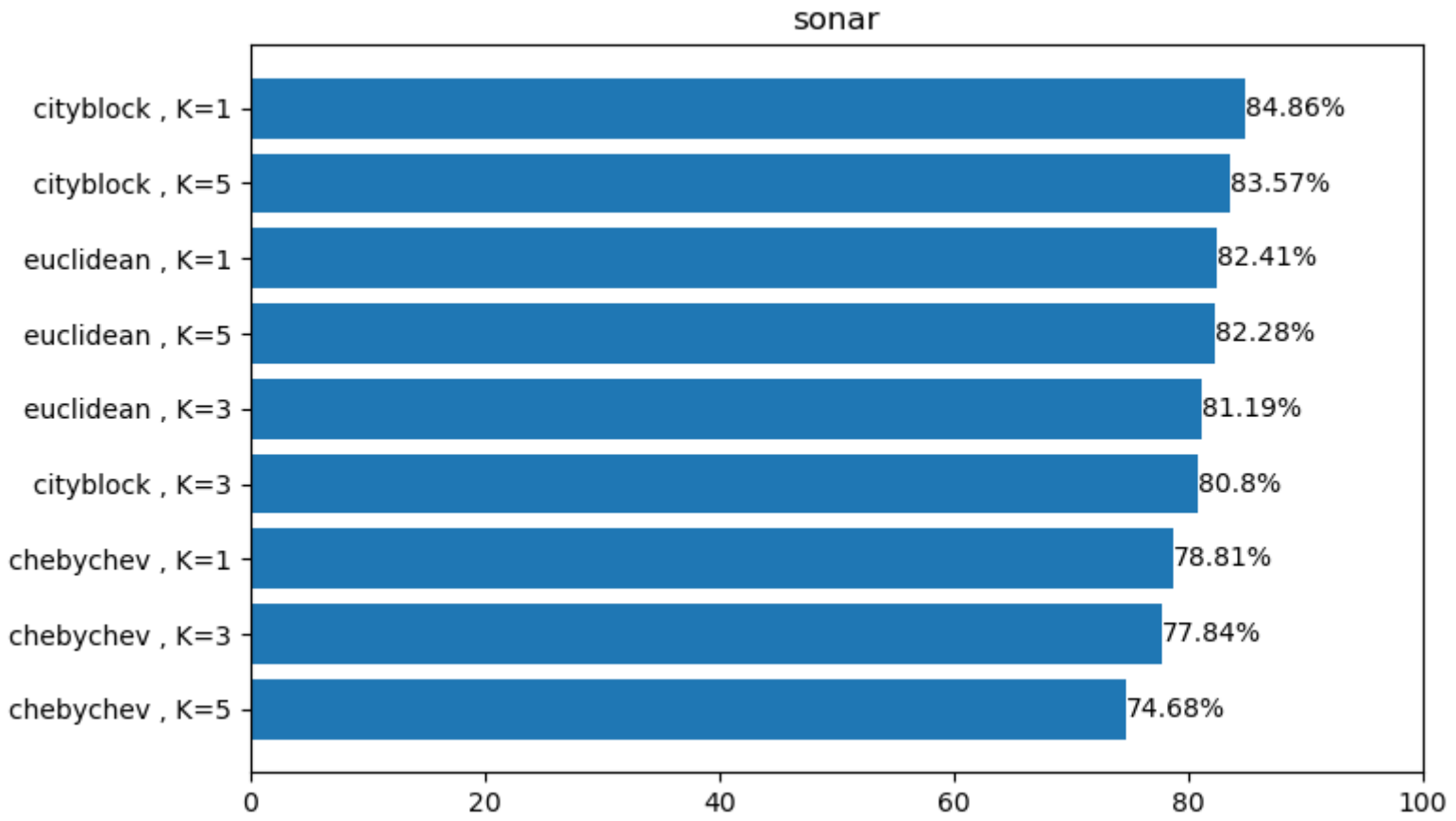
Apendicitis



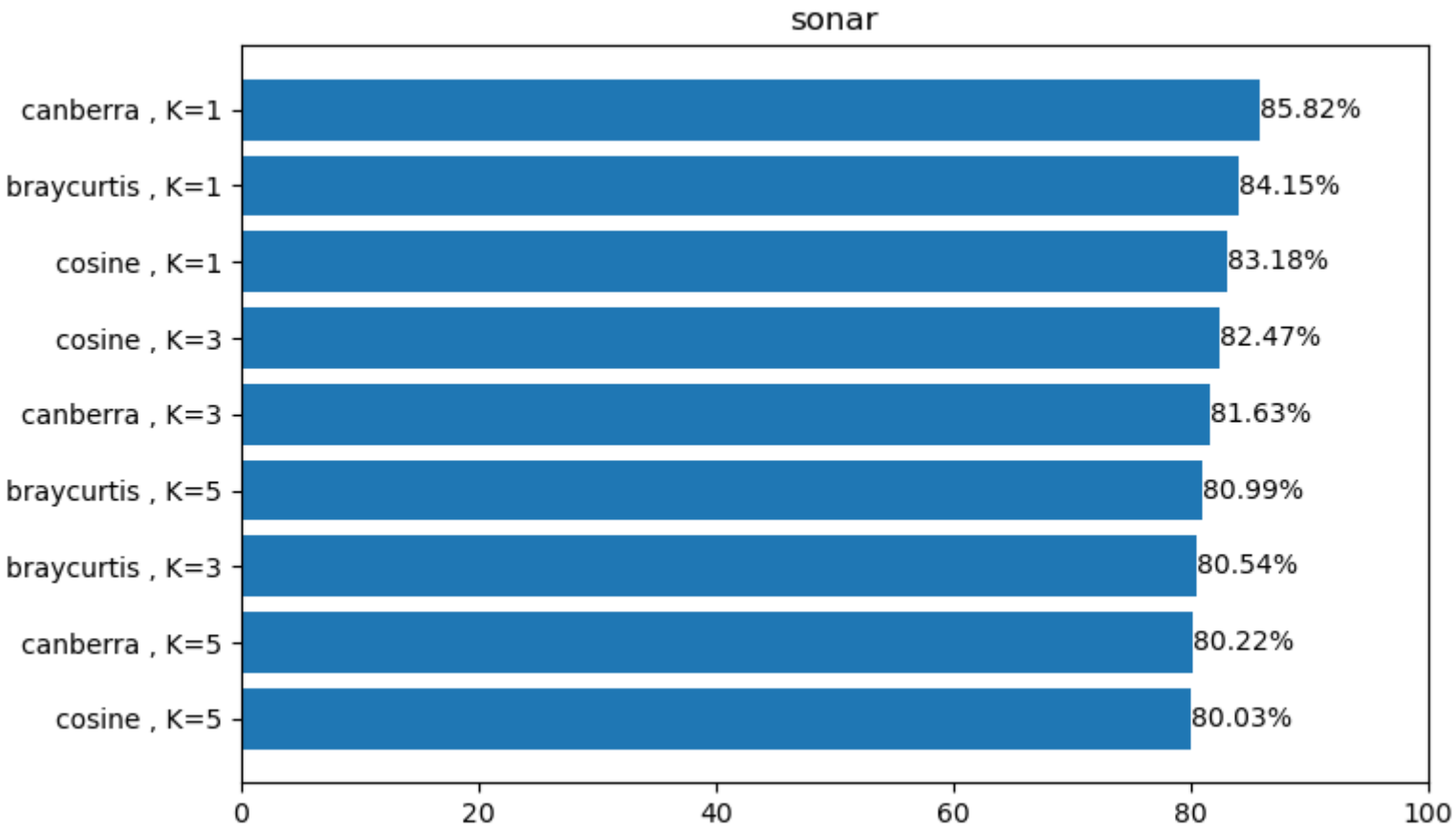
Apendicitis



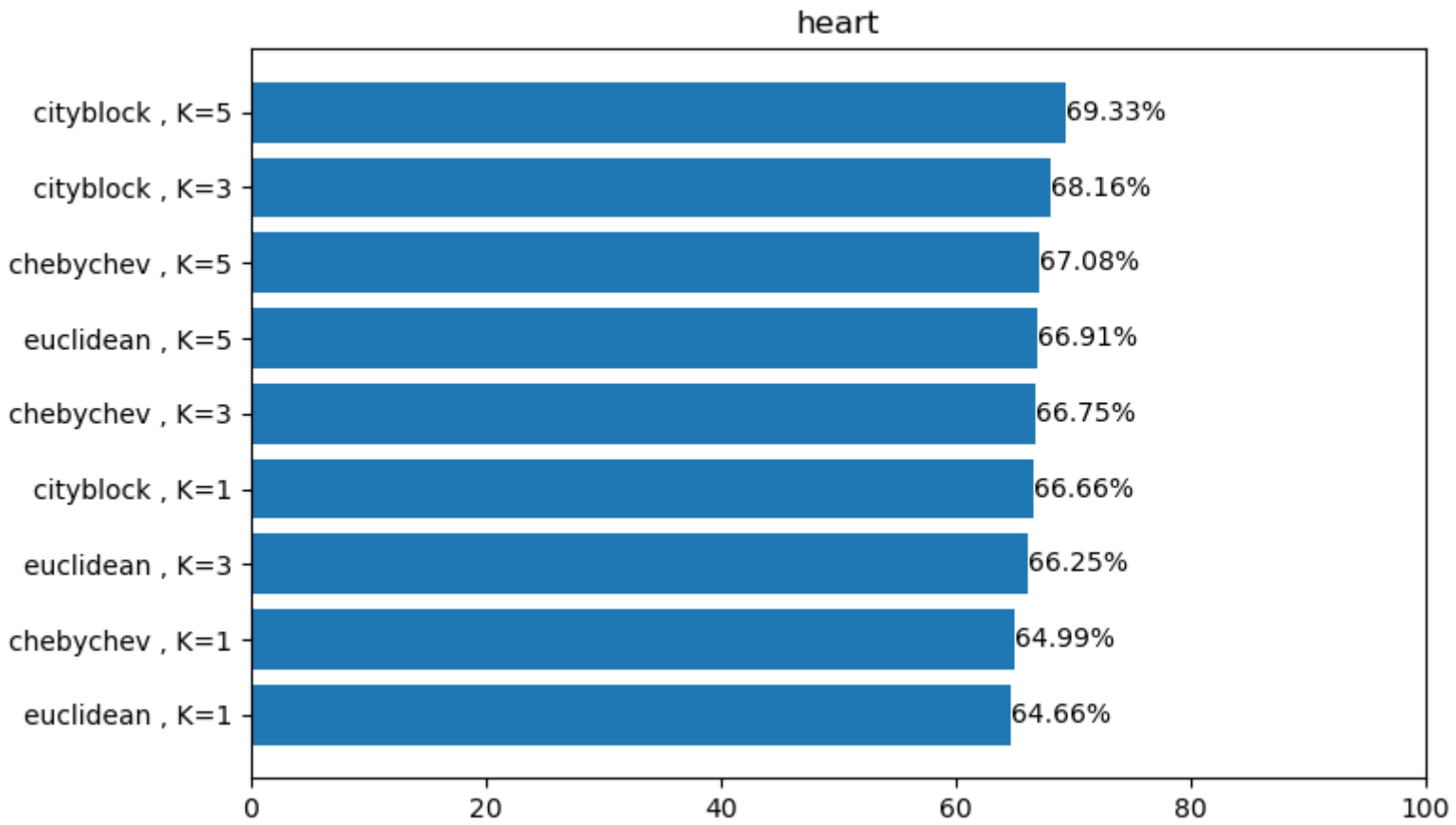
Sonar


sonar
cityblock , K=1
cityblock , K=5
euclidean , K=1
euclidean , K=5
euclidean , K=3
cityblock , K=3
chebychev , K=1
chebychev , K=3
chebychev , K=5
84.86%
83.57%
82.41%
82.28%
81.19%
80.8%
78.81%
77.84%
74.68%
0
20
40
60
80
100

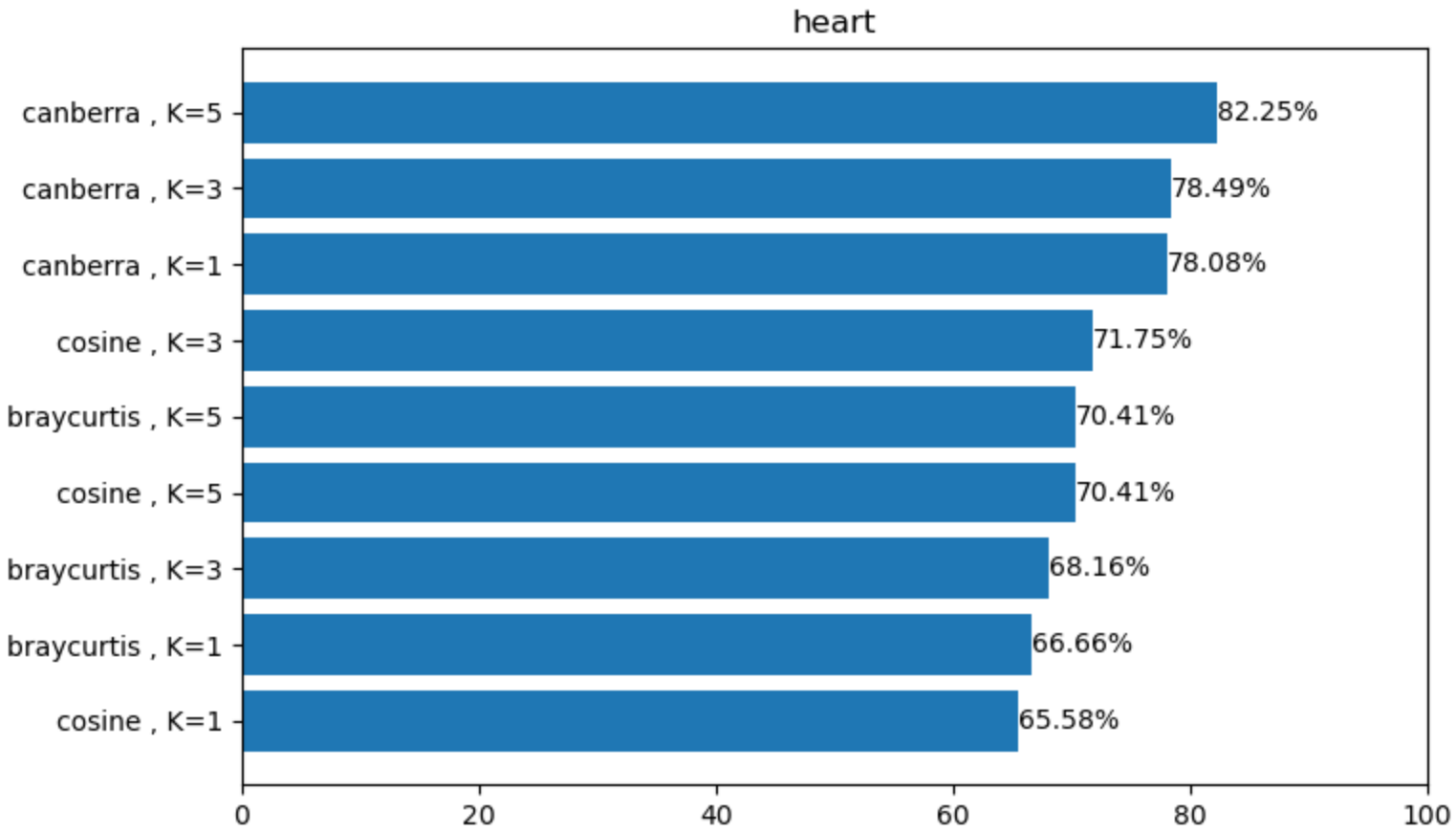
Sonar


sonar
canberra , K=1
braycurtis , K=1
cosine , K=1
cosine , K=3
canberra , K=3
braycurtis , K=5
braycurtis , K=3
canberra , K=5
cosine , K=5
85.82%
84.15%
83.18%
82.47%
81.63%
80.99%
80.54%
80.22%
80.03%
0
20
40
60
80
100

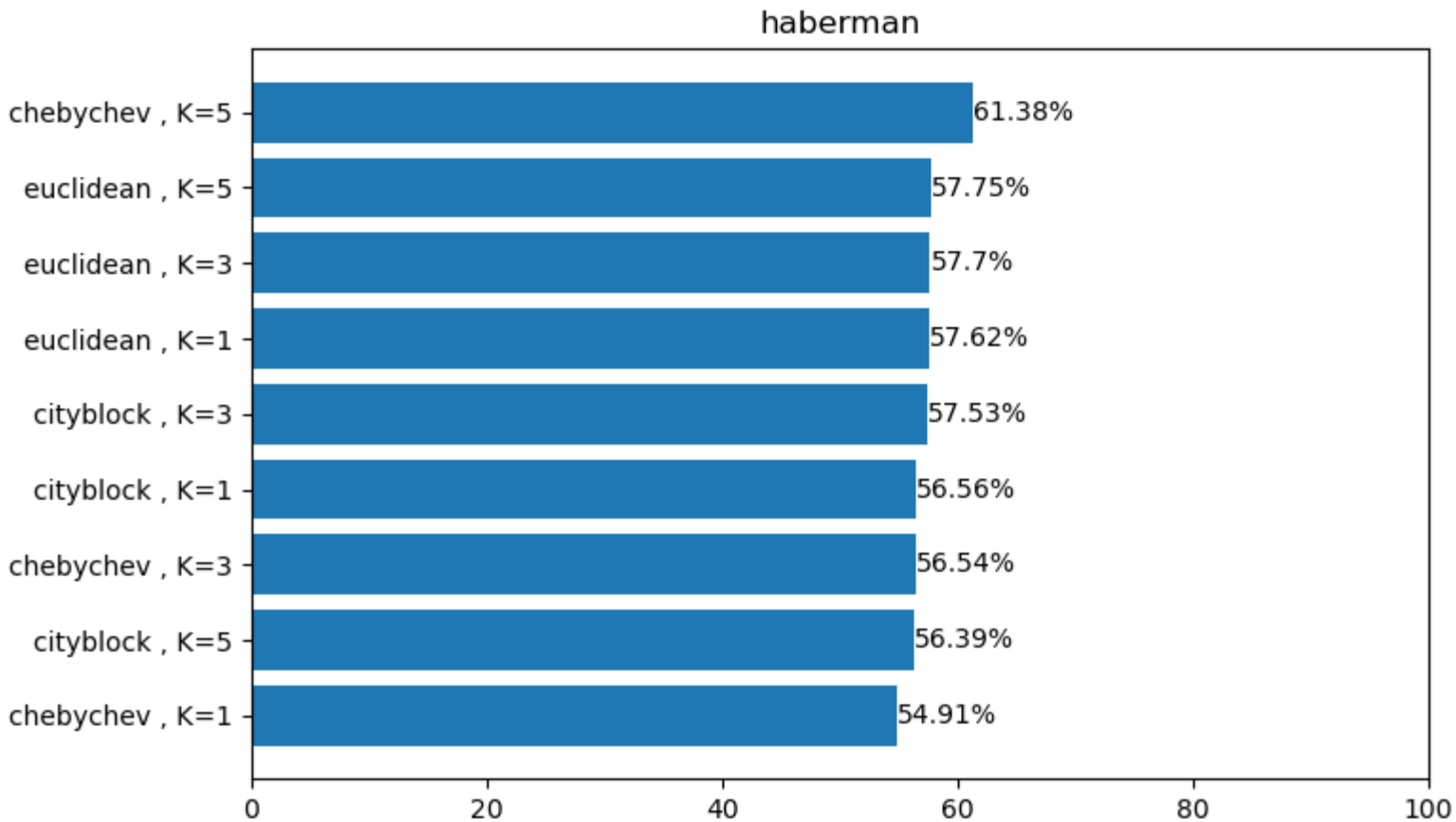
Heart


heart
cityblock , K=5
cityblock , K=3
chebychev , K=5
euclidean , K=5
chebychev , K=3
cityblock , K=1
euclidean , K=3
chebychev , K=1
euclidean , K=1
69.33%
68.16%
67.08%
66.91%
66.75%
66.66%
66.25%
64.99%
64.66%
0
20
40
60
80
100

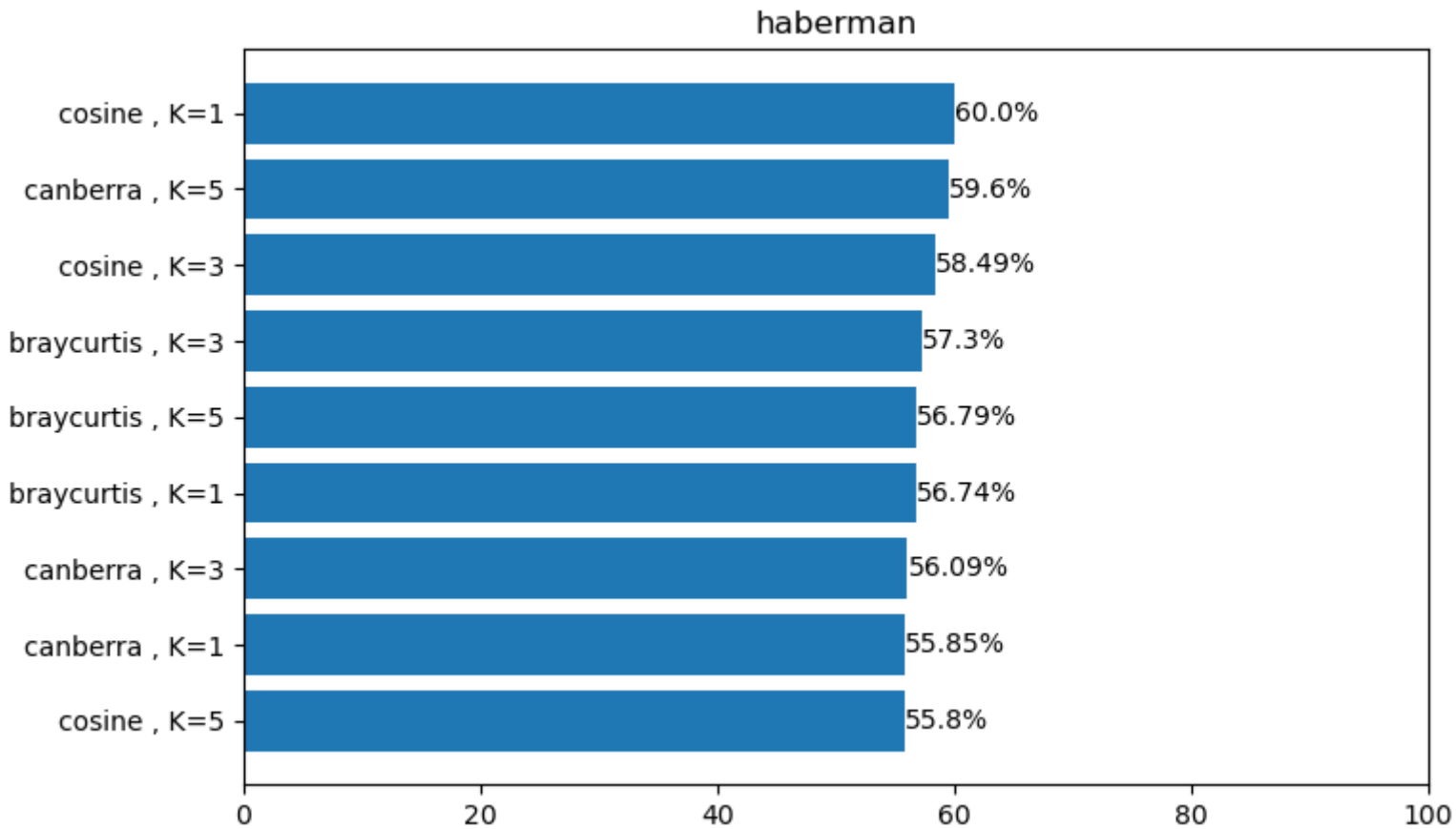
Heart


heart
canberra , K=5
canberra , K=3
canberra , K=1
cosine , K=3
braycurtis , K=5
cosine , K=5
braycurtis , K=3
braycurtis , K=1
cosine , K=1
82.25%
78.49%
78.08%
71.75%
70.41%
70.41%
68.16%
66.66%
65.58%
0
20
40
60
80
100

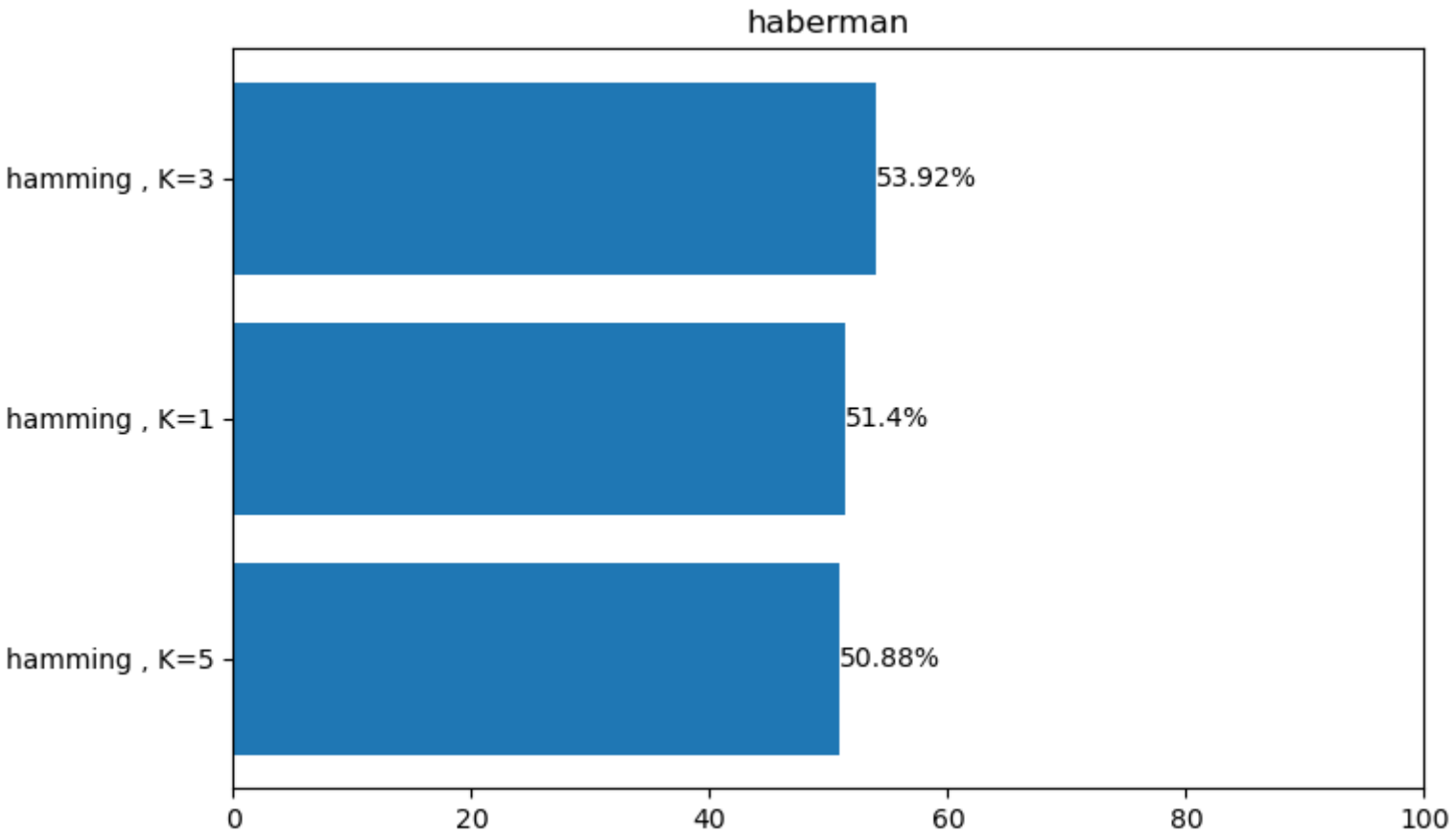
Haberman


haberman
chebychev , K=5
euclidean , K=5
euclidean , K=3
euclidean , K=1
cityblock , K=3
cityblock , K=1
chebychev , K=3
cityblock , K=5
chebychev , K=1
61.38%
57.75%
57.7%
57.62%
57.53%
56.56%
56.54%
56.39%
54.91%
0
20
40
60
80
100

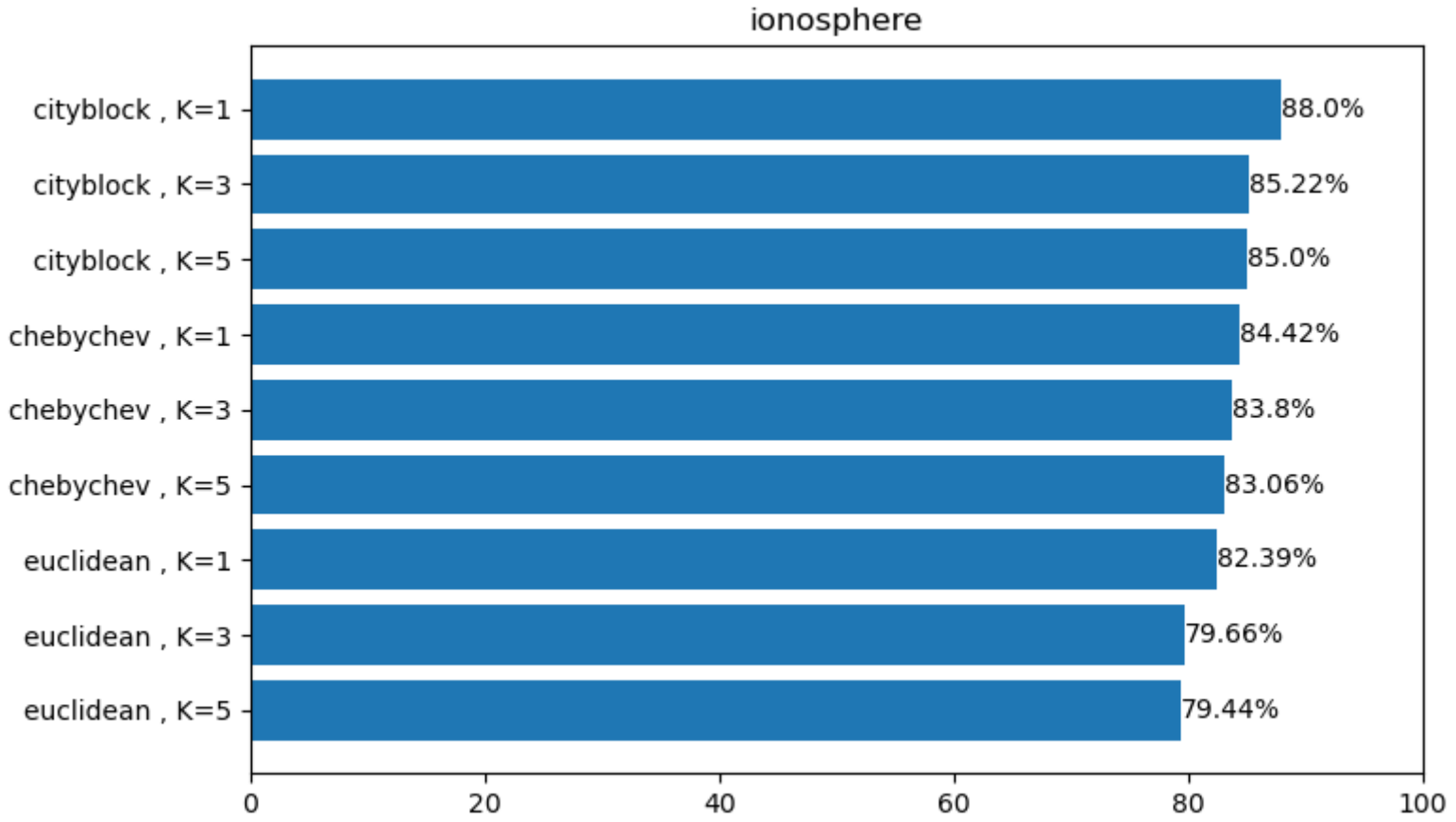
Haberman


haberman
cosine , K=1
canberra , K=5
cosine , K=3
braycurtis , K=3
braycurtis , K=5
braycurtis , K=1
canberra , K=3
canberra , K=1
cosine , K=5
60.0%
59.6%
58.49%
57.3%
56.79%
56.74%
56.09%
55.85%
55.8%
0
20
40
60
80
100

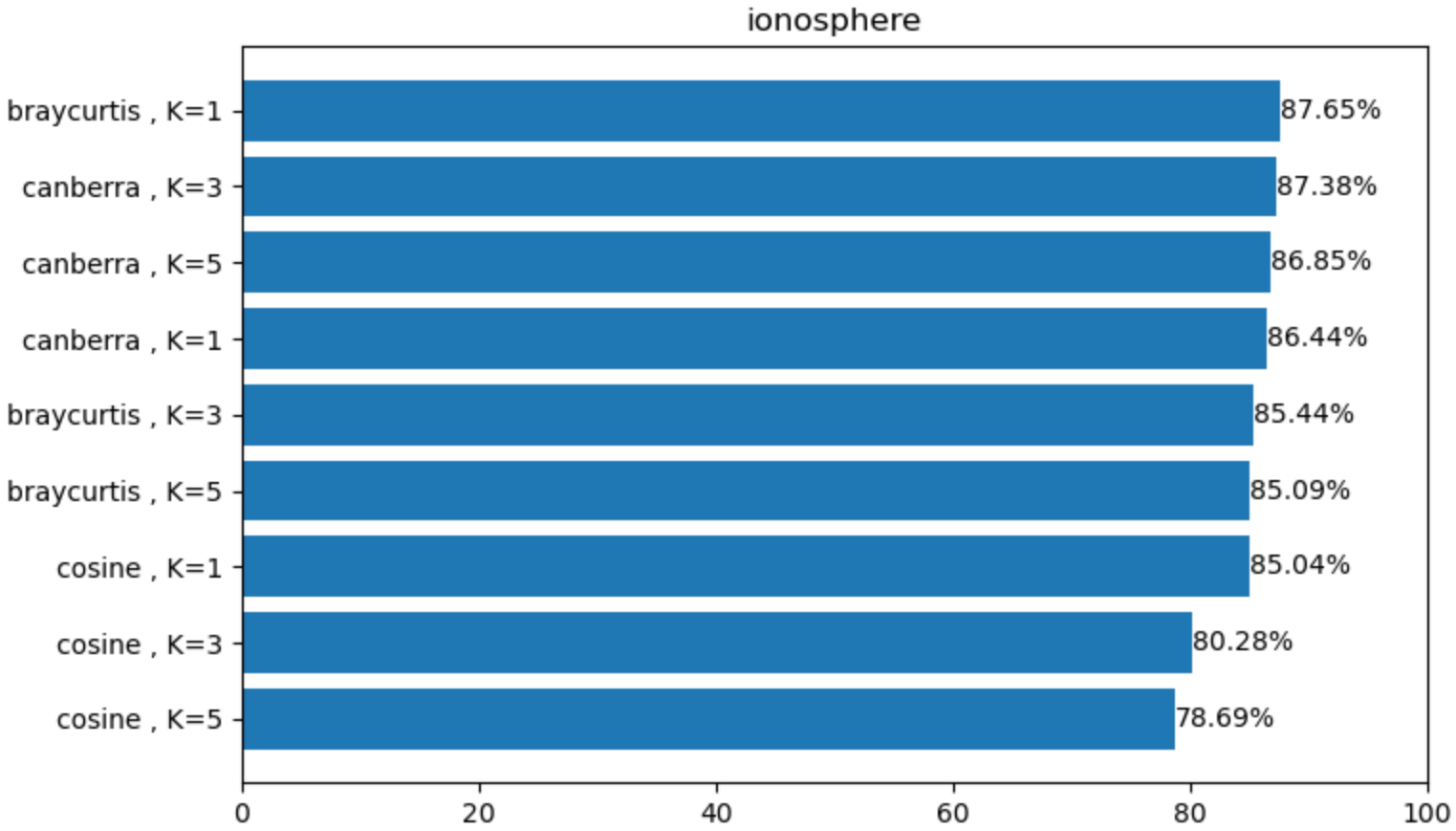
Haberman


haberman
hamming , K=3
53.92%
hamming , K=1
51.4%
hamming , K=5
50.88%
0
20
40
60
80
100

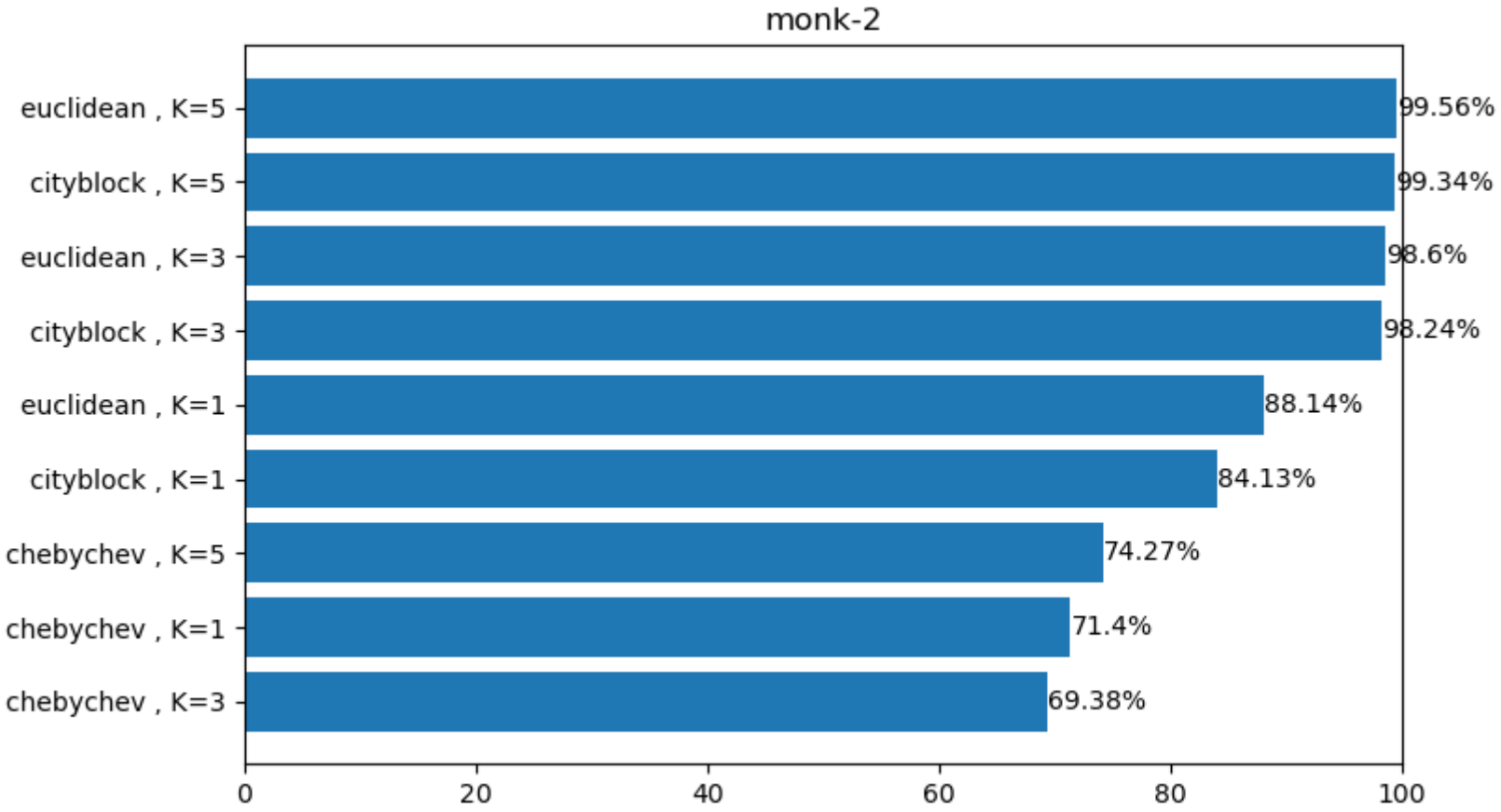
Ionosphere


ionosphere
cityblock , K=1
cityblock , K=3
cityblock , K=5
chebychev , K=1
chebychev , K=3
chebychev , K=5
euclidean , K=1
euclidean , K=3
euclidean , K=5
88.0%
85.22%
85.0%
84.42%
83.8%
83.06%
82.39%
79.66%
79.44%
0
20
40
60
80
100

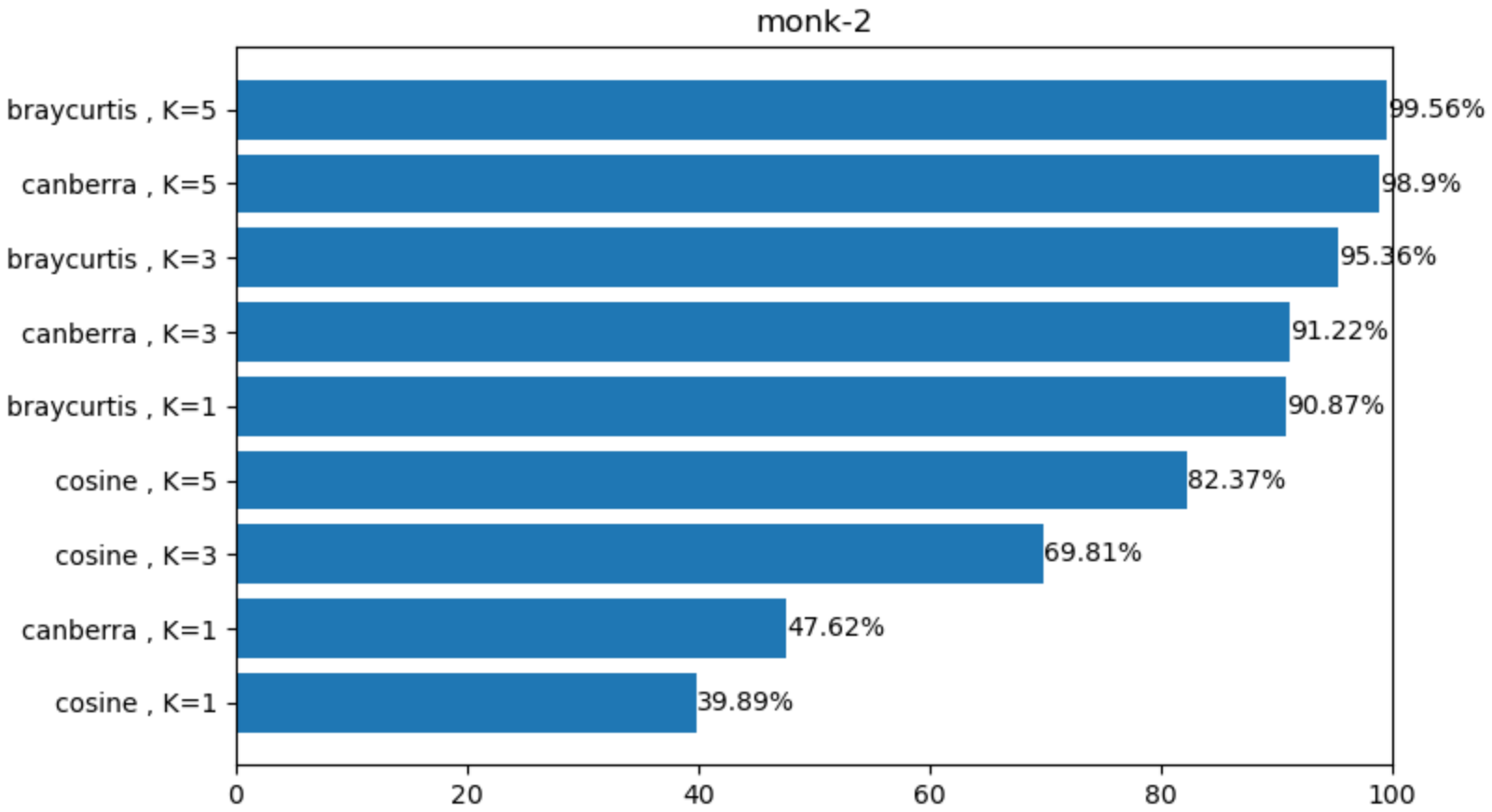
Ionosphere


ionosphere
braycurtis , K=1
canberra , K=3
canberra , K=5
canberra , K=1
braycurtis , K=3
braycurtis , K=5
cosine , K=1
cosine , K=3
cosine , K=5
87.65%
87.38%
86.85%
86.44%
85.44%
85.09%
85.04%
80.28%
78.69%
0
20
40
60
80
100

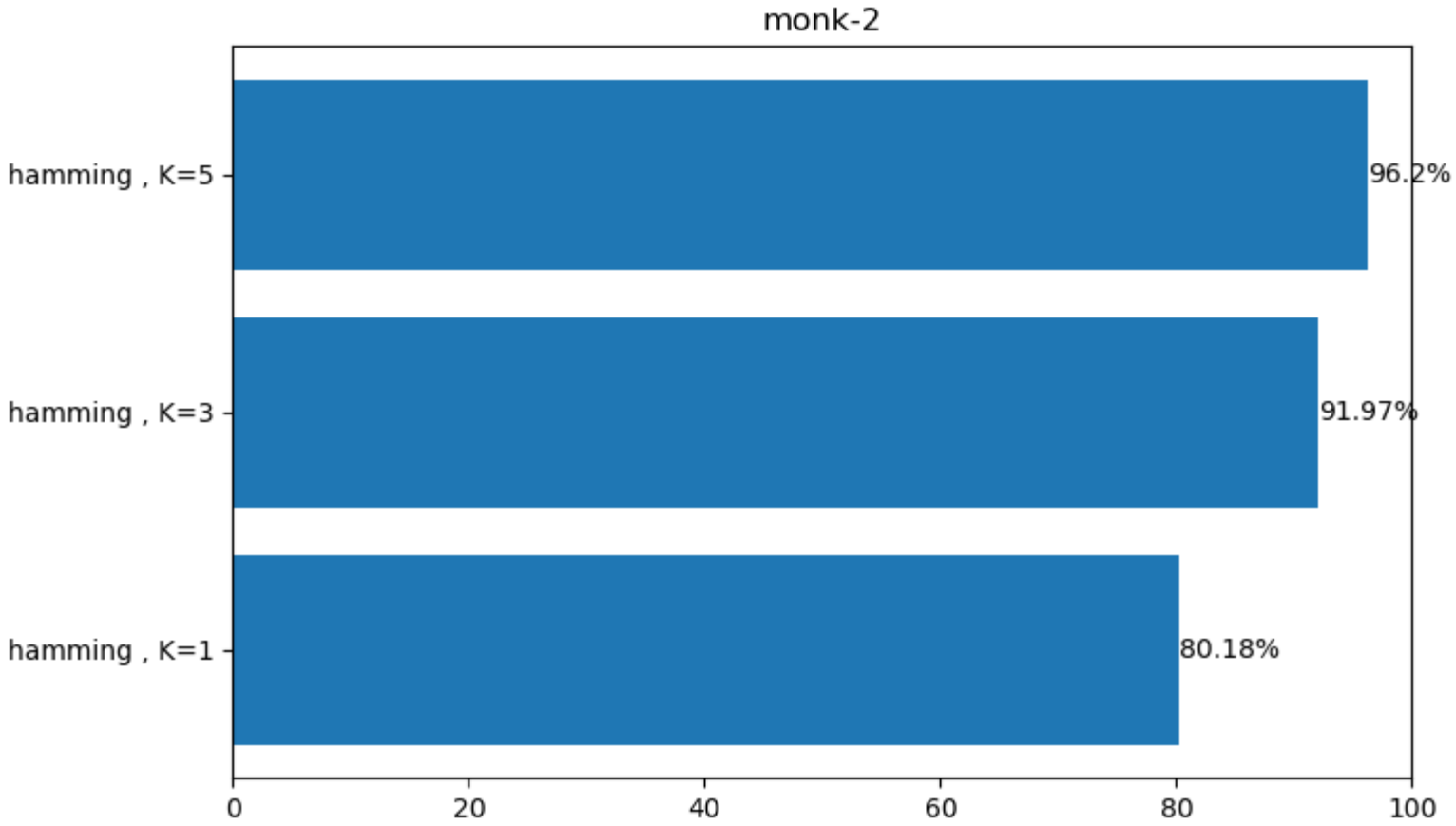
Monk-2


monk-2
euclidean , K=5
cityblock , K=5
euclidean , K=3
cityblock , K=3
euclidean , K=1
cityblock , K=1
chebychev , K=5
chebychev , K=1
chebychev , K=3
99.56%
99.34%
98.6%
98.24%
88.14%
84.13%
74.27%
71.4%
69.38%
0
20
40
60
80
100

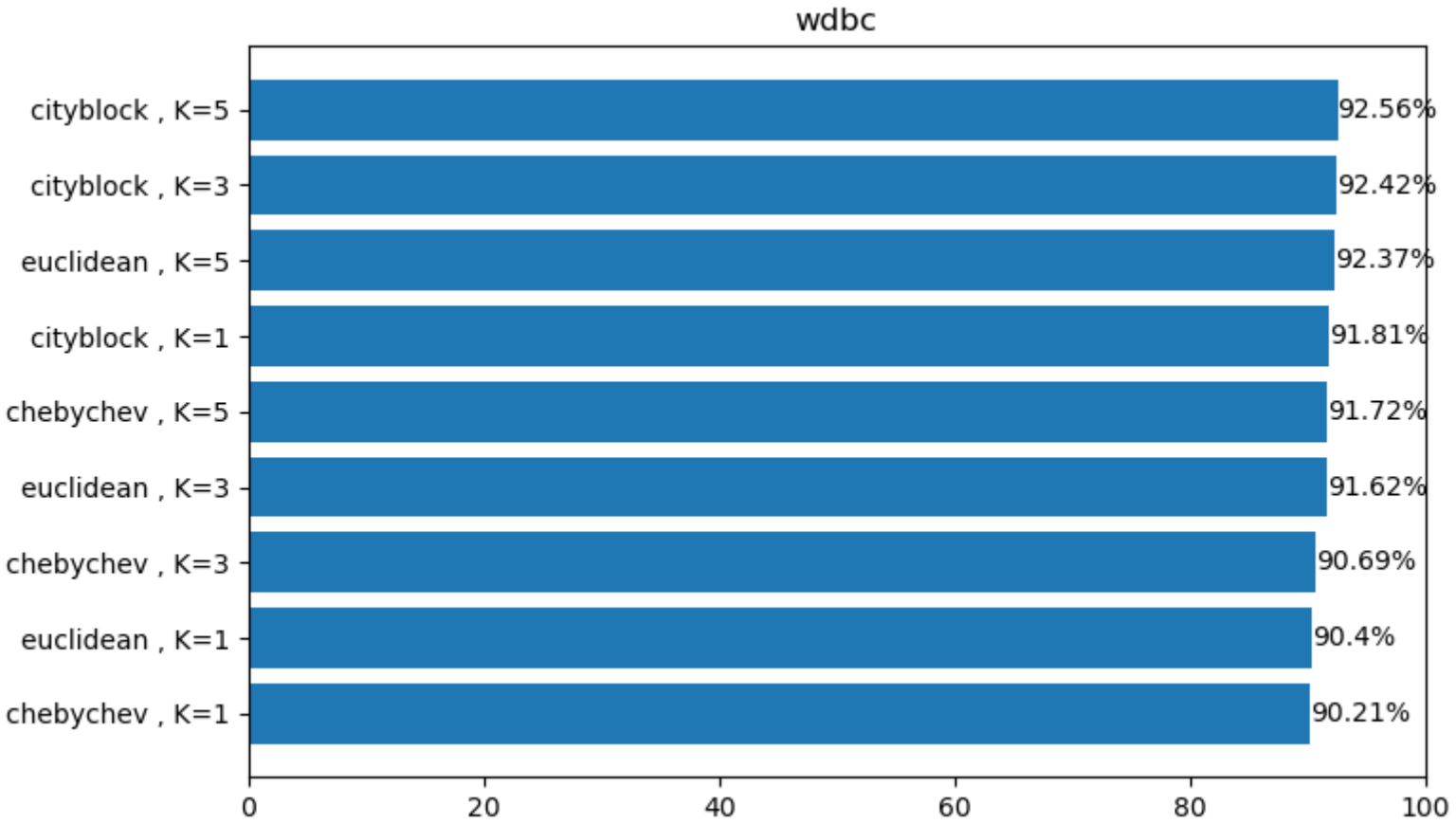
Monk-2


monk-2
braycurtis , K=5
canberra , K=5
braycurtis , K=3
canberra , K=3
braycurtis , K=1
cosine , K=5
cosine , K=3
canberra , K=1
cosine , K=1
99.56%
98.9%
95.36%
91.22%
90.87%
82.37%
69.81%
47.62%
39.89%
0
20
40
60
80
100

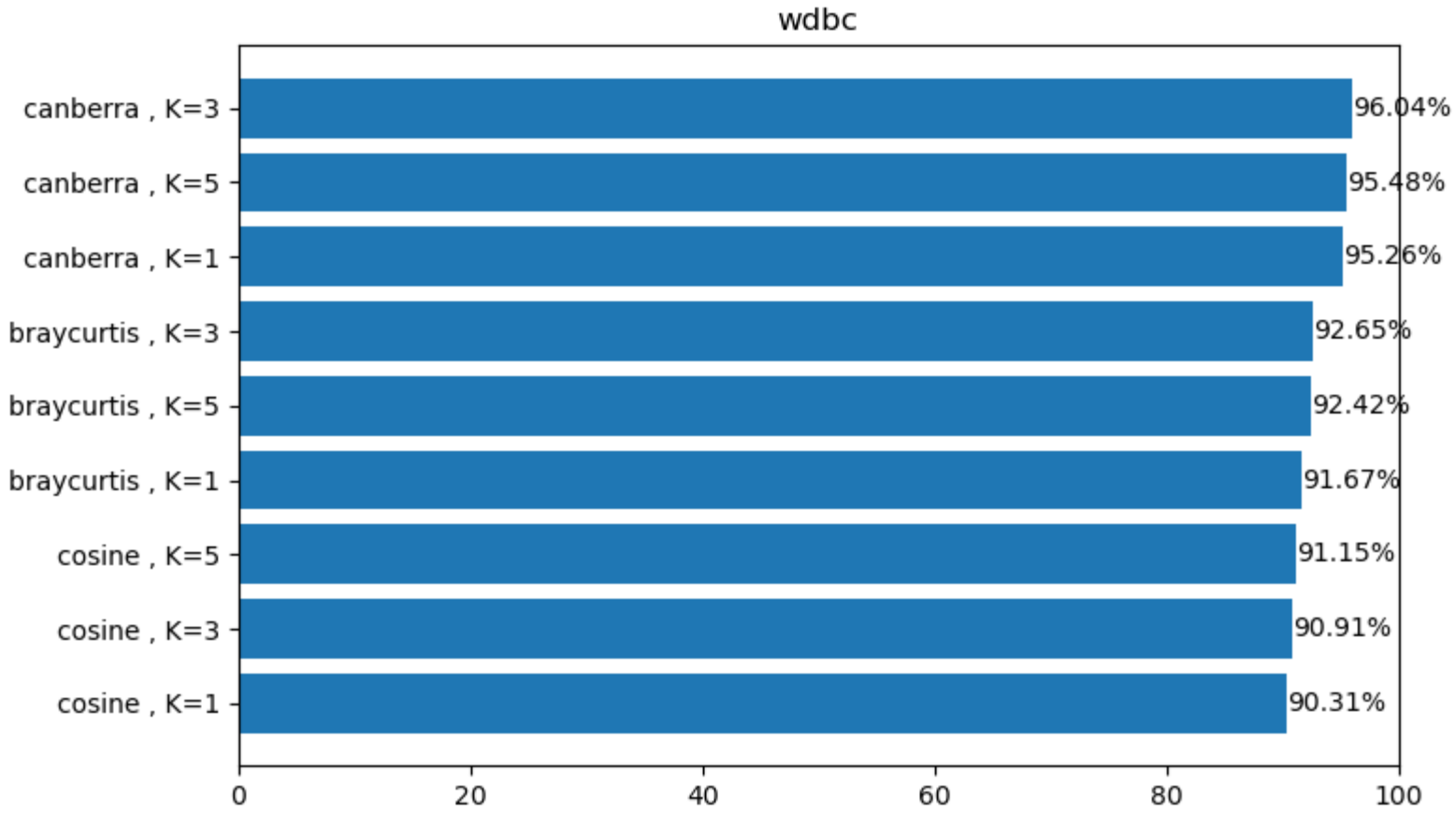
Monk-2


monk-2
hamming , K=5
hamming , K=3
hamming , K=1
96.2%
91.97%
80.18%
0
20
40
60
80
100

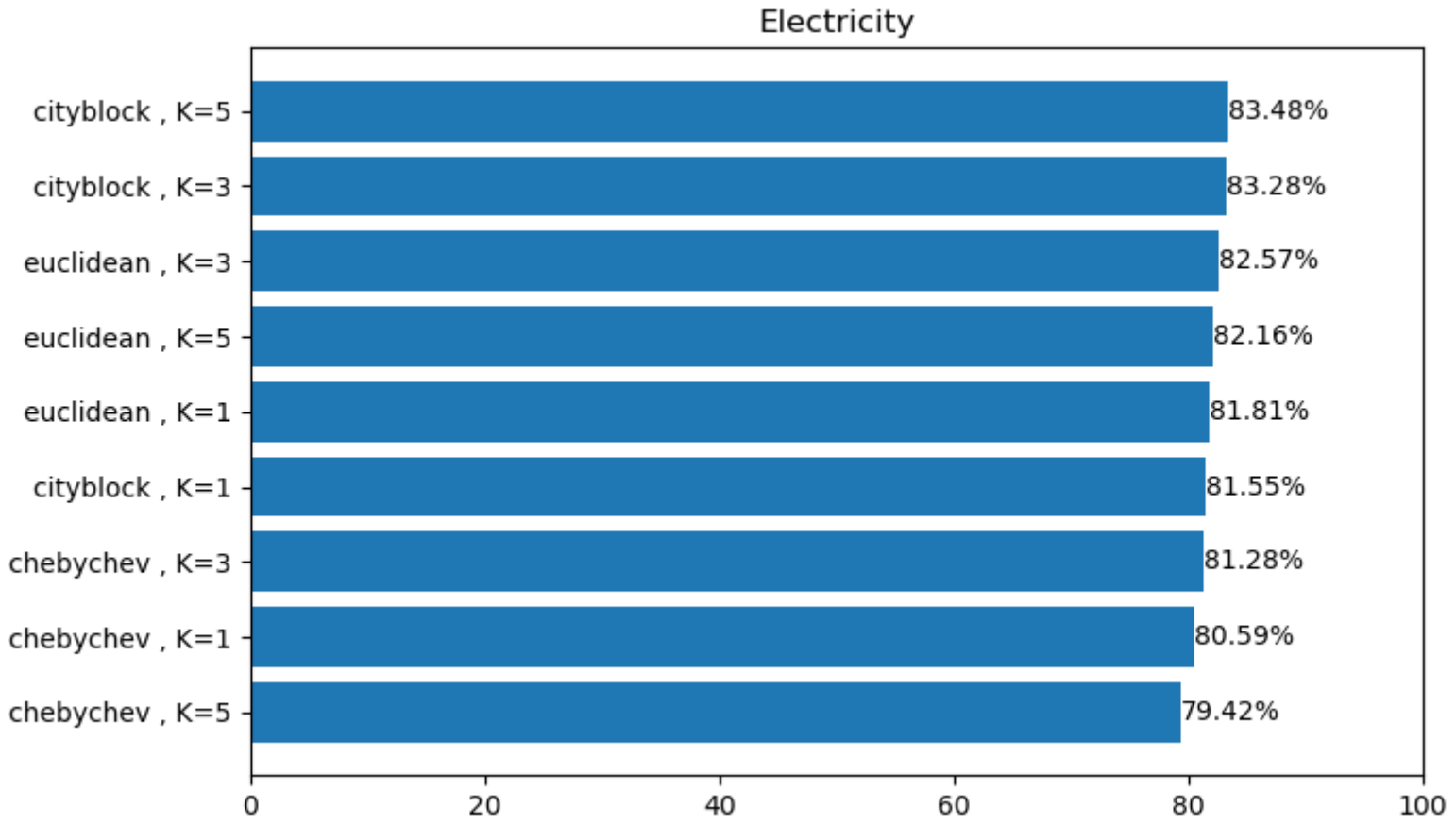
WDBC


wdbc
cityblock , K=5
cityblock , K=3
euclidean , K=5
cityblock , K=1
chebychev , K=5
euclidean , K=3
chebychev , K=3
euclidean , K=1
chebychev , K=1
92.56%
92.42%
92.37%
91.81%
91.72%
91.62%
90.69%
90.4%
90.21%
0
20
40
60
80
100

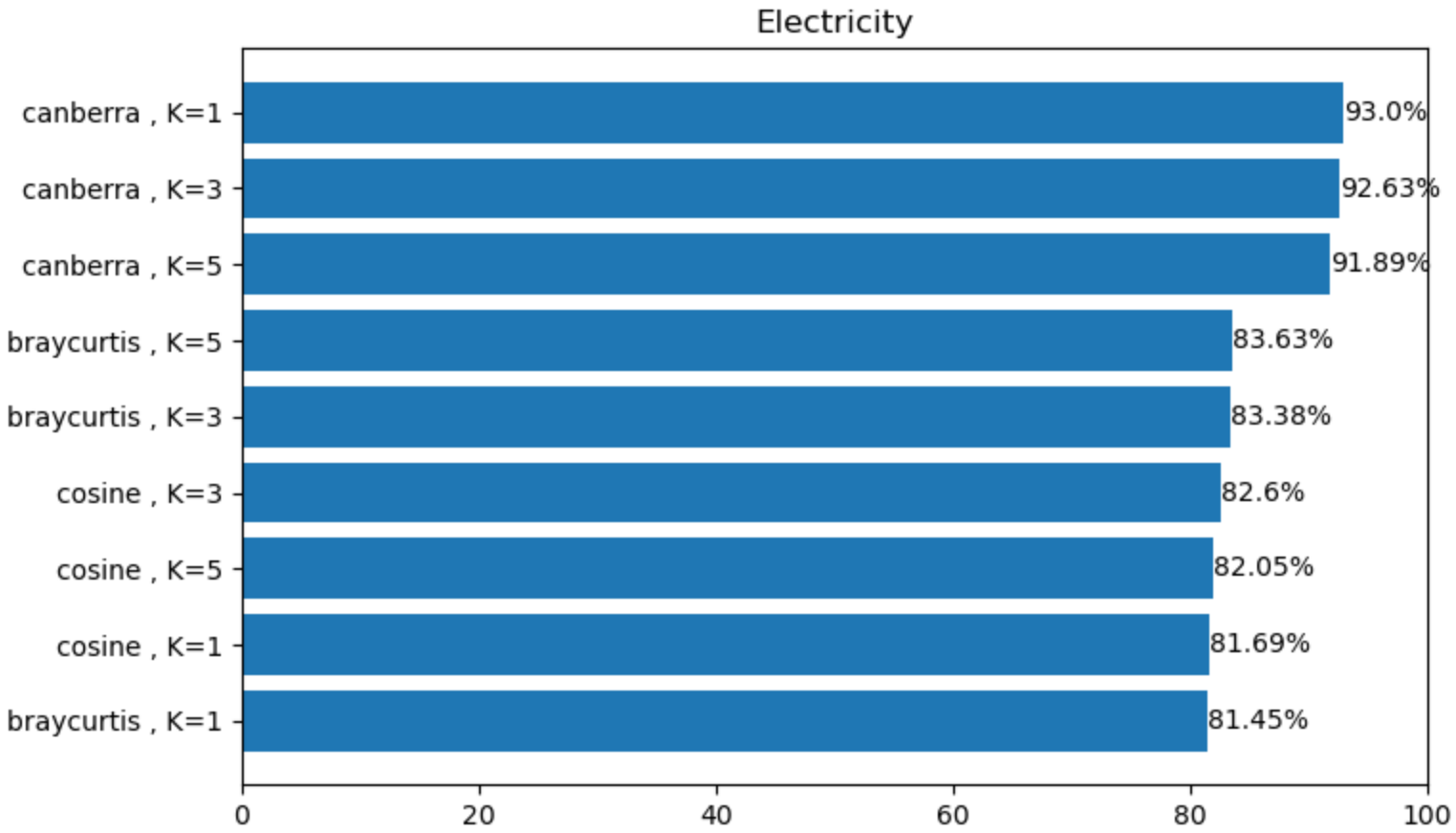
WDBC


wdbc
canberra , K=3
canberra , K=5
canberra , K=1
braycurtis , K=3
braycurtis , K=5
braycurtis , K=1
cosine , K=5
cosine , K=3
cosine , K=1
96.04%
95.48%
95.26%
92.65%
92.42%
91.67%
91.15%
90.91%
90.31%
0
20
40
60
80
100

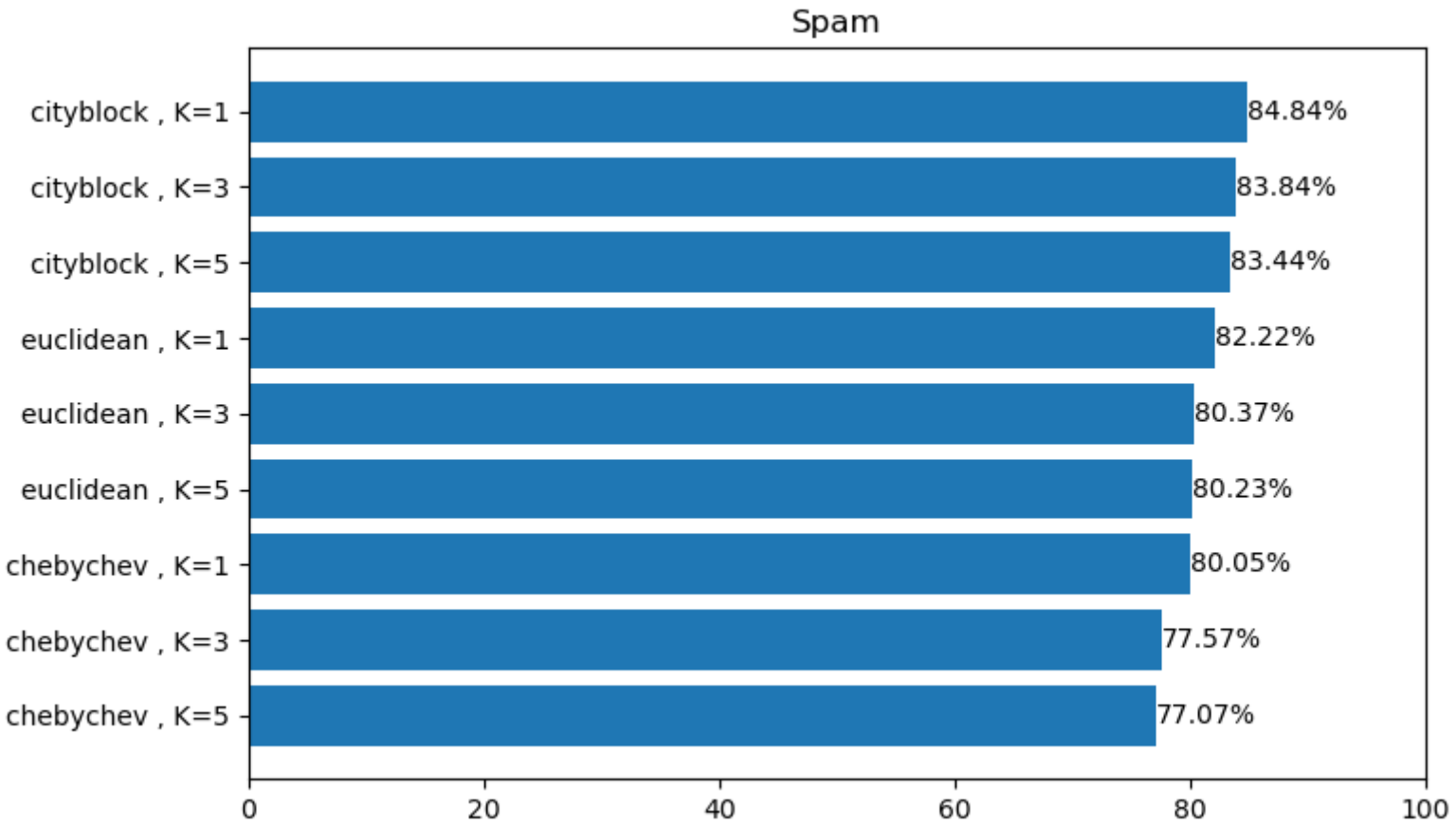
Electricity



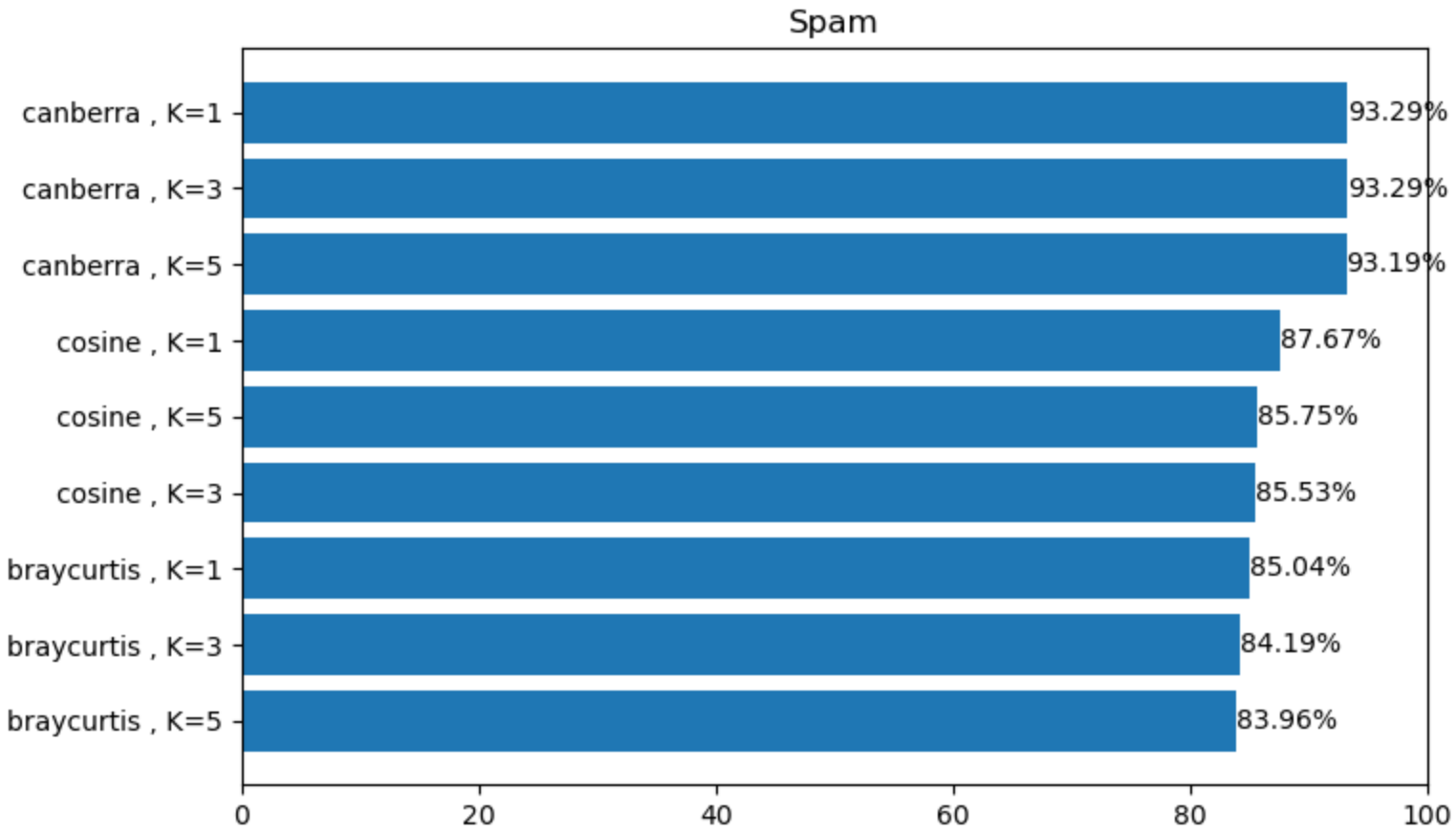
Electricity



Spam


Spam
cityblock , K=1
cityblock , K=3
cityblock , K=5
euclidean , K=1
euclidean , K=3
euclidean , K=5
chebychev , K=1
chebychev , K=3
chebychev , K=5
84.84%
83.84%
83.44%
82.22%
80.37%
80.23%
80.05%
77.57%
77.07%
0
20
40
60
80
100

Spam


Spam
canberra , K=1
canberra , K=3
canberra , K=5
cosine , K=1
cosine , K=5
cosine , K=3
braycurtis , K=1
braycurtis , K=3
braycurtis , K=5
93.29%
93.29%
93.19%
87.67%
85.75%
85.53%
85.04%
84.19%
83.96%
0
20
40
60
80
100

Top 3 Metrics

Dataset	Metric	K	BA
Nutt	Braycurtis	3	89.28
	Braycurtis	5	89.28
	Chebychev	1	82.14
Apendicitis	Canberra	1	79.21
	Canberra	5	77.98
	Cityblock	1	76.21
Sonar	Canberra	1	85.82
	Cityblock	1	84.86
	Braycurtis	1	84.15
Heart	Canberra	5	82.55
	Canberra	3	78.49
	Canberra	1	78.08
Haberman	Chebychev	5	61.38
	Cosine	1	60
	Canberra	5	59.60

Dataset	Metric	K	BA
Ionosphere	Citiblock	1	88
	Braycurtis	1	87.65
	Canberra	3	87.38
Monk-2	Euclidean	5	99.56
	Braycurtis	5	99.56
	Cityblock	5	99.34
WDBC	Canberra	3	96.04
	Canberra	5	95.48
	Canberra	1	95.26
Electricity	Canberra	1	93
	Canberra	3	92.63
	Canberra	5	91.89
Spam	Canberra	1	93.29
	Canberra	3	93.29
	Canberra	5	93.19

Frequency

Metric	K	Frequency
Canberra	1	6
	5	6
	3	5
Citiblock	1	3
Braycurtis	1	2
	5	2
Braycurtis	3	1
Chebychev	1	1
	5	1
Cityblock	5	1
Cosine	1	1
Euclidean	5	1

Metric	Frequency
Canberra	17
Braycurtis	5
Cityblock	4
Chebychev	2
Cosine	1
Euclidean	1

K	Frequency
1	13
3	11
5	6

Thank you !

Questions?