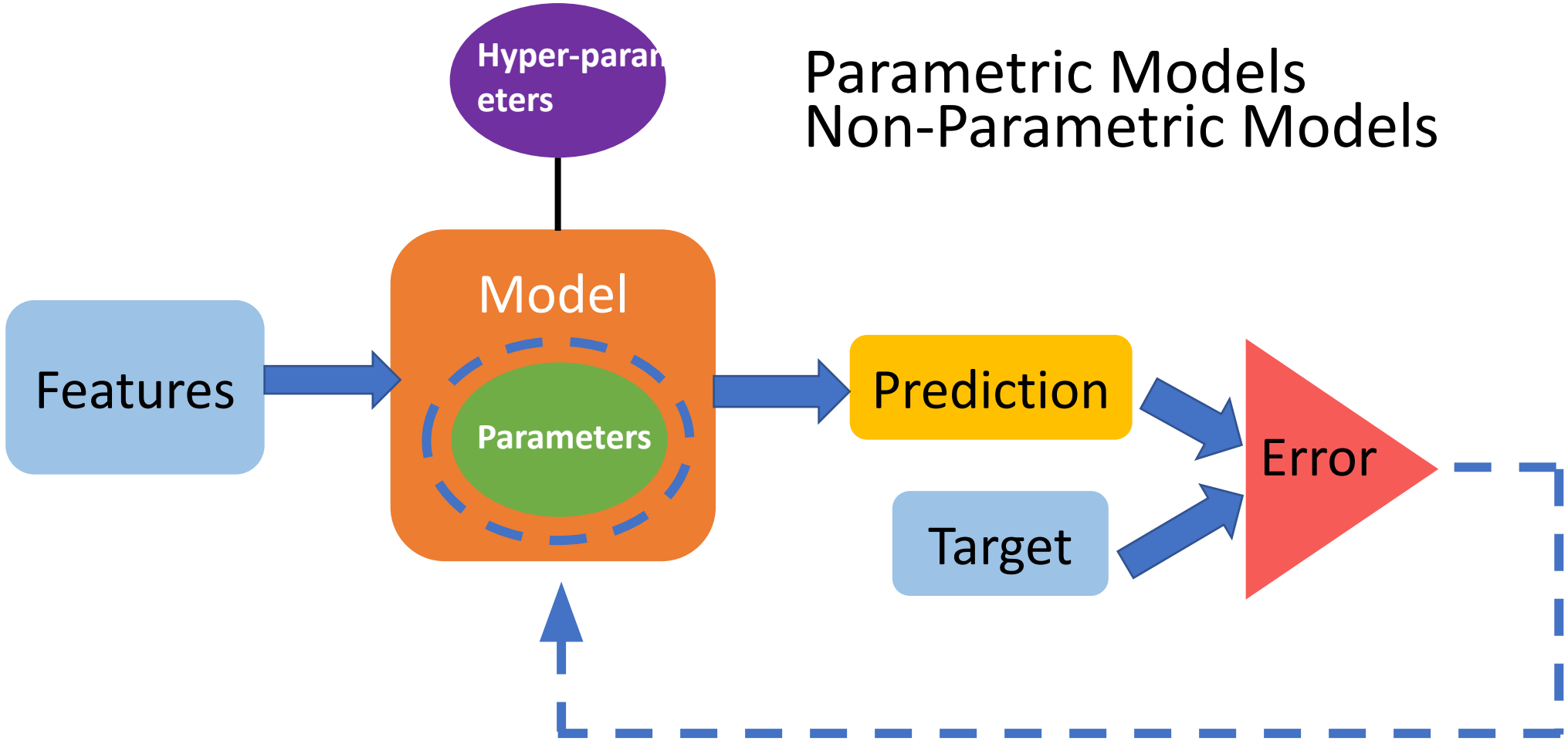
Non-parametric Models Intro

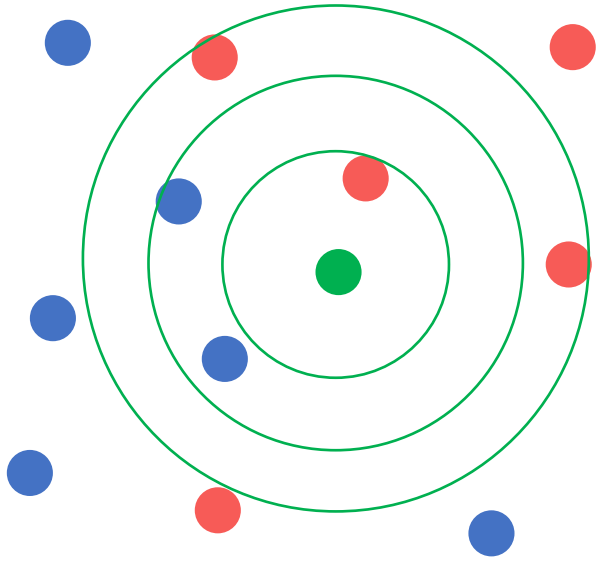
Non-parametric models



Non-parametric models

- K-Nearest Neighbors
- Decision Trees
- Support Vector Machine

K-Nearest Neighbors



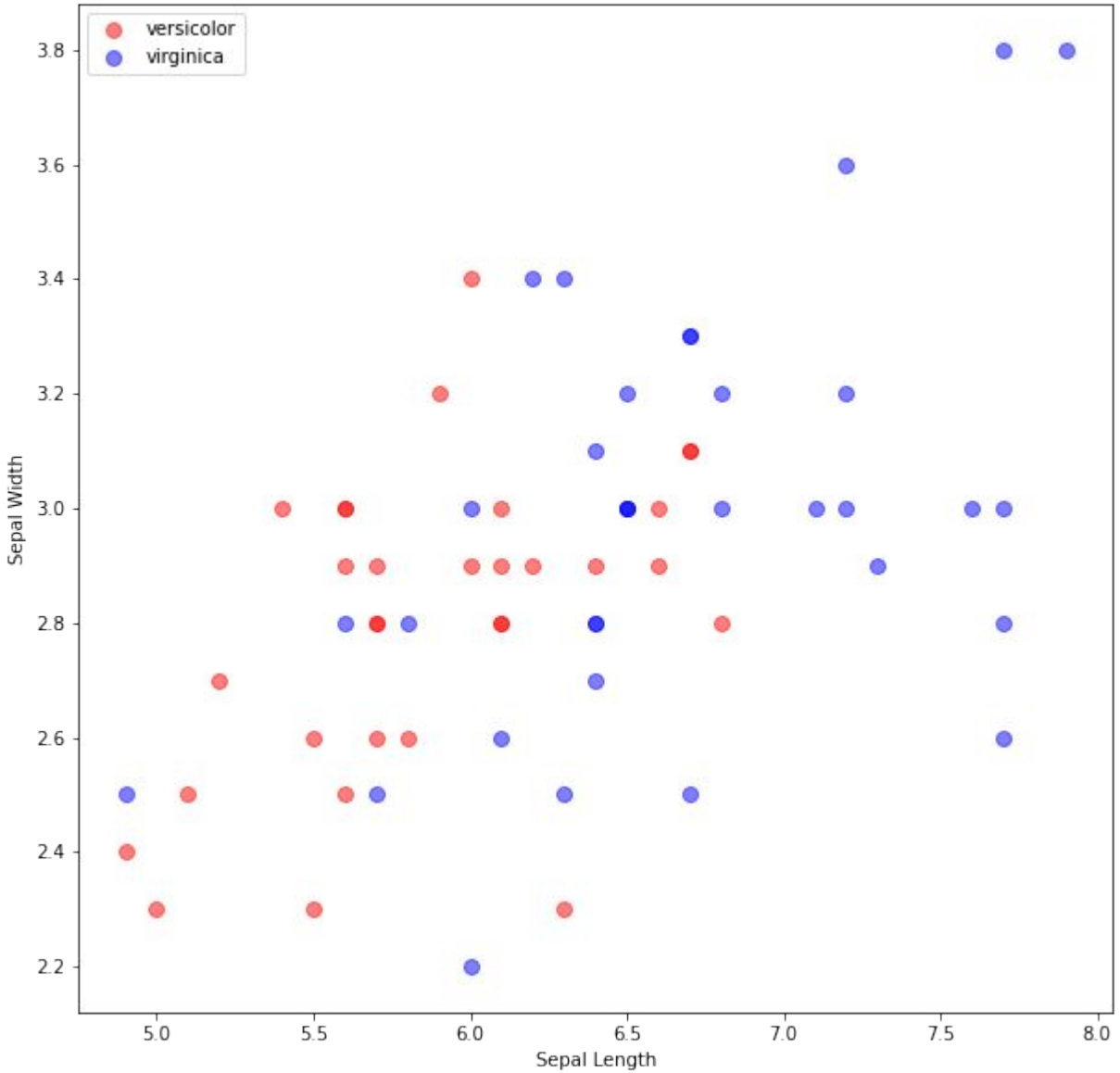
K=1
K=3
K=5

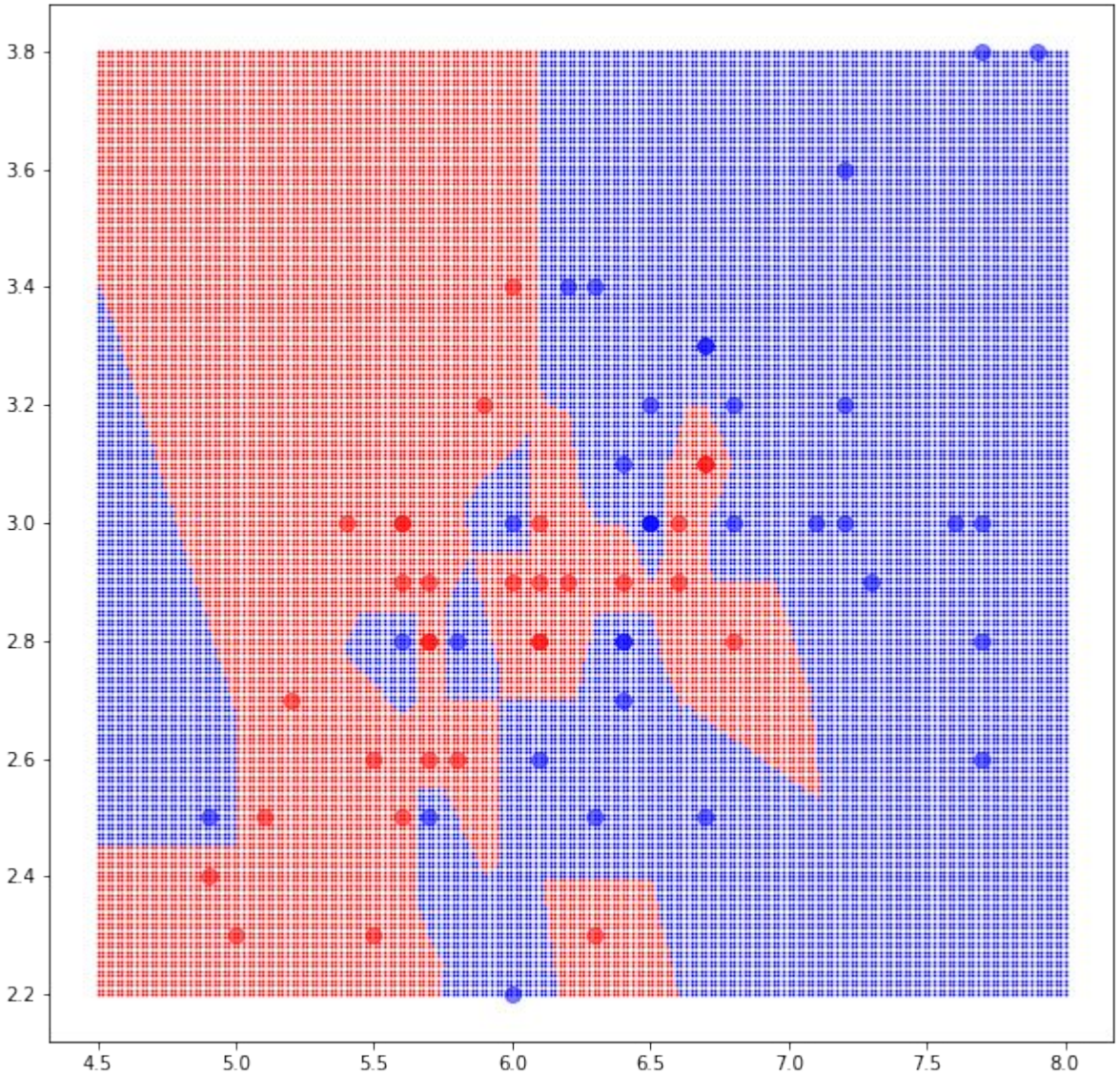
Manhattan distance

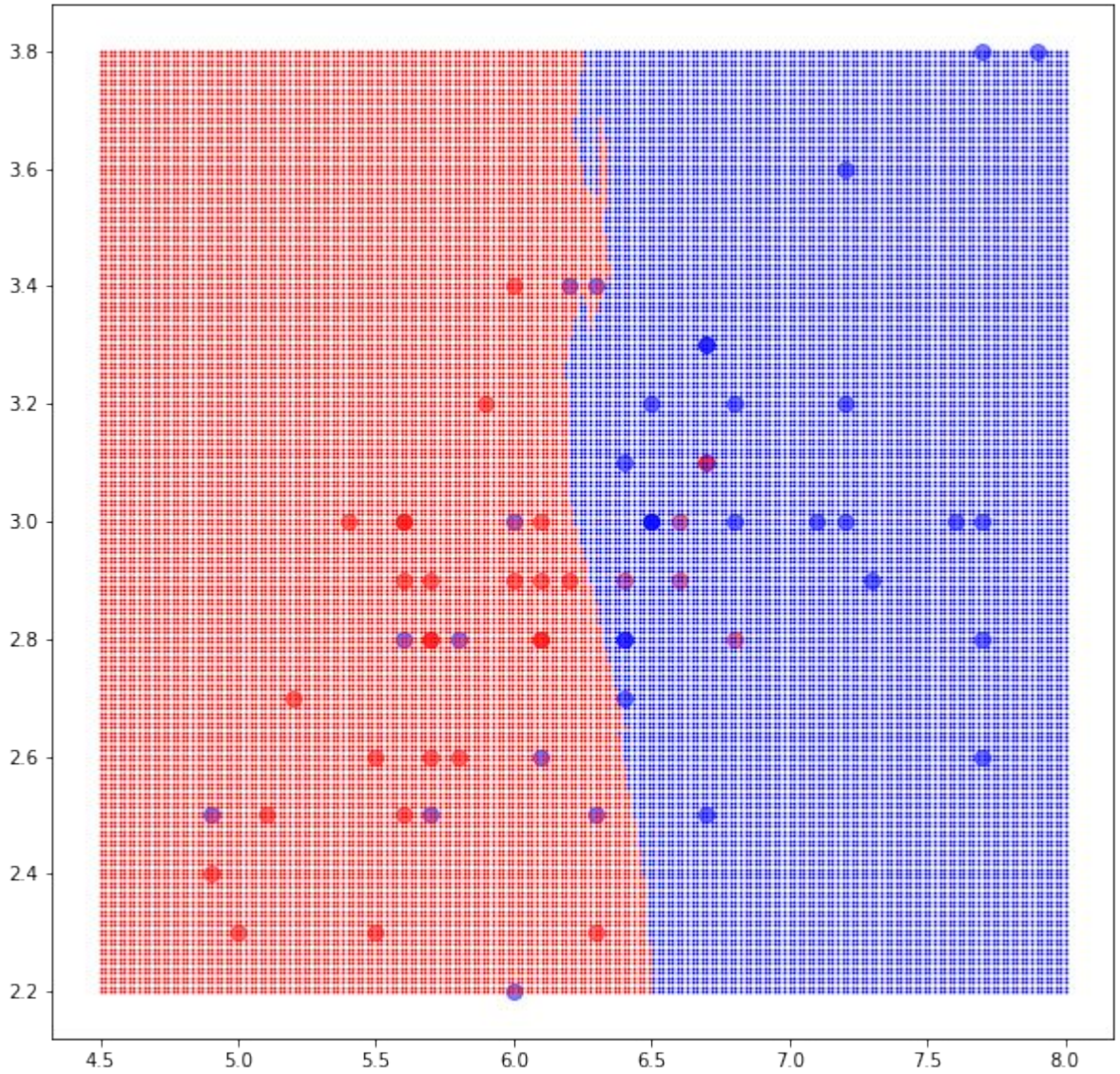
$$||\vec{X}_1 - \vec{X}_2||_1$$

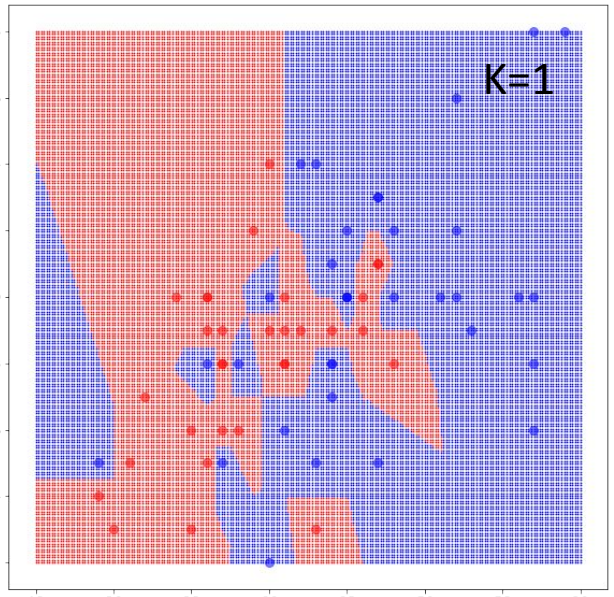
Euclidean distance

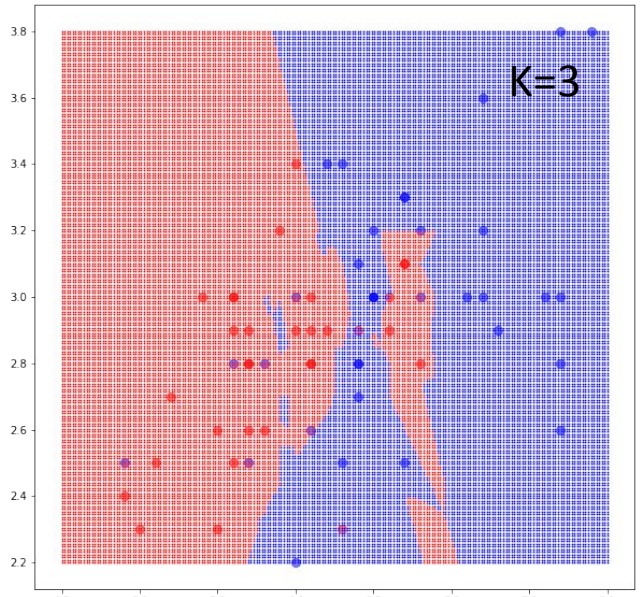
$$||\vec{X}_1 - \vec{X}_2||_2$$

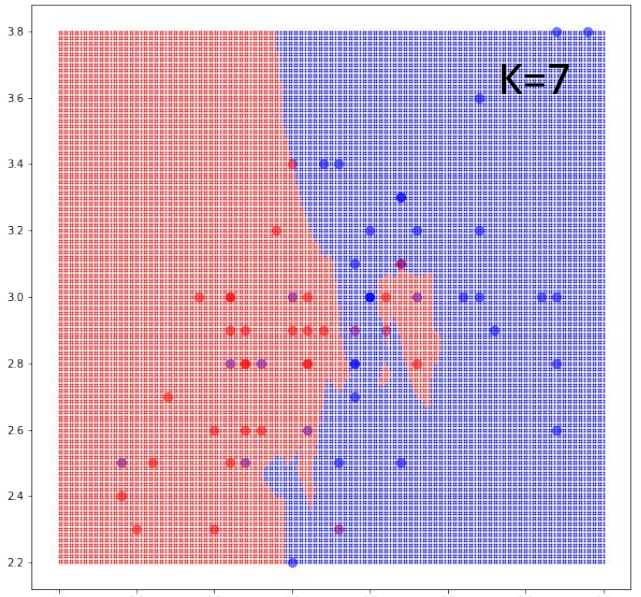

versicolor
virginica
Sepal Width
Sepal Length
3.8
3.6
3.4
3.2
3.0
2.8
2.6
2.4
2.2
5.0
5.5
6.0
6.5
7.0
7.5
8.0

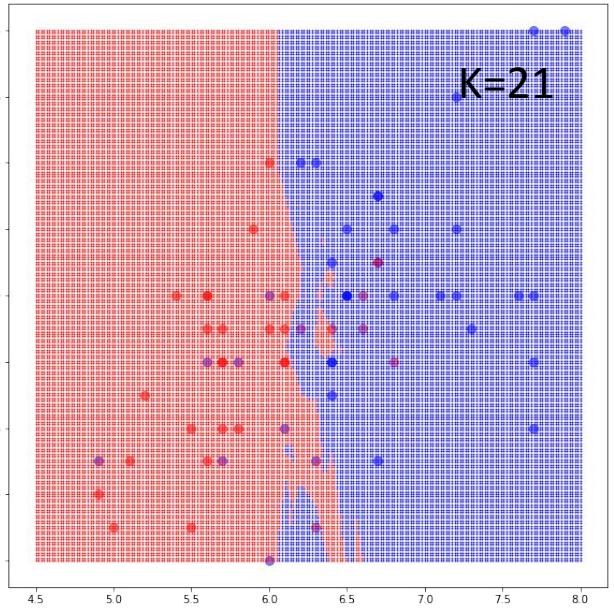

3.8
3.6
3.4
3.2
3.0
2.8
2.6
2.4
2.2
4.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0

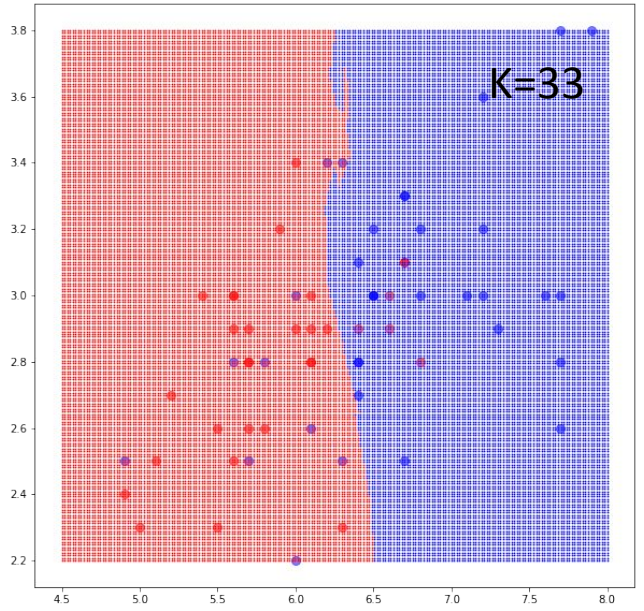

3.8
3.6
3.4
3.2
3.0
2.8
2.6
2.4
2.2
4.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0

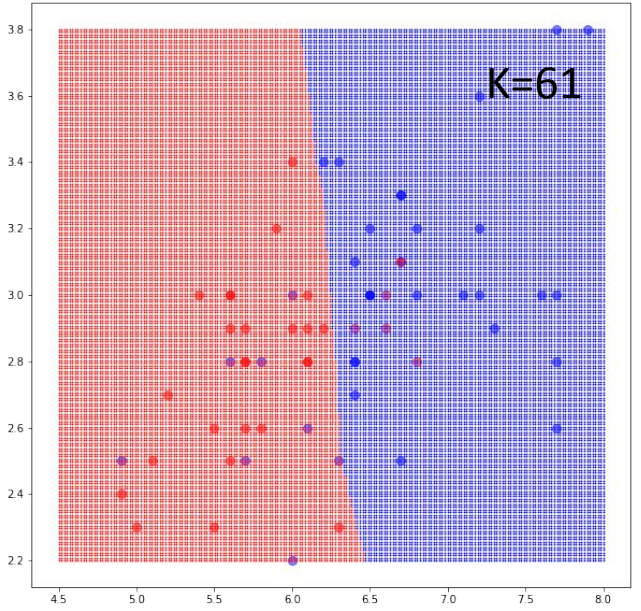

K=1


K=3


K=7


K=21


K=33


K=61

Bias and Variance of KNN

1. Which model has larger bias?

    a. smaller k

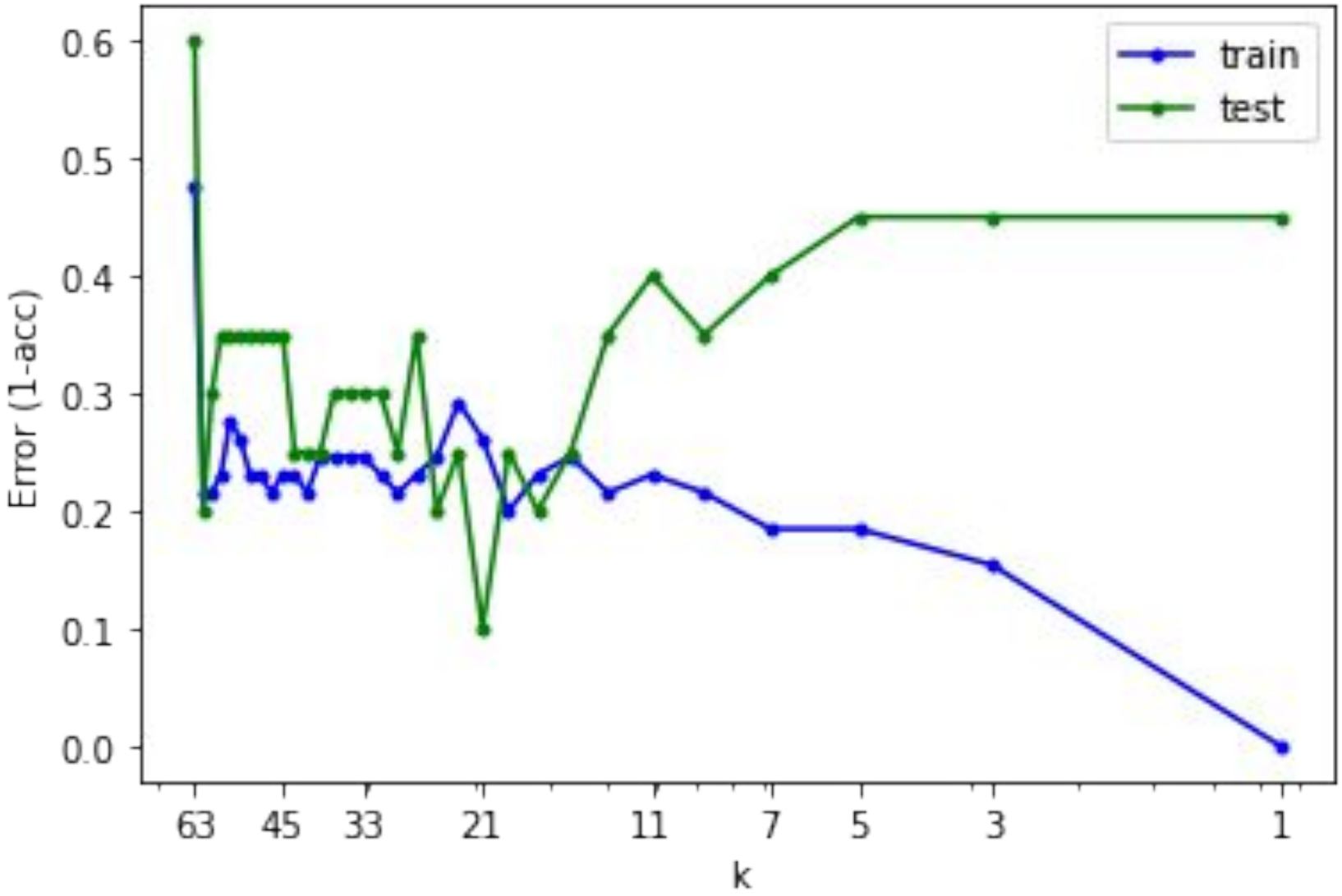
    b. larger k

Bias and Variance of KNN

2. Which model has larger variance?

a. smaller k

b. larger k

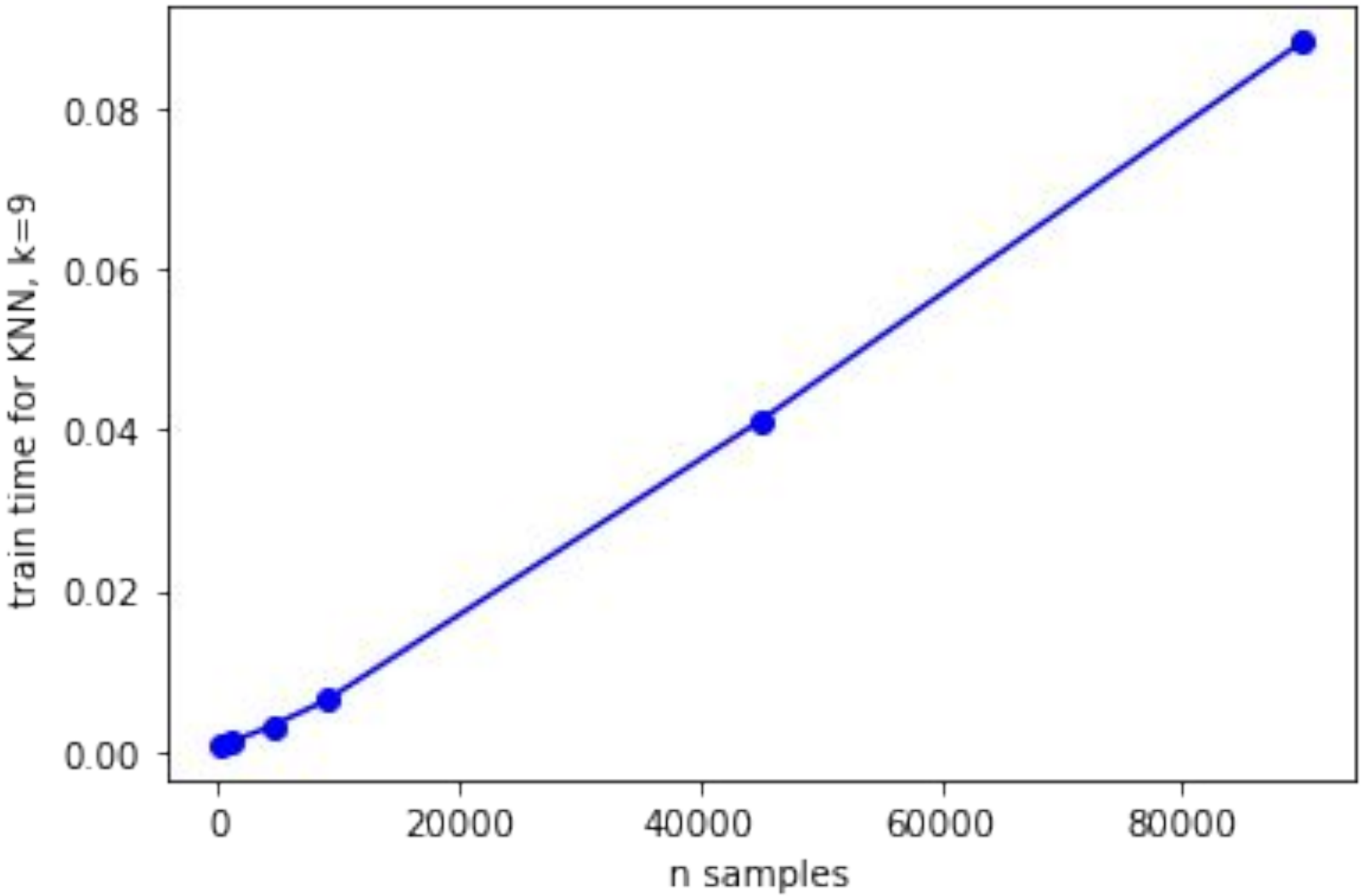

train
test
Error (1-acc)
0.6
0.5
0.4
0.3
0.2
0.1
0.0
63
45
33
21
11
7
5
3
1
k

KNN properties

- Simple, memory-based algorithm
- Time complexity $O(nm)$

KNN properties

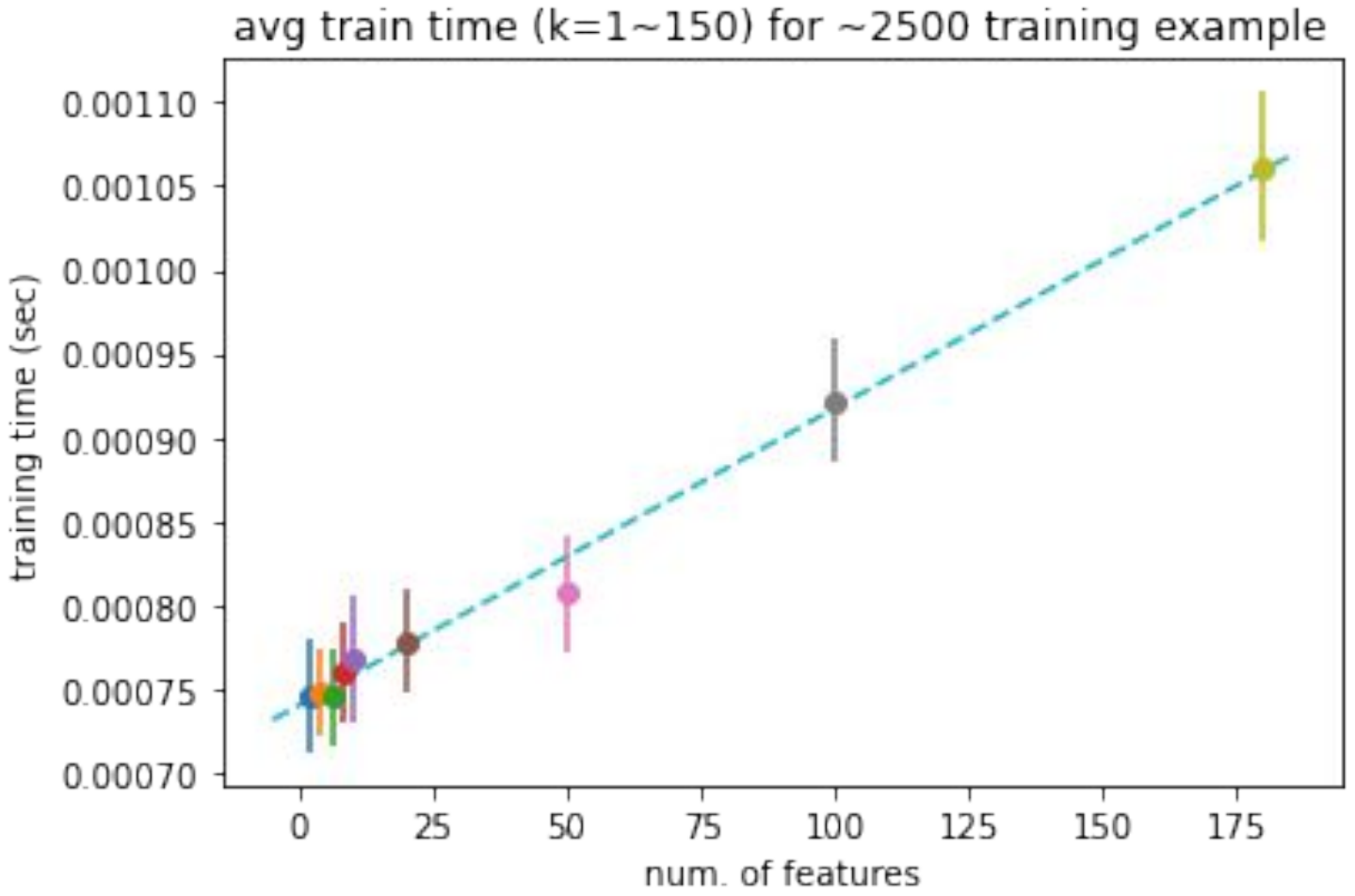
SensIT Vehicle dataset
90k+ samples
100 features
binary class



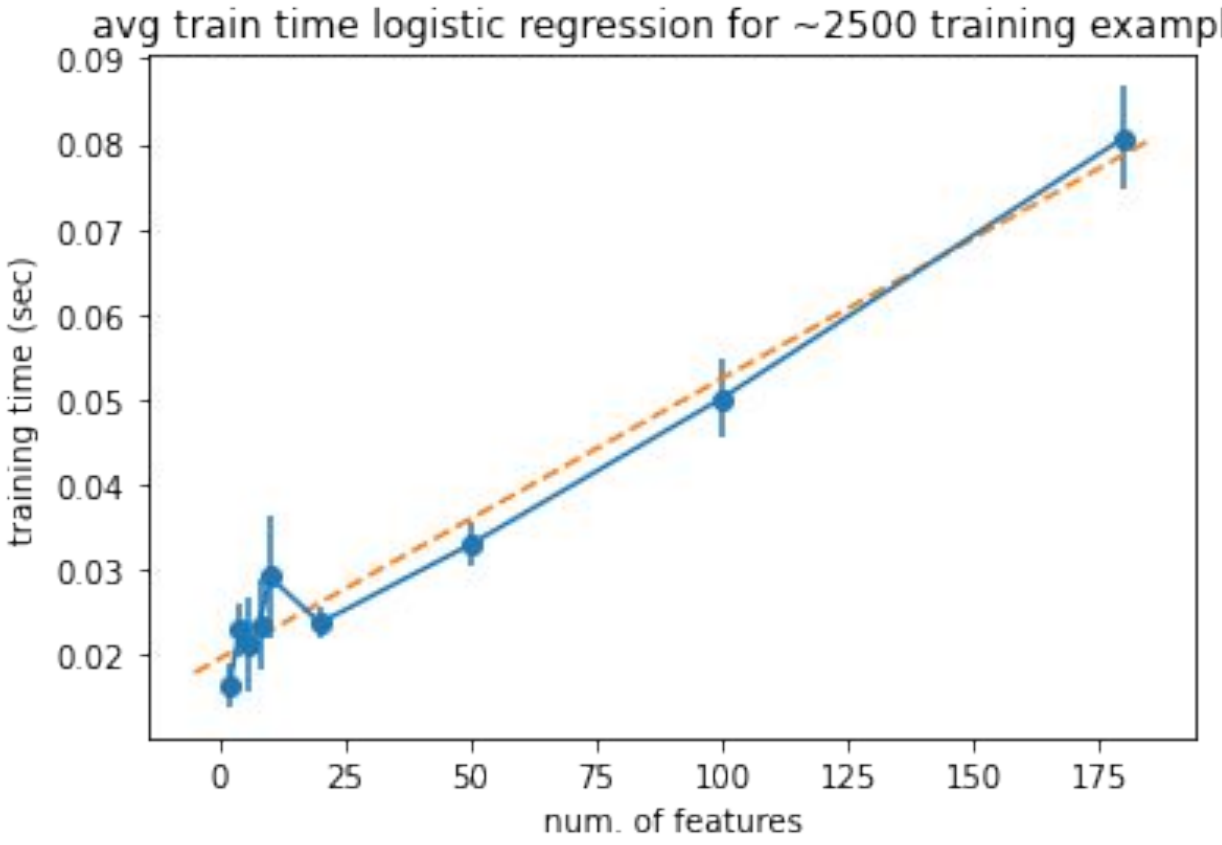
KNN properties- Example

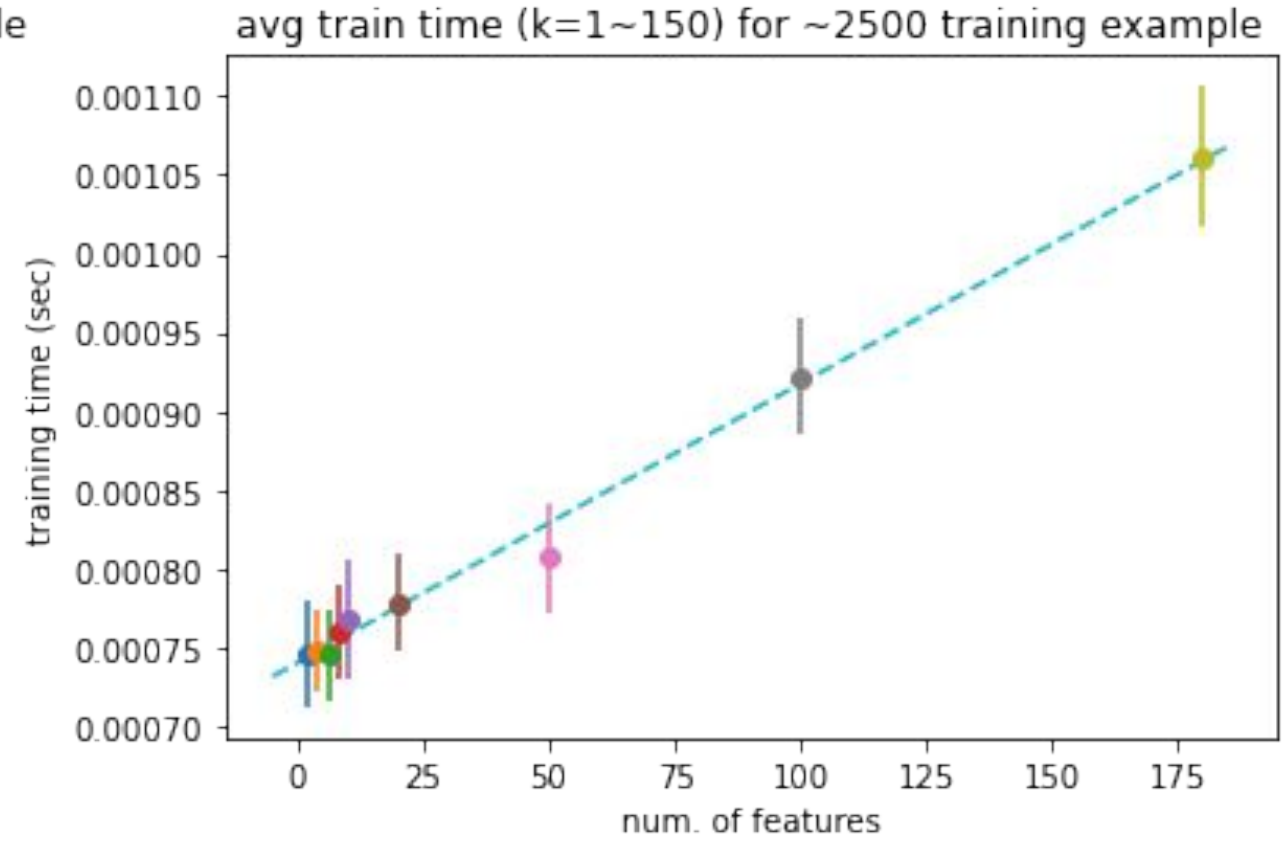
Splice-Junction Gene Sequences (DNA) dataset

180 binary features coding 60 ACGT nucleotide
3 classes (exon boundaries)

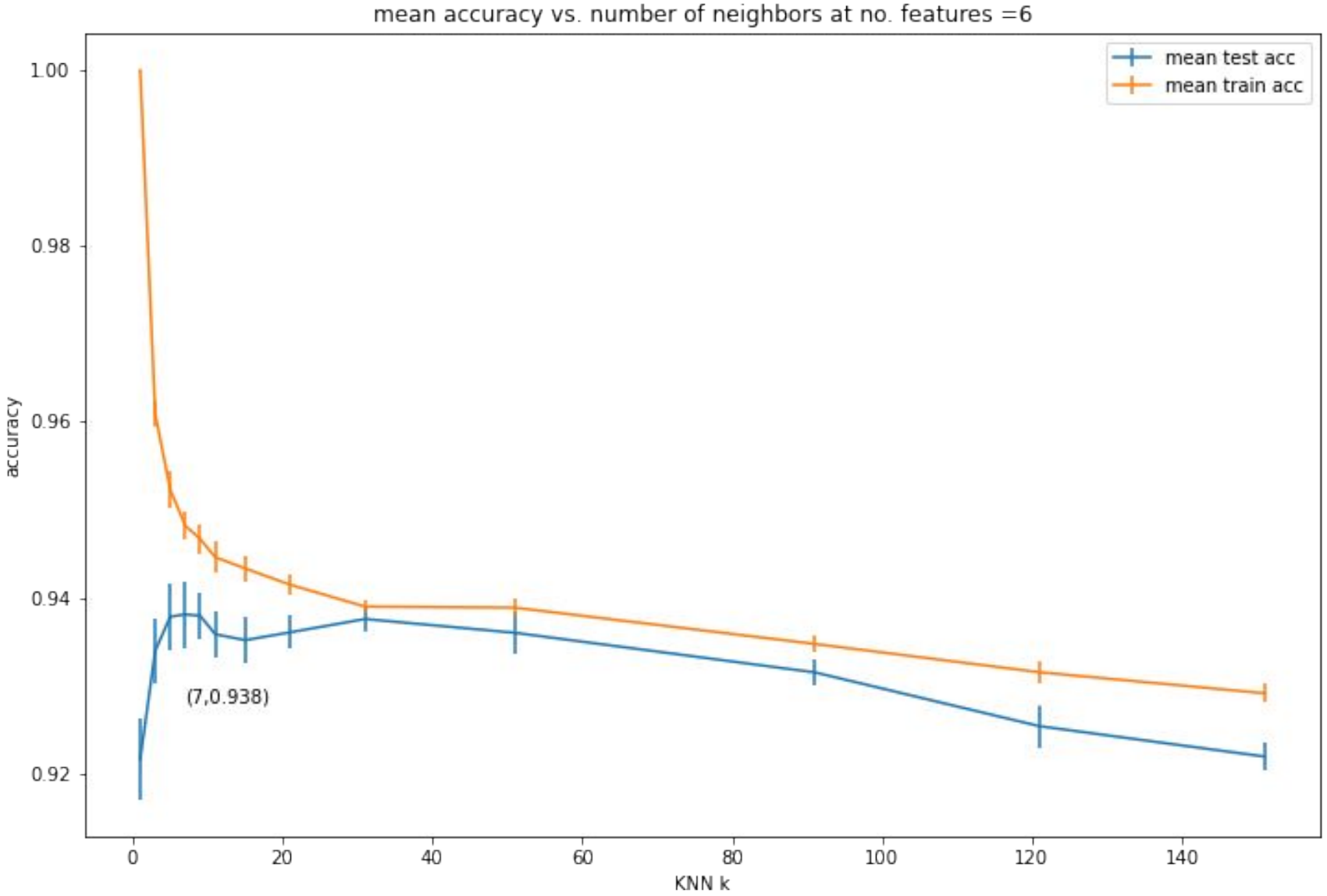


KNN properties

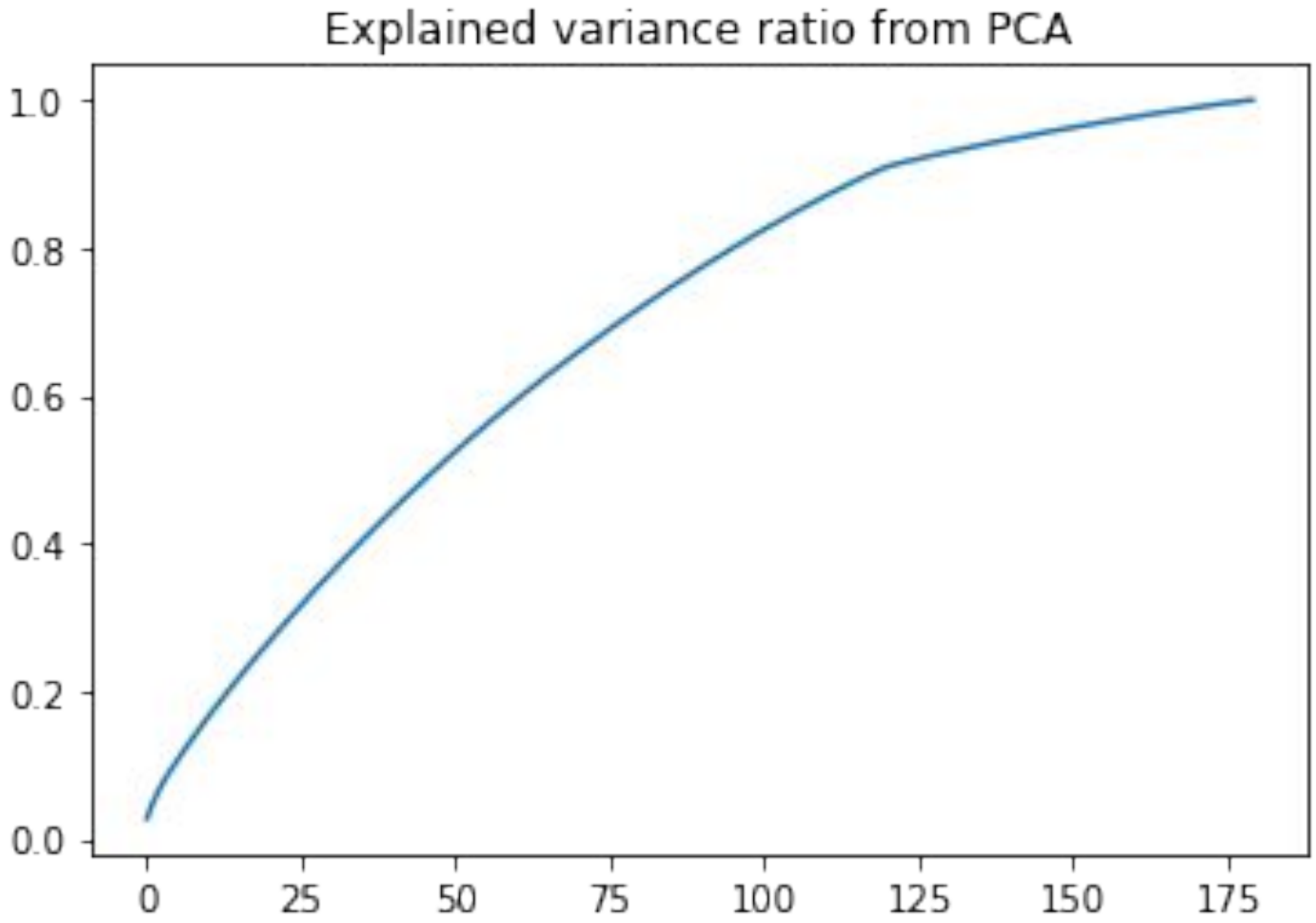




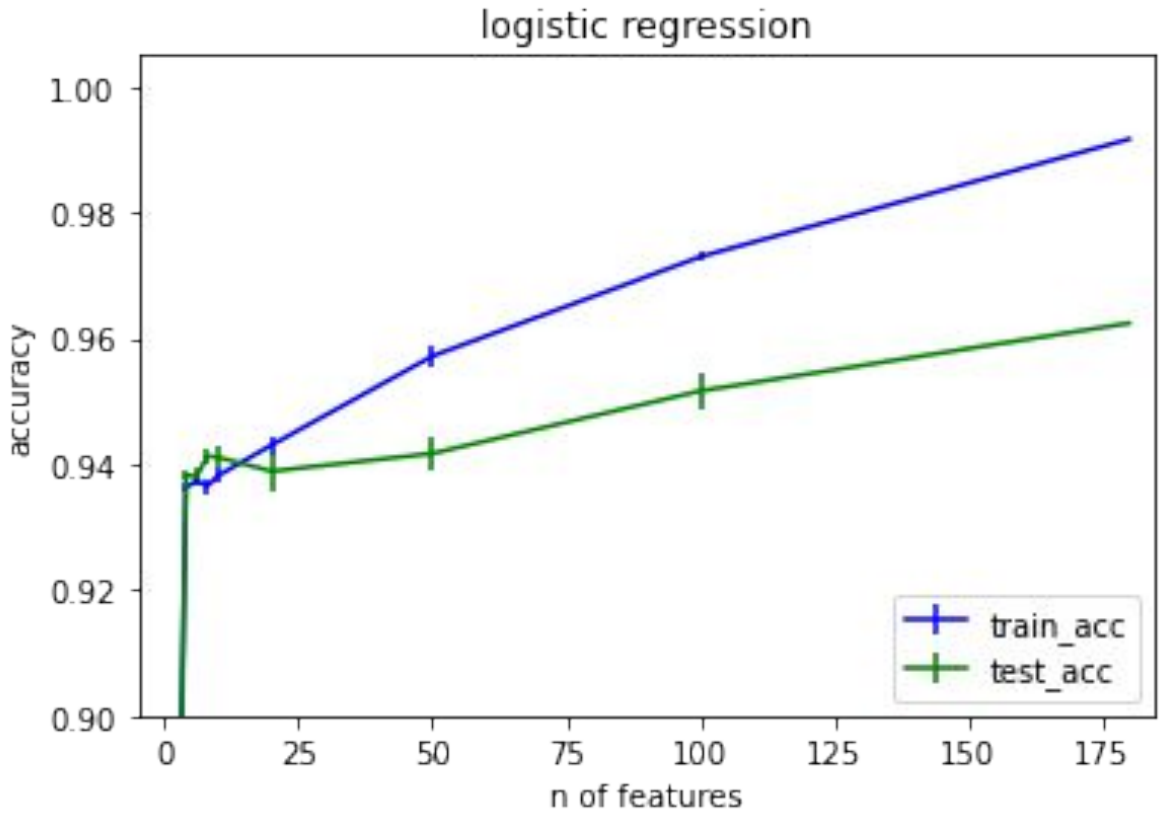
KNN properties

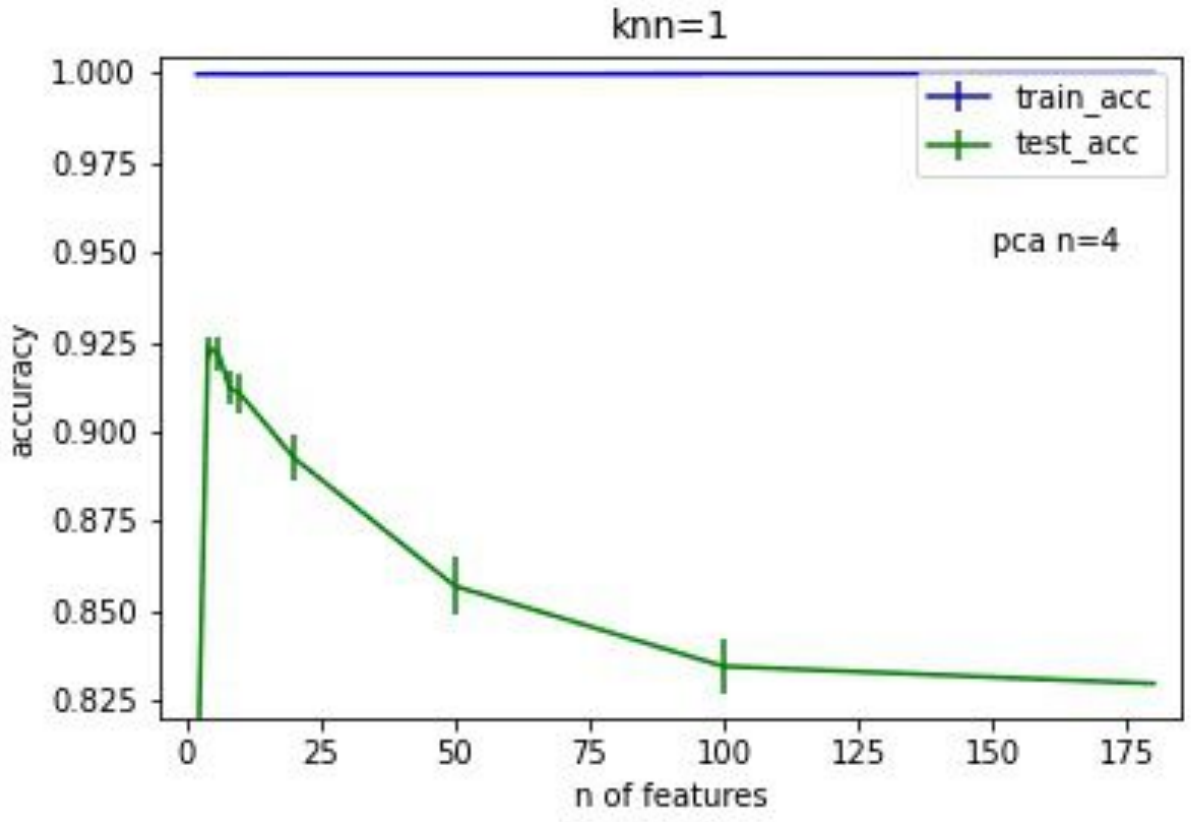


KNN Curse of dimensionality?



KNN Curse of dimensionality?





KNN Curse of dimensionality?