K Nearest Neighbors (KNN)

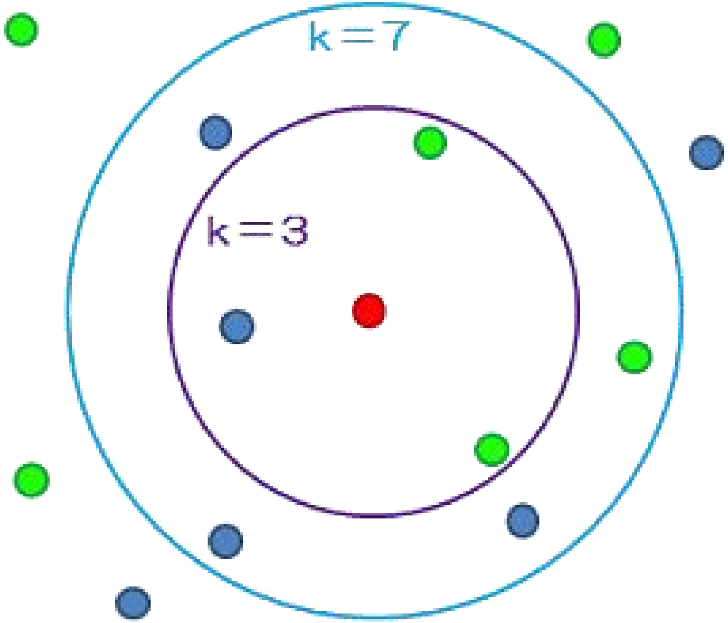
Swara Salih

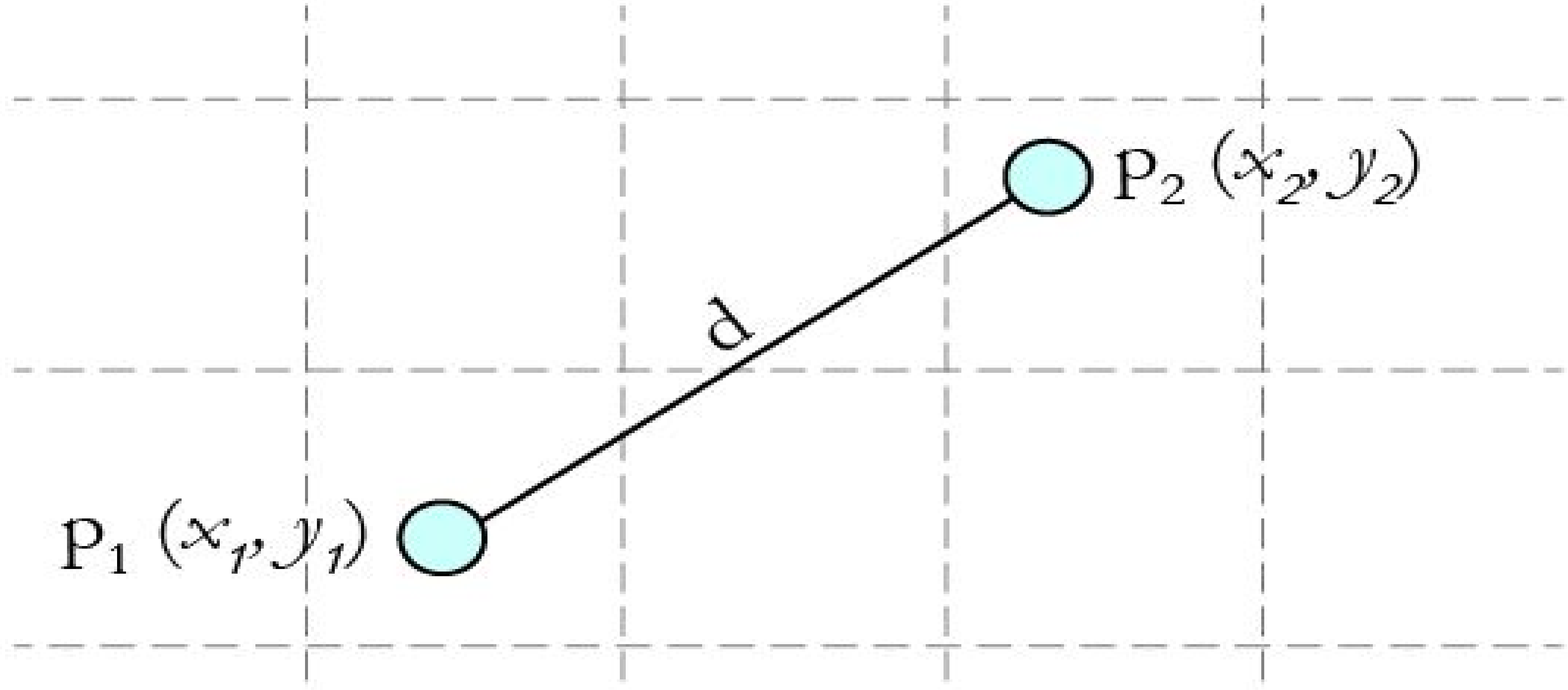
Basic Features

- Supervised learning classification AND regression technique
- Non-parametric
- “Lazy”--uses all/nearly all of the training data
- Feature similarity
- Groups
- Uses Euclidean Distance for calculating distance between centroid and *k* points

Application

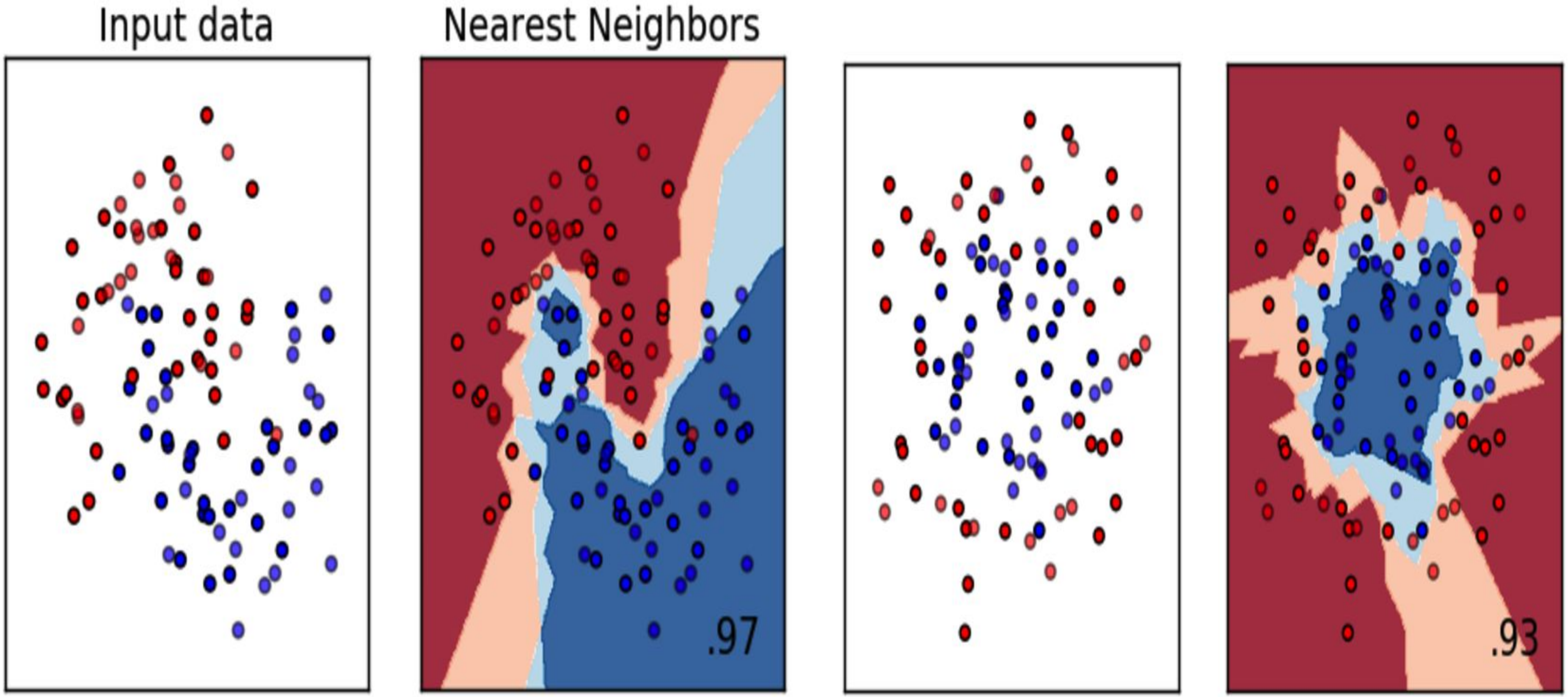
- “N-neighbors” = “k”, the number of data points we set our centroid to group with





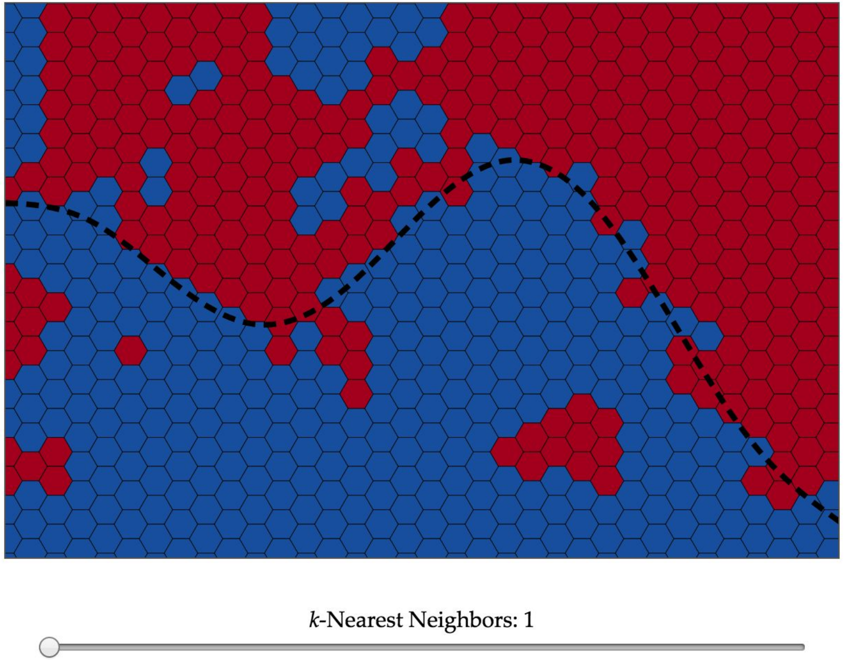
Euclidean distance (d) $= \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$

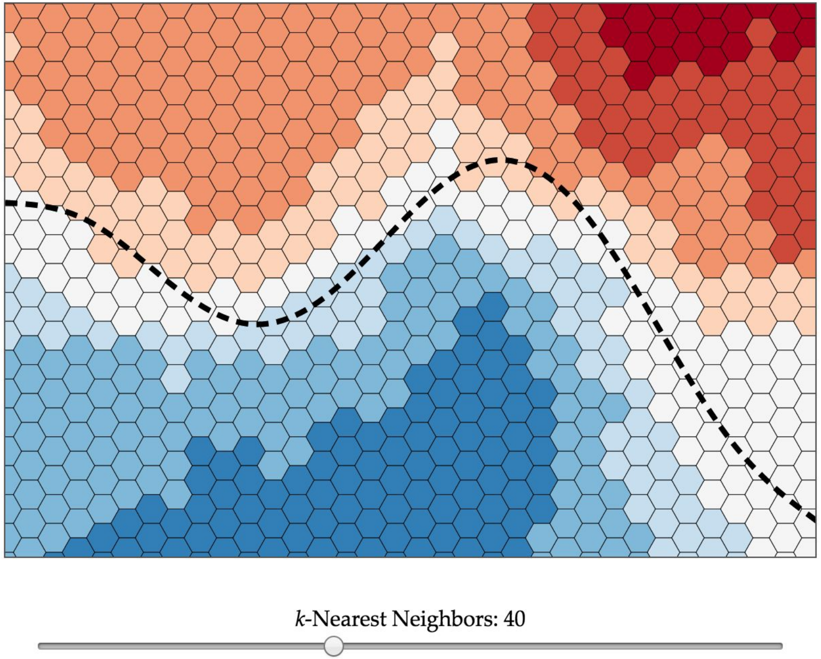
Examples from scikit-learn

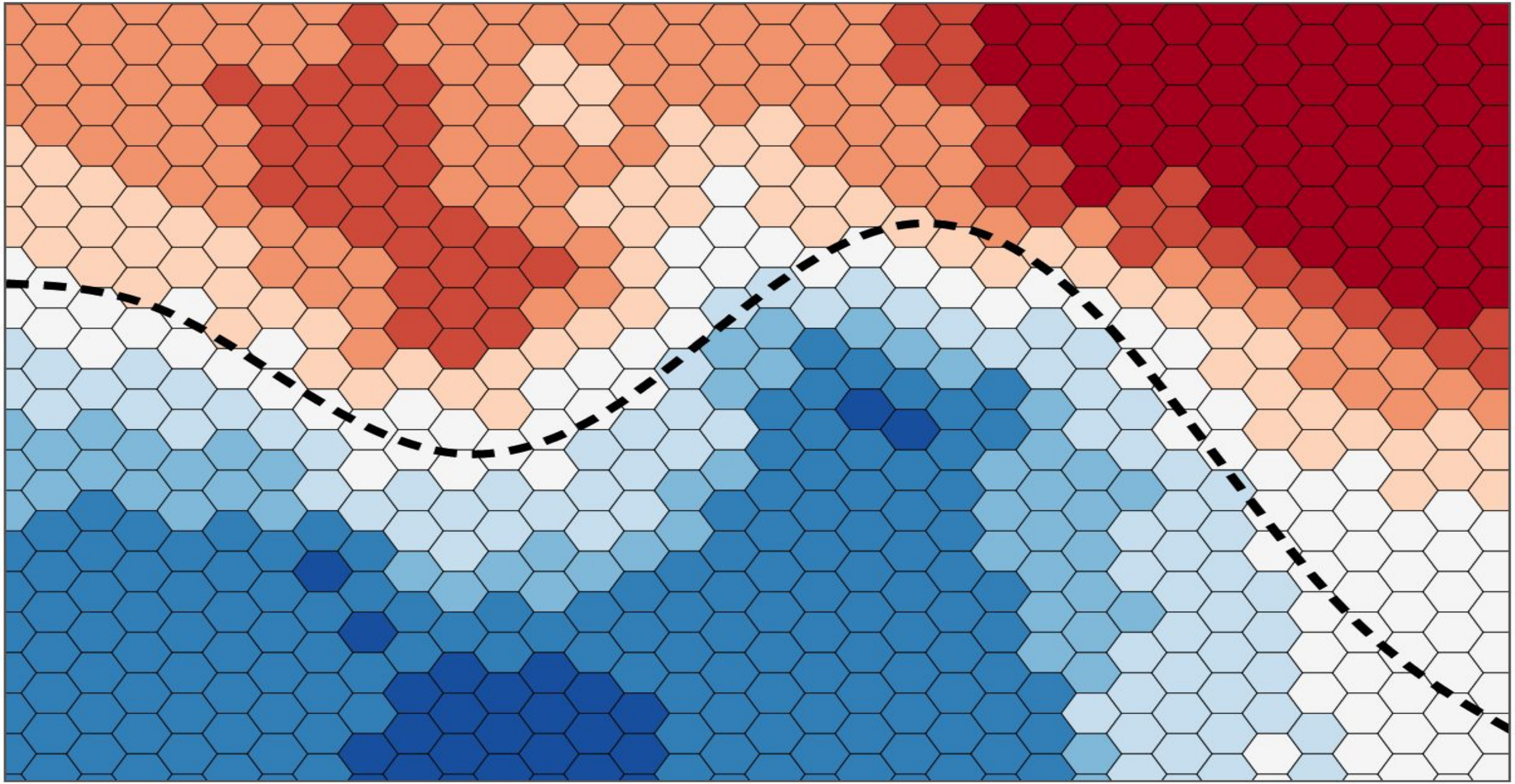


Overfitting vs. Underfitting

- Low k vs high k







k-Nearest Neighbors: 25

Use Case: Predicting Landslides in Greece

- Predicted landslide susceptibility in the basin of the Selinounda River in Achaia County, Greece with KNN
- Study area:

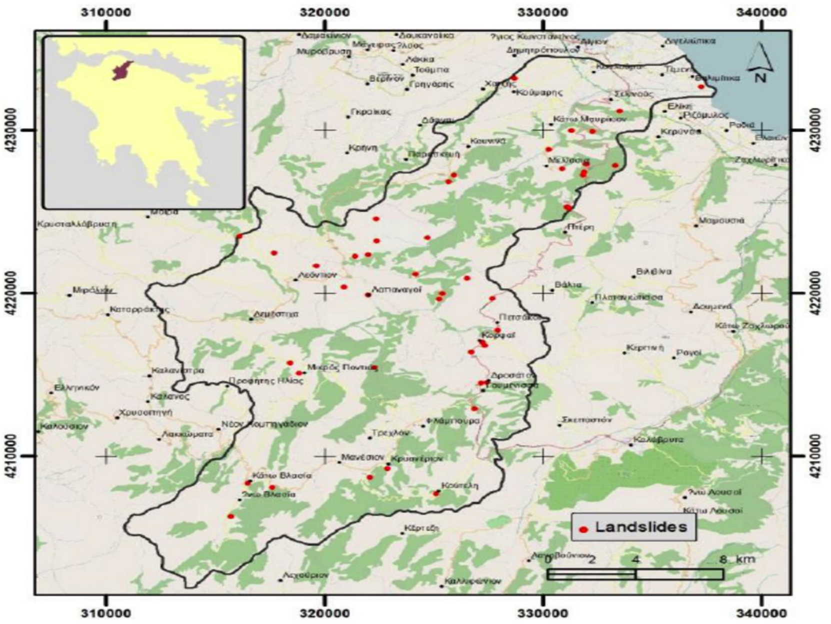


Figure 1 Study area

Landslide Variables

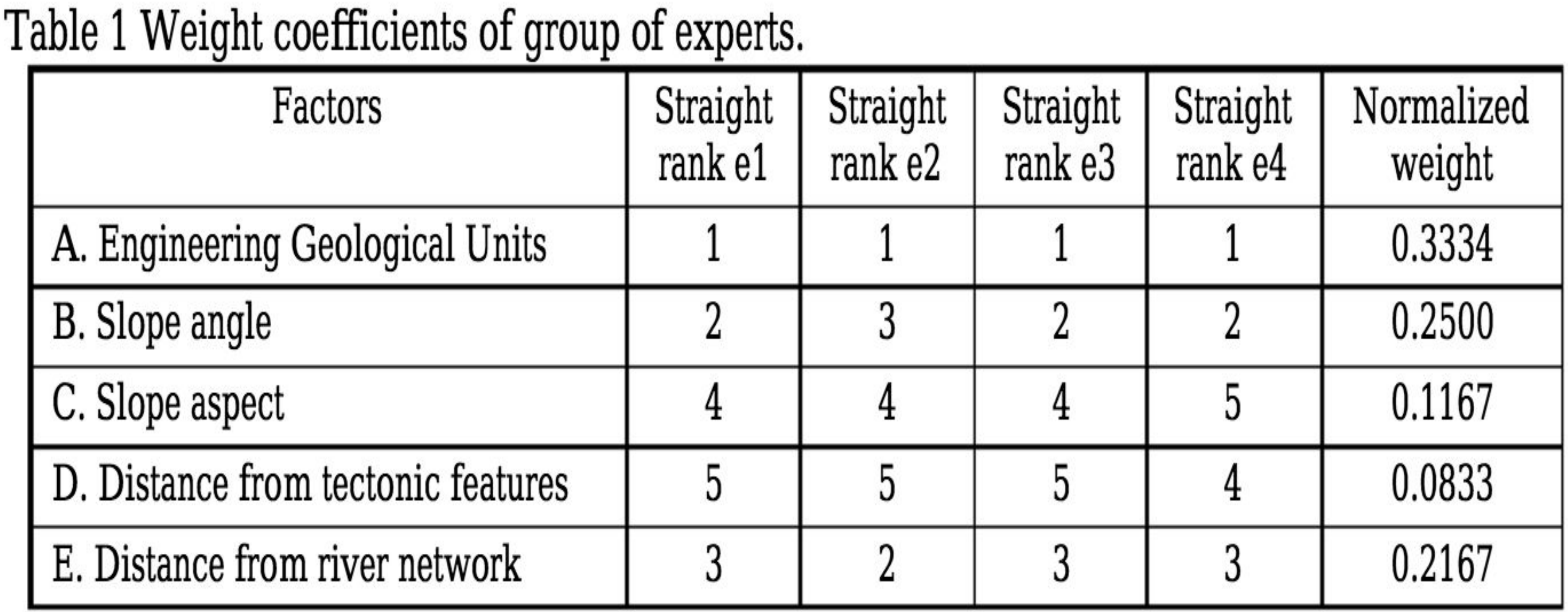


Table 1 Weight coefficients of group of experts.

Factors	Straight rank e1	Straight rank e2	Straight rank e3	Straight rank e4	Normalized weight
A. Engineering Geological Units	1	1	1	1	0.3334
B. Slope angle	2	3	2	2	0.2500
C. Slope aspect	4	4	4	5	0.1167
D. Distance from tectonic features	5	5	5	4	0.0833
E. Distance from river network	3	2	3	3	0.2167

Use of KNN to predict landslides

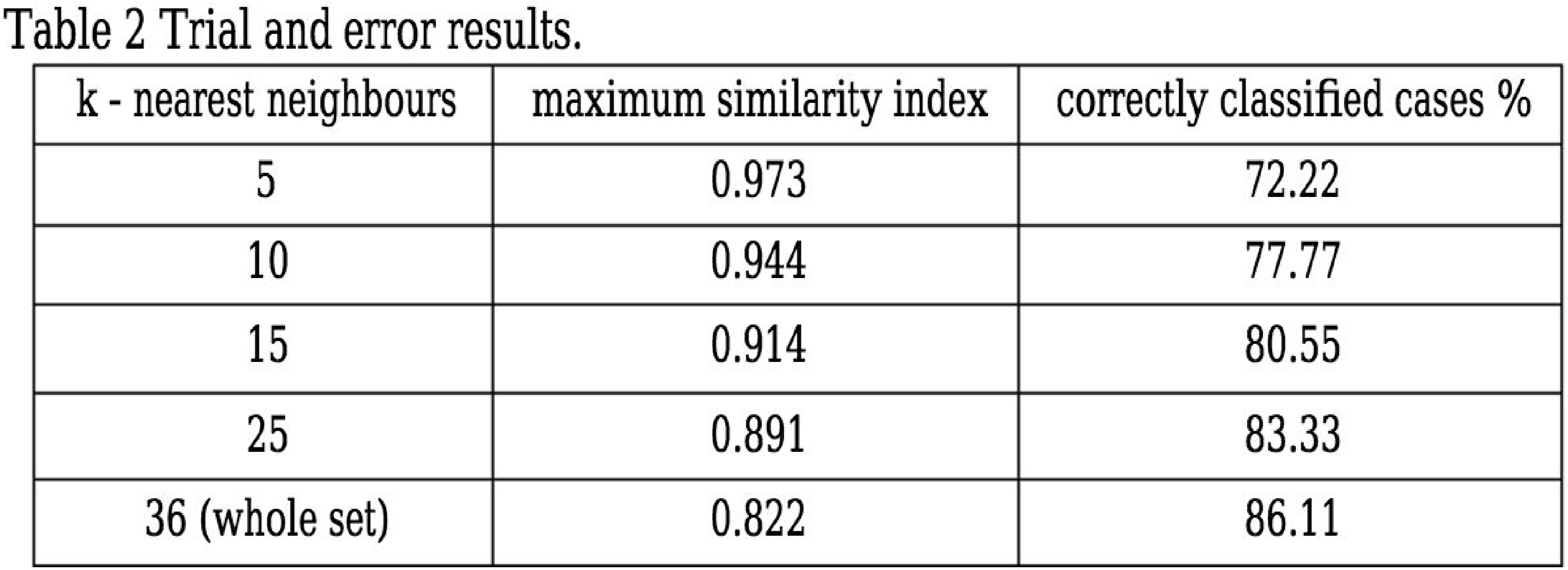


Table 2 Trial and error results.

k - nearest neighbours	maximum similarity index	correctly classified cases %
5	0.973	72.22
10	0.944	77.77
15	0.914	80.55
25	0.891	83.33
36 (whole set)	0.822	86.11

Results mapped and ROC Curve

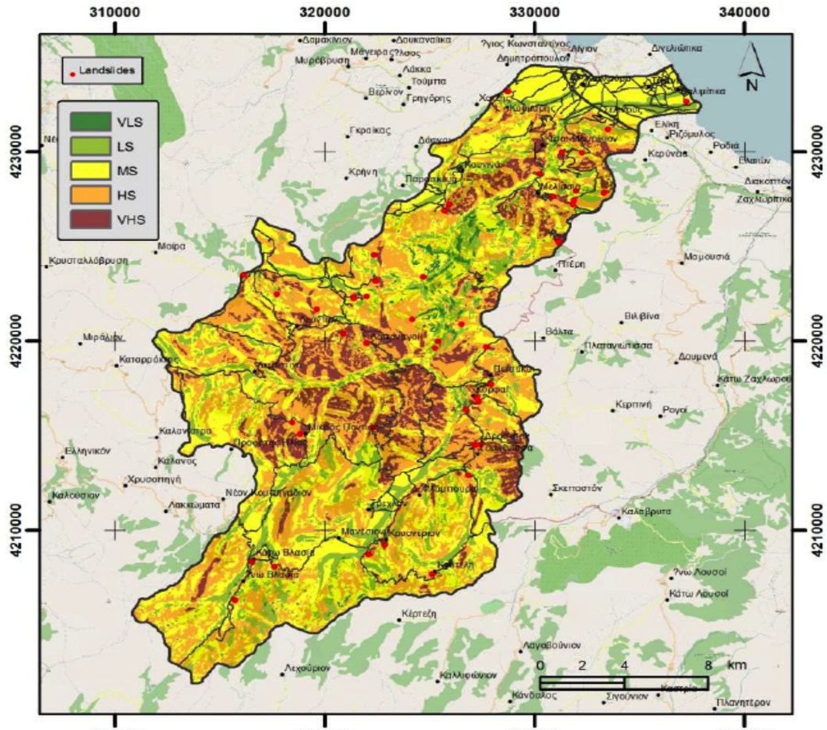


Figure 2 Landslide susceptibility map

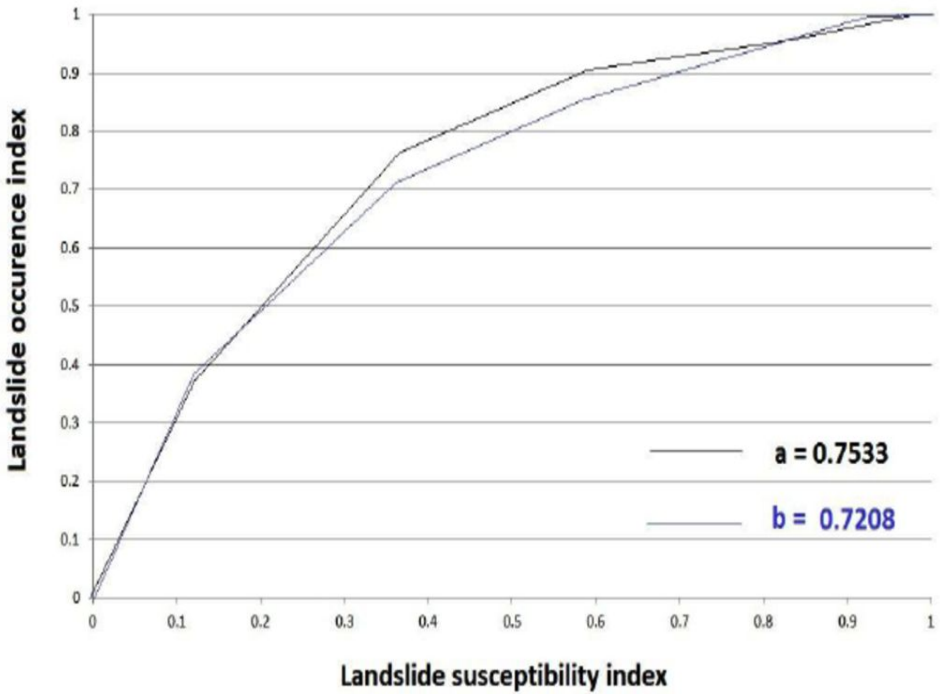


Figure 3 ROC curve: a success rate, b: prediction rate

Why do we use it?

Intuitive

Straightforward

Powerful

Fast