Teaching Core Principles of Machine Learning with **a Simple Machine Learning Algorithm**: The Case of the **KNN** Algorithm

Orit Hazzan & Koby Mike

TECHNION
Israel Institute of Technology

Introduction

- Data science is a new interdisciplinary science; data science education is even newer.
  - Data science pedagogy is currently shaped.
- Introductory data science courses usually include several machine learning algorithms of different kinds.
  - Our message:
    - Only one simple algorithm may be sufficient for introductory data science courses
    - Illustrating our approach with the KNN algorithm.

The K-Nearest Neighbors (KNN) algorithm

- To classify a new object *x*:

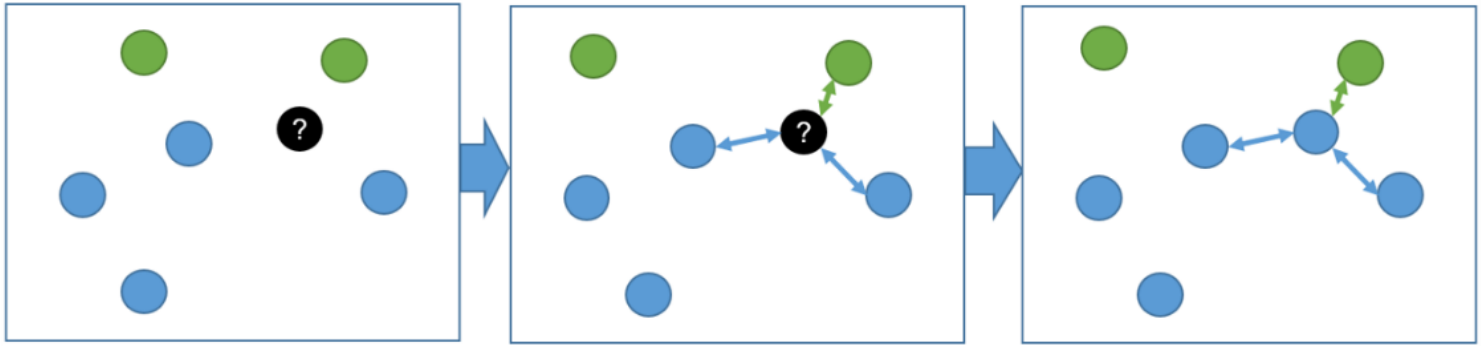
  1: Calculate its Euclidean distance from all examples in the training set.

  2: Locate the K nearest neighbors.

  3: For the K nearest neighbors, check the frequency of each category.

  4: Predict the x' category based on the category of the majority of nearest neighbors.

Graphic illustration of the KNN classification process for K=3



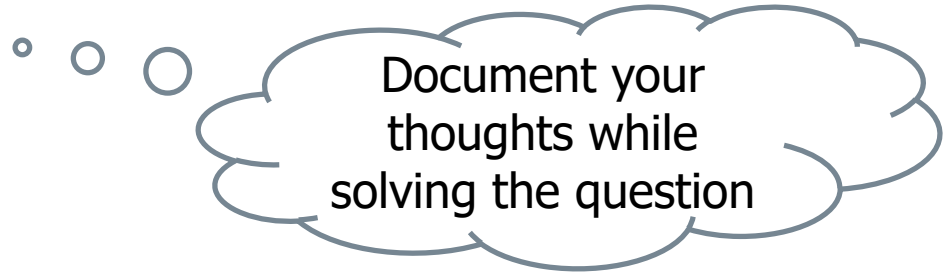
Main Message

- Only one simple algorithm may be sufficient for introductory data science courses
- Illustrating our approach with the KNN algorithm.

- Explanation from three perspectives:
  - **Algorithmic**
    - Simple for calculation and learning
    - Despite its simplicity, many core machine learning concepts can be explained intuitively using the KNN algorithm
  - **Cognitive**
    - Gradual mental construction
      - Process-object duality
  - **Pedagogical**
    - Implementation in the basic level of the data science unit of the Israeli high school computer science curriculum
    - Questions by Bloom's taxonomy
    - Elimination of barriers, which new teachers may encounter, to learning and teaching data science.

Illustrative Question

- In order to classify dogs as Poodles or Labradors, four characteristics were selected: height, weight, tail length, and ear length.
  The training set included 1,000 dogs, 500 of each kind.
  Based on this data set, we wish to classify an unknown dog using the KNN classifier.
  a. For K=5: How many times is the square operation executed?
  b. For K=11: How many times is the square operation executed?
  c. What conclusion can you draw from your answers to the above two questions?
  d. In your opinion, when are the chances of a correct classification higher?
     I. K=5
     II. K=11
     III. It is impossible to decide
     IV. I do not know
  e. Please explain your answer.



The KNN-algorithm from three perspectives

- Algorithmic
- **Cognitive**
- Pedagogical

- The algorithmic and pedagogical perspectives are discussed in:
  - Hazzan, O. and Mike, K. (2022). Teaching core principles of machine learning with a simple machine learning algorithm: The case of the KNN algorithm in a high school introduction to data science course, *ACM Inroads* **13**(1), pp. 18-25. https://dl.acm.org/doi/10.1145/3514217
- Additional details will be published:
  - Mike, K. and Hazzan, O. (2022, in press). Teaching machine learning: A white box approach, *Statistical Education Research Journal* (SERJ).

Algorithmic perspective: Simple algorithm

- The KNN algorithm is simple to learn and calculate.
  - It can be understood intuitively (like **real new neighbors**).
  - The KNN algorithm **does not involve the creation of a model** based on which new objects are classified.
    - It eliminates the need to understand both the meaning of the model as well as its construction process.
    - Instead of a disadvantage, the KNN can be viewed as a bridge to more advanced data science courses.
  - From a mathematical perspective, the KNN algorithm:
    - requires only **the calculation of the distance function**
    - **does not require an understanding of advanced concepts** such as vector dot product, partial derivatives, or function extrema.

The KNN-algorithm from three perspectives

- **Algorithmic**
  - Simple to calculate and learn
  - Despite its simplicity, many core machine learning concepts can be explained intuitively using the KNN algorithm
    - Hyperparameter
    - Classification
    - Training set
    - Test set
    - Model performance metrics
    - Underfit and overfit
    - the process of a data science project

Illustrations:

Hazzan, O. and Mike, K. (2022). Teaching core principles of machine learning with a simple machine learning algorithm: The case of the KNN algorithm in a high school introduction to data science course, *ACM Inroads* **13**(1), pp. 18-25. https://dl.acm.org/doi/10.1145/3514217

The KNN-algorithm from three perspectives

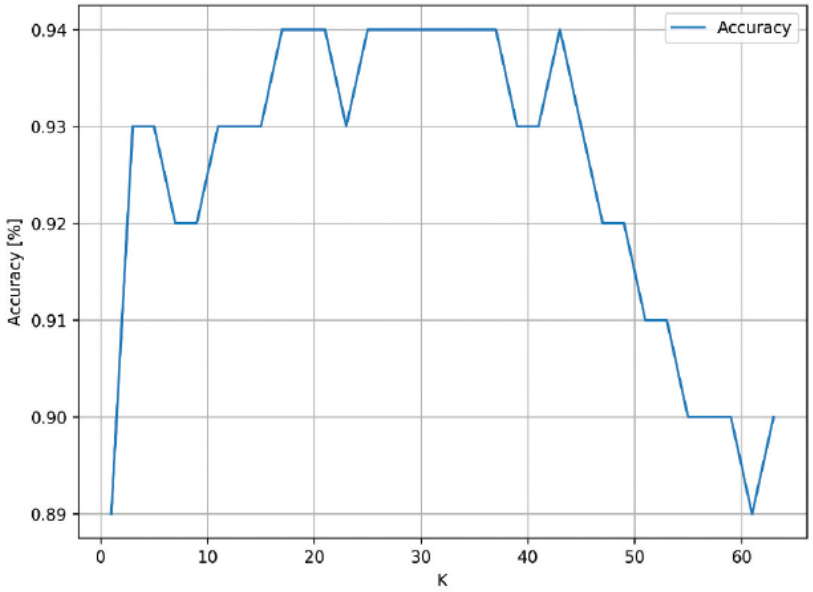


Figure 3: KNN hyperparameter tuning: Accuracy vs. K

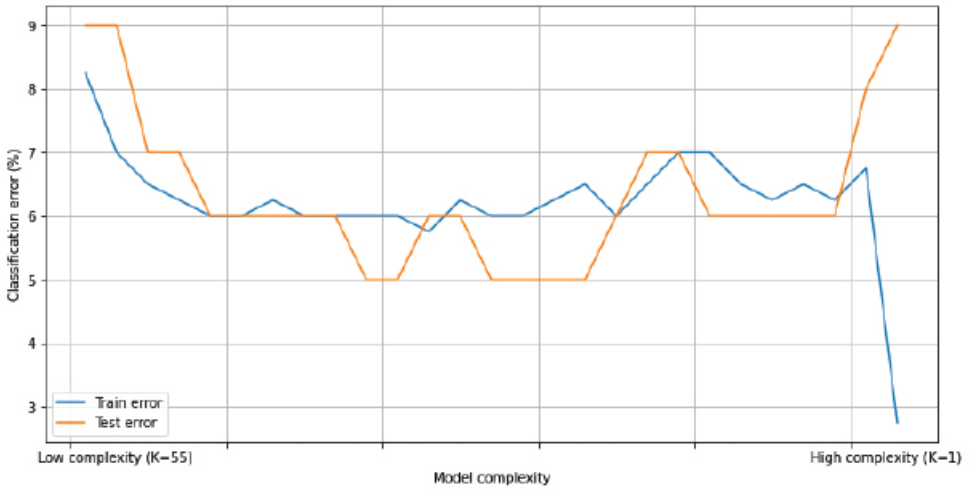


Figure 4: The underfit and overfit phenomena in KNN

The KNN-algorithm from three perspectives

- **Cognitive**
  - Gradual mental construction
    - Process-object duality

Process-object conception of mathematical concepts

- Abstract mathematical concepts can be represented in the human mind as **either objects or processes** (Sfard, 1991).
  - As an **object**, an abstract mathematical concept is conceived of as a fixed construct,
  - As a **process**, an abstract mathematical concept is conceived of as an algorithm or a computation that generates an output from an input.

  Examples:
  - Derivative
  - A proof by induction

Process-object duality

- In the learning processes of most mathematical concepts, the learner passes through three phases.
  - First, the concept is conceived of as a **process**.
  - Then, the process is mentally repacked (encapsulated) and an **object** representation is created in the learner's mind.
  - In the final step, the concept can be used as an element of a more complex process.

Understanding a mathematical concept as a process is an essential step toward understanding it as an object.

Object-process duality: The KNN algorithm

Q1. How would you explain to a friend what the KNN algorithm is?

- Discussion:
  - The question formulation
  - What can we learn from students' answers about their conception of the KNN algorithm?

White-box understanding and the process-object duality

White box understanding:

- is the understanding of how something works
- is important for improving algorithm performance (by hyper parameter tuning)
- requires both a process conception of the algorithm and an object conception of its properties (behavior of the parameter)

Gradual construction based on the process-object duality

- Three steps:
  - Visual
  - Process
    - Ritual
      - the execution of a mathematical routine on a lower level of thinking
      - a simple repetition of the mathematical procedure
    - Exploration
      - the execution of a mathematical routine on a higher level of thinking
      - constructing the mathematical concept as an object.
  - Object

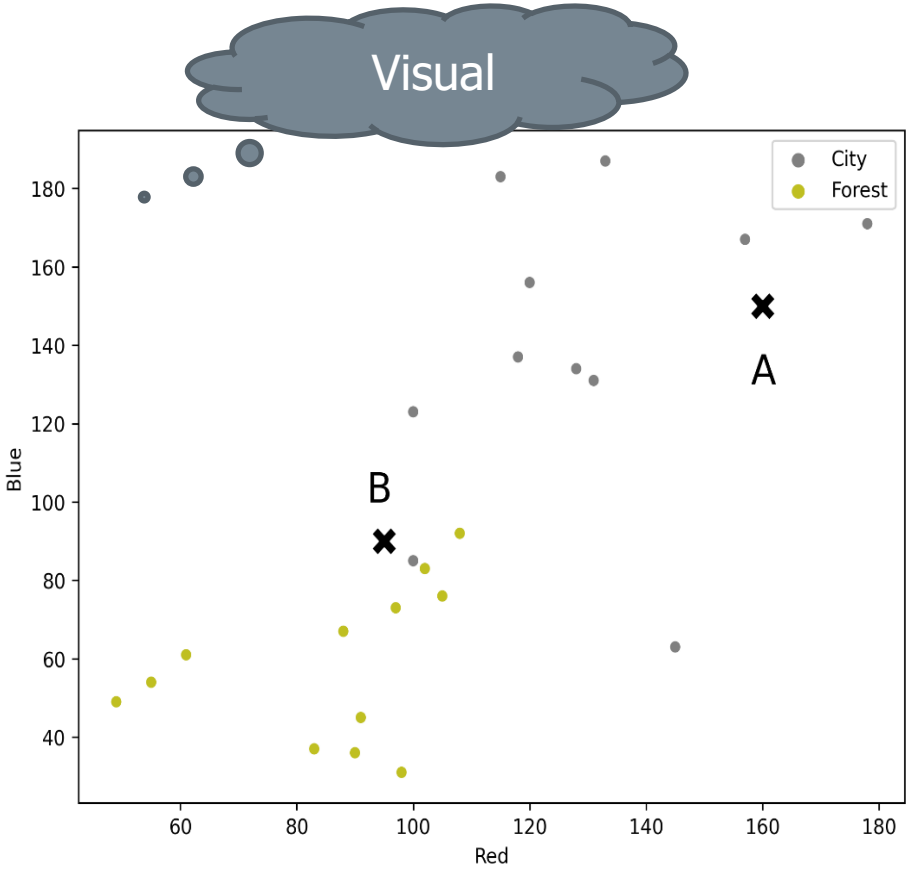
Sources: Sfard & Lavie, 2005; Heyd-Metzuyanim & Graven, 2019; Lavie et al., 2019.

Part 1: Image classification (worksheet)

A KNN algorithm is designed to classify images into two types: urban or forest, based on two features: Red - the mean level of the red color in the image, and Blue - the mean level of the blue color in the image.

The following graph presents the training dataset images (urban images in gray and forest images in green). The graph also shows two unknown images, A and B.

Classify images A and B using a KNN algorithm, with K=1 and K=3.



1.For K=1:

a. The classification of image A is:

Explain your answer:

b. The classification of image B is:

Explain your answer:

2. For K=3:

a. The classification of image A is:

Explain your answer:

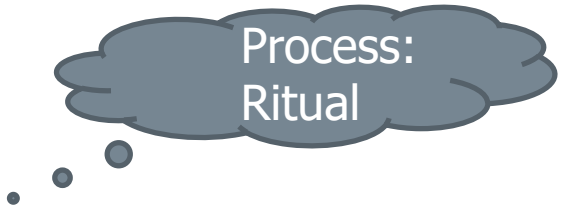
b. The classification of image B is:

Explain your answer:

Part 2: Iris classification (worksheet)

A KNN algorithm is designed to classify Iris flowers into two classes: setosa and versicolor.
Four features are given for each flower: sepal length, sepal width, petal length, and petal width.
There are six samples of flowers in the training set.
A researcher found a new flower, U, with the following features: U_{SL}=5, U_{SW}=3, U_{PL}=2, U_{PW}=2.
Calculate the distance of the new flower from each sample in the training set.
2. Classify this new flower using the KNN algorithm with K=1 and with K=3.
For K=1:

a. The indexes of the K closest training examples:
b. The labels of the K closest training examples:
c. The final classification is:



For K=3:

a. The indexes of the K closest training examples:
b. The labels of the K closest training examples:
c. The final classification is:

Sample	sepal length (SL)	sepal width (SW)	petal length (PL)	petal length (PW)	Label	$d^{(i)}$
1	5.1	3.5	1.4	0.2	Setosa	
2	4.9	3	1.4	0.2	Setosa	
3	4.7	3.2	1.3	0.2	Setosa	
4	7	3.2	4.7	1.4	Versicolor	
5	6.4	3.2	4.5	1.5	Versicolor	
6	6.9	3.1	4.9	1.5	Versicolor	

KNN Questionnaire

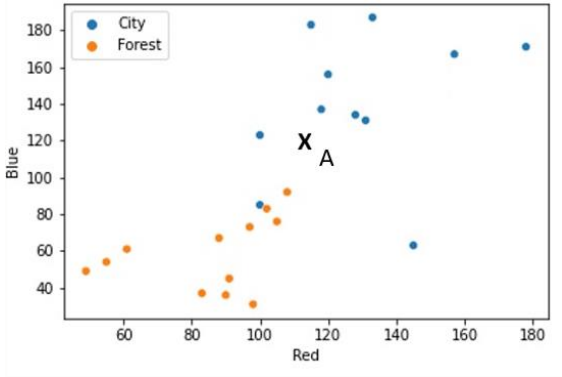
Q2. Students were asked to classify Example A in the figure below, using the KNN algorithm for K=5.

Alice claims that A's classification is Forest.

a. Is Alice right?
b. In your opinion, how did Alice explain her answer?

Bob claims that A's classification is City.

a. Is Bob right?
b. In your opinion, how did Bob explain his answer?



Q3. Students were asked to classify Example B in the figure below, using the KNN algorithm.

Carol claims that for K=5, B's classification is City.

a. Is Carol right?
b. In your opinion, how did Carol explain her answer?

Dave claims that for K=5, B's classification is Forest.

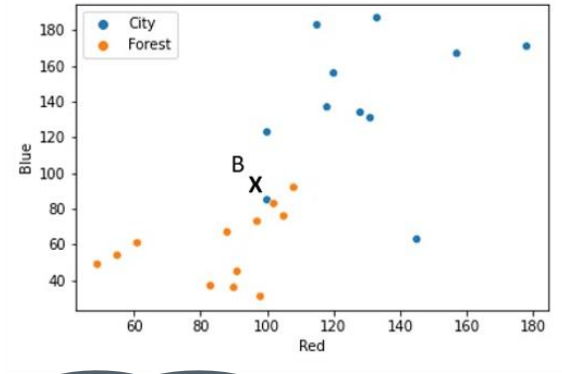
c. Is Dave right?
d. In your opinion, how did Dave explain his answer?

Eve claims that for any K, B's classification is City.

e. Is Eve right?
f. In your opinion, how did Eve explain her answer?

Frank claims that for any K, B's classification is Forest.

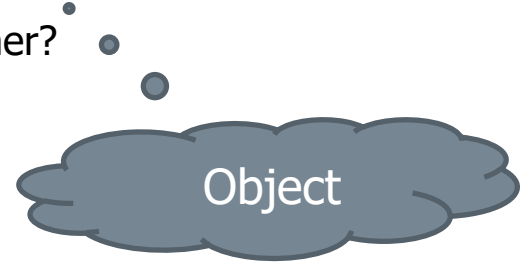
g. Is Frank right?
h. In your opinion, how did Frank explain his answer?



Process: Exploration

Q4. In order to classify dogs as Poodles or Labradors, four characteristics were selected: height, weight, tail length, and ear length. The training set included 1,000 dogs, 500 of each kind. Based on this data set, we wish to classify an unknown dog using the KNN classifier.

- a. For K=5: How many times is the square operation executed?
- b. For K=11: How many times is the square operation executed?
- c. What conclusion can you draw from your answers to the above two questions?
- d. In your opinion, when are the chances of a correct classification higher?
  I. K=5
  II. K=11
  III. It is impossible to decide
  IV. I do not know
- e. Please explain your answer.



Object-process duality: The KNN algorithm

Students' conception of the KNN algorithm as a process / object can be categorized by their examination of the following properties of the KNN algorithm

- **Process conception:**
  - (P1) Calculate distance from all samples
  - (P2) Pick the K nearest samples
  - (P3) Find the label of the majority
  - (P4) Tune the hyperparameter K to improve performance (to avoid underfitting and overfitting)

- **Object conception:** O1-O3 address the classification, O4 addresses the algorithm's performance and the O5-O7 address the algorithm's complexity (determined by the number of distance calculations):
  - (O1) Classification depends on similarity
  - (O2) Classification is determined by distance
  - (O3) The classification of a specific unknown example depends on K
  - (O4) The performance of the KNN algorithm for a specific K depends on the distribution of the data
  - (O5) The number of distance calculations depends on the number of training samples
  - (O6) The number of distance calculations depends on the number of features
  - (O7) The number of distance calculations does not depend on K

Mapping questions of the KNN algorithm according to the KNN process steps and object properties each question elicits

		Question			
Type of conception	Expression of conception	1	2	3	4
Process conception	(P1) Calculate distance from all samples	X			X
	(P2) Pick the K nearest samples	X	X	X	
	(P3) Find the label of the majority	X	X	X	
	(P4) Tune the hyper parameter K to improve performance				X
Object conception	(O1) Classification depends on similarity	X			
	(O2) Classification is determined by distance	X			
	(O3) The classification of a specific unknown example depends on K	X		X	
	(O4) The performance of the KNN algorithm for a specific K depends on the distribution of the data				X
	(O5) The number of distance calculations depends on the number of training samples				X
	(O6) The number of distance calculations depends on number of features				X
	(O7) The number of distance calculations does not depend on K				X

Coded answers

			Student number											
Type of conception	Expression of conception	Total	1	2	3	4	5	6	7	8	9	10	11	12
Process conception	(P1) Calculate distance from all samples	6												
	(P2) Pick the K nearest samples	12												
	(P3) Find the label of the majority	12												
	(P4) Tune the hyperparameter K to improve performance	3												
Object conception	(O1) Classification depends on similarity	4												
	(O2) Classification is determined by distance	11												
	(O3) The classification of a specific unknown example depends on K	4												
	(O4) The performance of the KNN algorithm for specific K depends on the distribution of the data	6												
	(O5) The number of distance calculations depends on the number of training samples	6												
	(O6) The number of distance calculations depends on the number of features	6												
	(O7) The number of distance calculations does not depend on K	6												

Process-object duality and white-box understanding

- **Cognitive perspective at the KNN algorithm**
  - Gradual mental construction
    - The process-object duality guides
      - gradual understanding – first as a process and then ad an object
      - teaching white box understanding

The KNN-algorithm from three perspectives

- **Pedagogical**
  - Implementation in the basic level of the data science unit of the Israeli high school computer science curriculum.
  - Illustrative questions by (the revised) Bloom's taxonomy:
    - Remember, Understand, Apply, Analyze, Evaluate, and Create
  - Elimination of learning and teaching data science barriers, which new teachers may encounter.

Table 2: Updated data science for high school curriculum with only the KNN algorithm – Topics and number of hours

Topic	Hours	Comments
Introduction to data science	3	Data science application in real life
Python	18	Basic data types, lists, loops and conditions in Python
Introduction to machine learning	3	Types of machine learning
Tabular data and pandas	6	Import csv, select rows and columns
Exploratory data analysis	3	Visualization with matplotlib and seaborn
The K-nearest neighbors (KNN) algorithm	6	The KNN algorithm and implementation in Python
Image as data	3	Read images and classify based on dominant colors
Evaluating classifiers	3	Confusion matrix and accuracy, imbalanced classes
Core concepts in machine learning	3	Hyperparameters, overfit and underfit
Data collection and curation	3	Find errors in data and missing values, normalization
Managing data project	3	The data processing cycle
Principles of modern machine learning algorithms	6	Introduction to neural networks
Final project development	30	Real-life project based on pupils' choice
TOTAL	**90**	

Examples of tasks involving the KNN algorithm, sorted according to Bloom's revised taxonomy.

4. ANALYZE

The following question reflects the analysis level of Bloom's revised taxonomy since it requires learners to analyze the fit between a classifier and a given labeled dataset by examining the classifier's performance metrics.

For a given data set, build a series of KNN classifiers with K=1, 2, 3, ... 11.

a. Calculate the performance metrics of each of the classifiers.
b. Which classifier is the best for the given data set?

6. CREATE

For this level of understanding, we propose a "Give an example" task (see, e.g., [11]). Here is an example.

For each of the following conditions, design a data set containing three kinds of objects, with two features each:

(a) For K=3, the KNN algorithm classifies new examples correctly;
For K=5, the KNN algorithm classifies new examples incorrectly.
(b) The recall is always higher than 0.9.
(c) The KNN algorithm classifies only one new example incorrectly.

Do your answers depend on the number of kinds of objects and/or on the number of features of each example in the data set? Explain your answer.

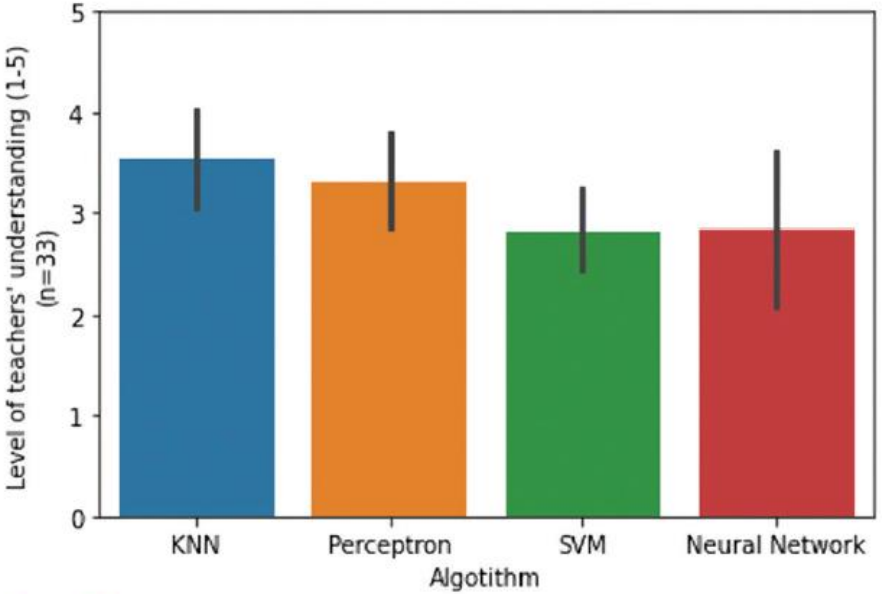


Figure 5A: Computer science teachers' perception of their understanding of the four algorithms

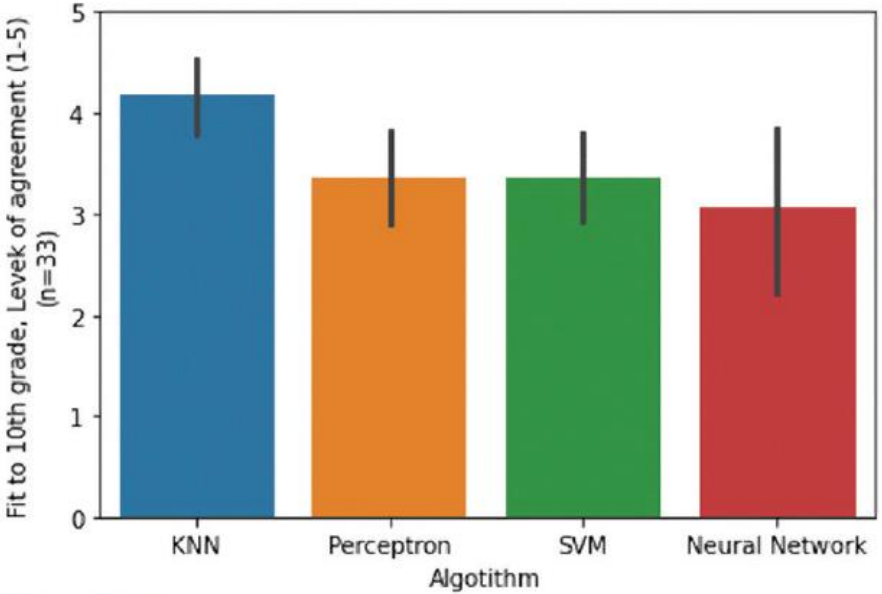


Figure 5B: Computer science teachers' perception of the suitable algorithm for teaching in the 10th grade

The KNN-algorithm from three perspectives

- Algorithmic
  - Simple for calculation
  - Despite its simplicity, the KNN algorithm enables to expose novice data science learners to main ideas of machine learning
- Cognitive
  - Process-object duality
- Pedagogical
  - Implementation in the basic level of the data science unit of the Israeli high school computer science curriculum
  - A variety of questions on different level of complexity
  - Elimination of learning and teaching data science barriers, which new teachers may encounter

Discussion

- Suggest other algorithms that may serve as the sole algorithm taught in introductory data science courses.