ALGONQUIN COLLEGE

CST8502
MACHINE LEARNING

Week 2

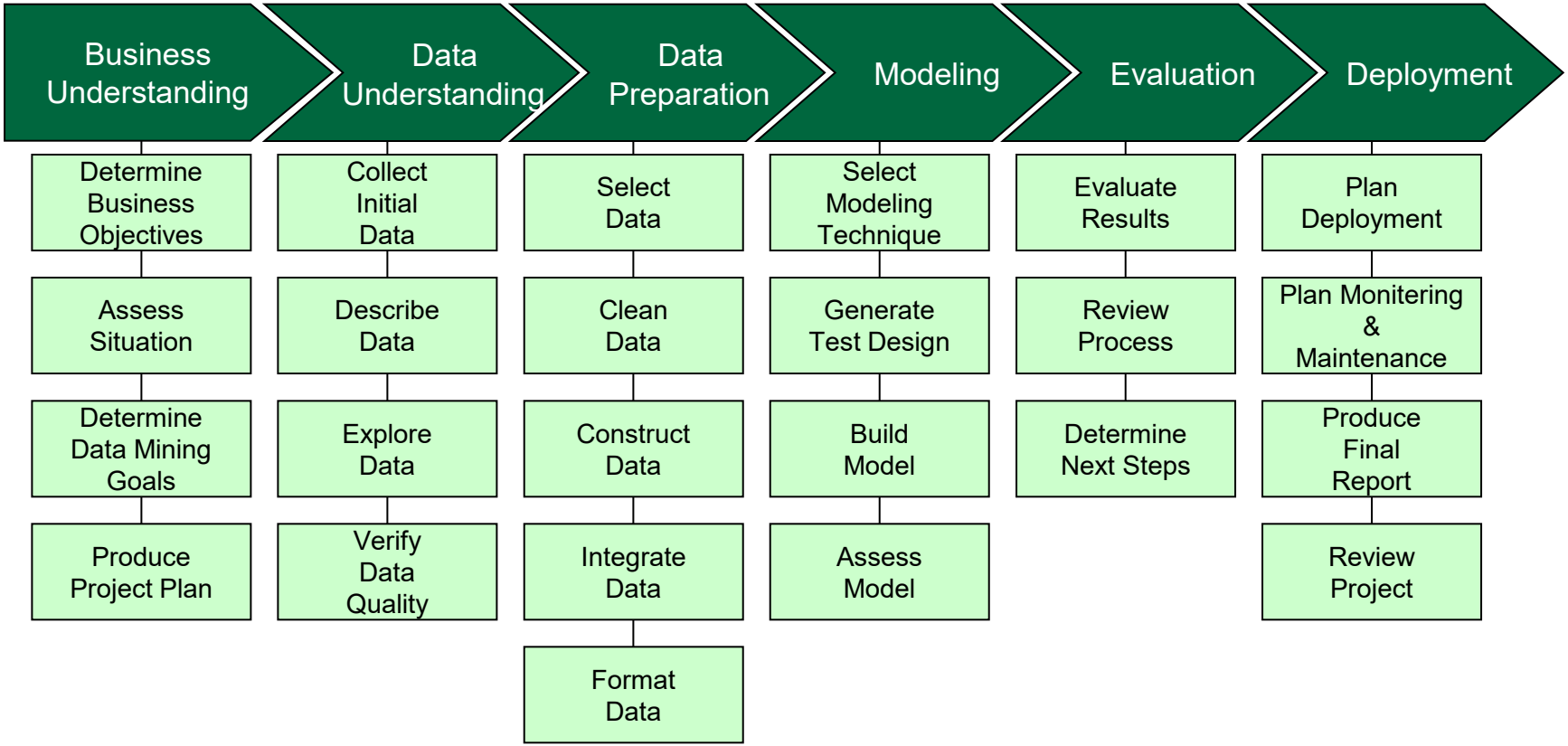
Data Understanding and Preparation

Classification by kNN

Professor: Dr. Anu Thomas

CRISP-DM

- Cross-industry Standard Process for Data Mining
  - Business Understanding
  - Data Understanding
  - Data Preparation
  - Modeling
  - Evaluation
  - Deployment


Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Determine Business Objectives
Assess Situation
Determine Data Mining Goals
Produce Project Plan
Collect Initial Data
Describe Data
Explore Data
Verify Data Quality
Select Data
Clean Data
Construct Data
Integrate Data
Format Data
Select Modeling Technique
Generate Test Design
Build Model
Assess Model
Evaluate Results
Review Process
Determine Next Steps
Plan Deployment
Plan Monitering & Maintenance
Produce Final Report
Review Project

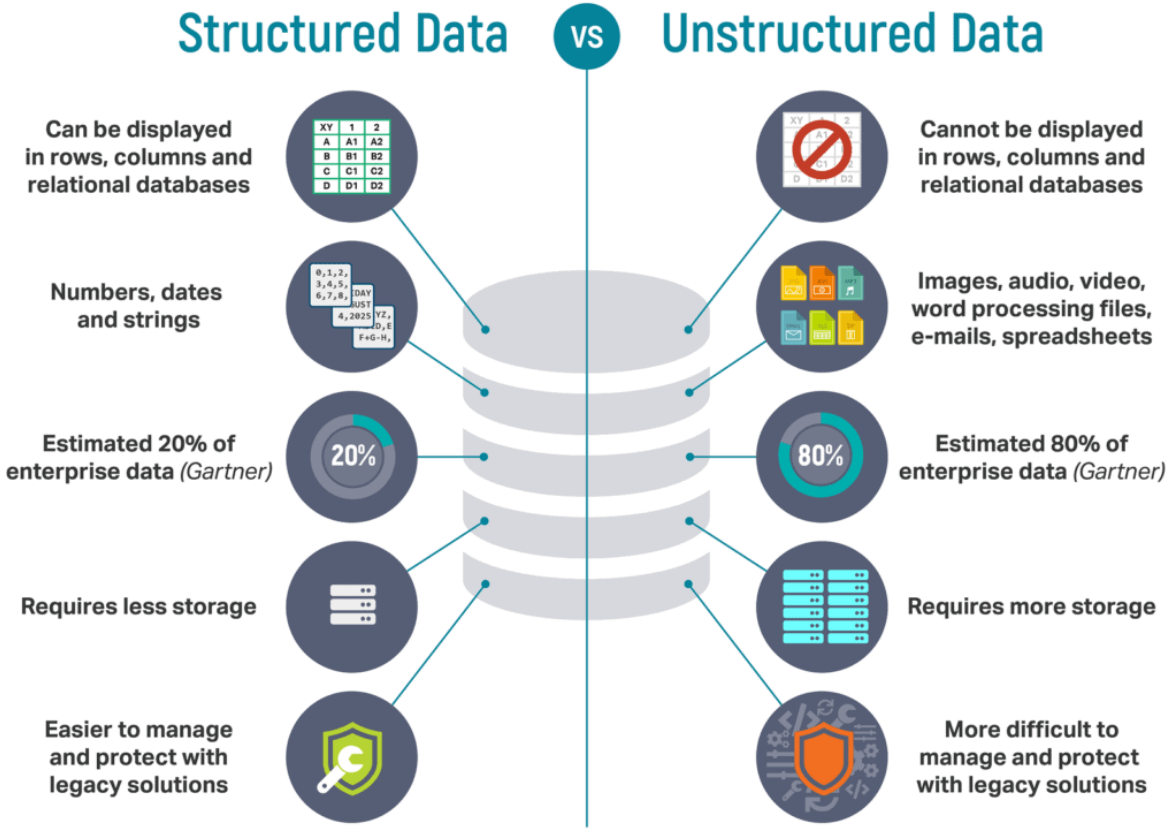
Machine Learning

- Supervised learning – classification, regression
- Unsupervised learning – clustering, outlier detection

Structure of Data

Typically, data fall into one of 3 categories:

- **Structured** – The data are highly structured, where every element has the same fields: age, first name, last name, address, etc. Databases, objects are good examples
- **Unstructured** – The data have no common structure. News articles, websites, video, audio and photographs.
- **Semi-structured** – The data use some structure like tags and keywords, but without a proper schema. This includes tree-type data like XML and JSON.

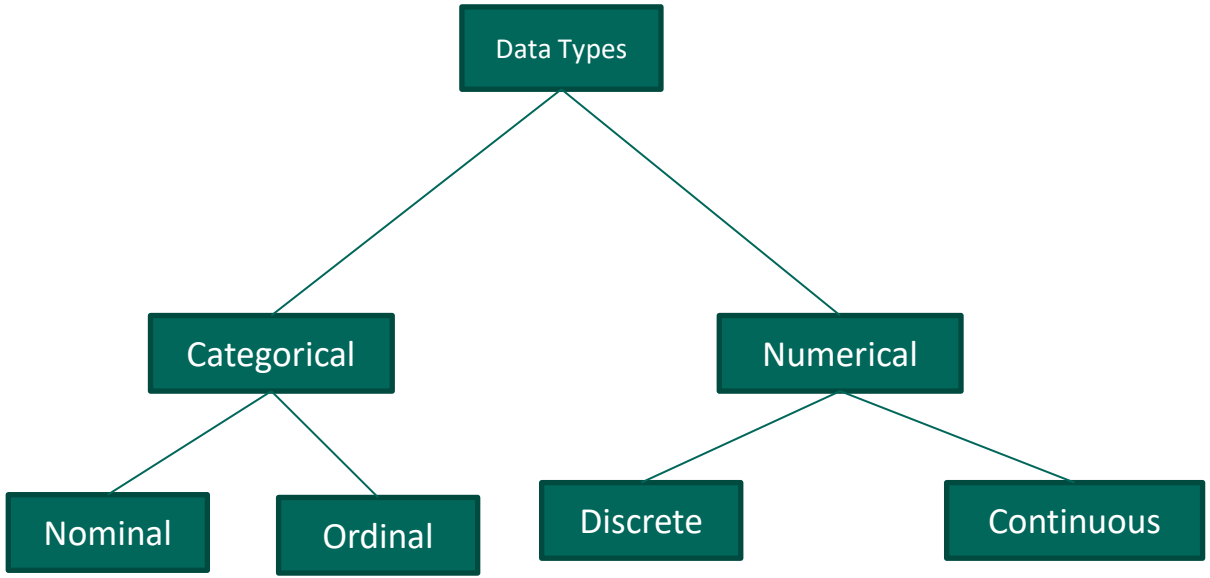


Taken from: https://k21academy.com/microsoft-azure/dp-900/structured-data-vs-unstructured-data-vs-semi-structured-data/

Semi-structured data

```
## Document 1 ##
{
  "customerID": "103248",
  "name":
  {
    "first": "AAA",
    "last": "BBB"
  },
  "address":
  {
    "street": "Main Street",
    "number": "101",
    "city": "Acity",
    "state": "NY"
  },
  "ccOnFile": "yes",
  "firstOrder": "02/28/2003"
}
```

Data Types



Nominal:
Gender: male, female

Ordinal:
survey questions: Strongly agree, agree, not sure, disagree, strongly disagree

Discrete:
if values are distinct and separate.
Cannot be measured but can be counted
Ex: number of heads in 100 coin flips

Continuous:
represents measurements. Values cannot be counted but can be measured
Ex: Height, salary

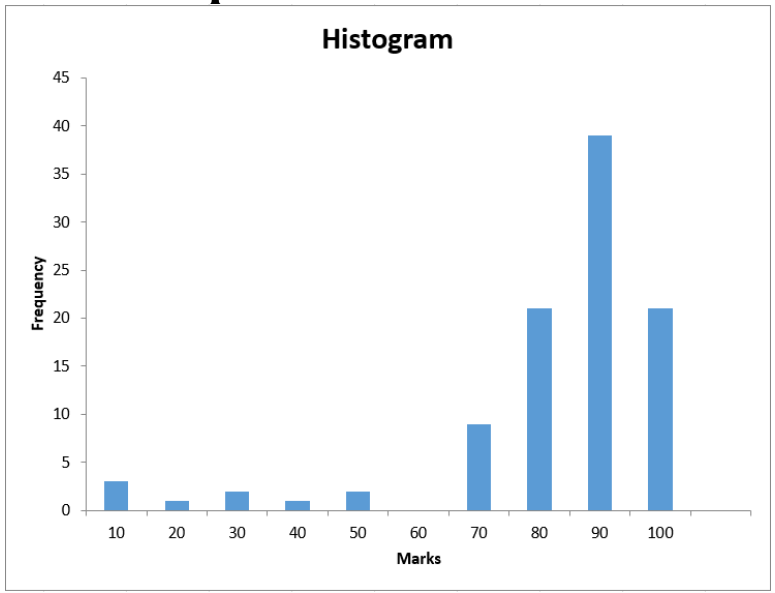
Data Understanding

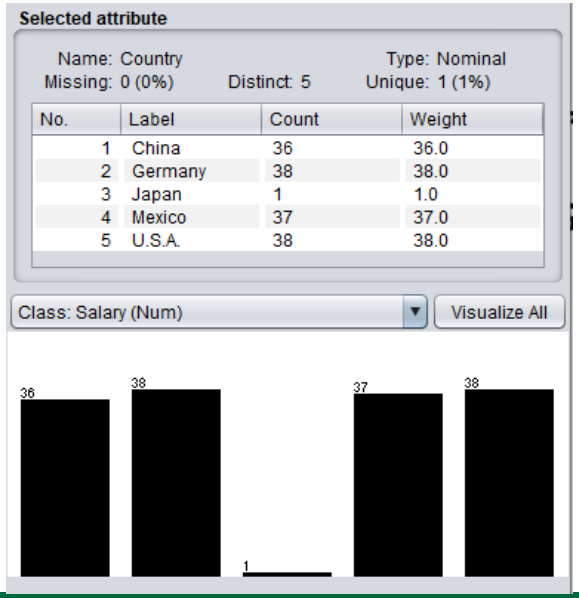
Getting to know the data

- For categorical features, we should:
  - Examine the mode, count etc. as these tell us the most common levels within these features and will identify if any levels dominate the dataset.
- For continuous features we should:
  - Examine the mean and standard deviation of each feature to get a sense of the central tendency and variation of the values within the dataset for the feature.
  - Examine the minimum and maximum values to understand the range that is possible for each feature.

Visualizations

- Histogram for each continuous feature
- Bar plot for each categorical feature

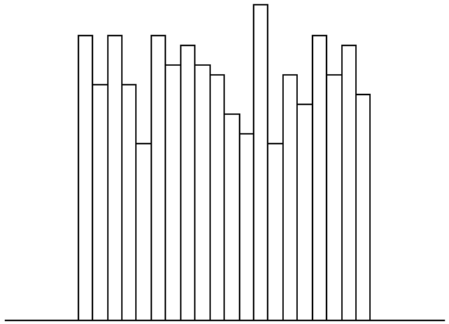


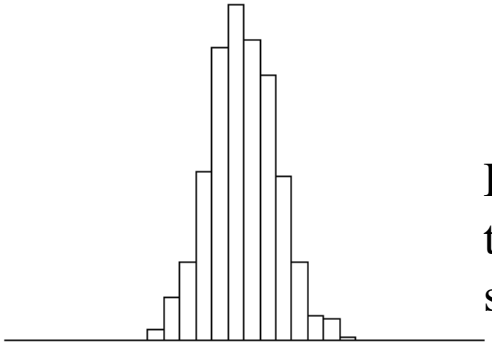


Histograms

Distribution may be uniform, normal, skewed, exponential, multimodal etc.

Uniform distribution indicates that a feature is equally likely to take a value in any of the ranges present

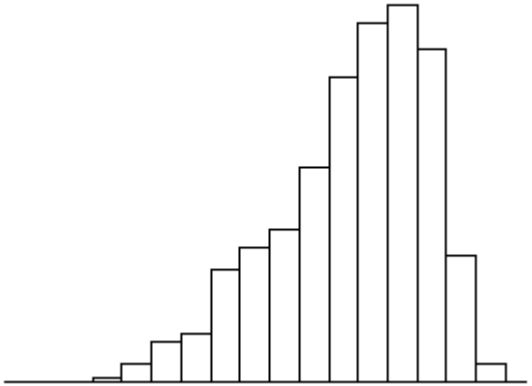




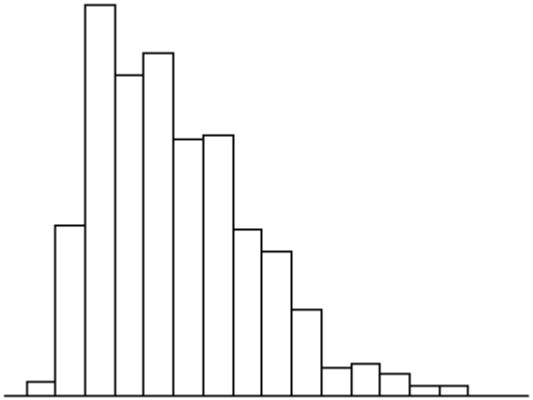
Features following a normal distribution are characterized by a strong tendency towards a central value and symmetrical variation to either side of this.

Histograms

Skew is a tendency towards very high (right skew) or very low (left skew) values



Unimodal (skewed left)



Unimodal (skewed right)

Data Quality Issues

- Missing values
- Cardinality
  - Incorrect cardinality: when the cardinality of a feature does not match to expected values (Assume you have a Gender column with 2 options – Male, Female. If you create 2 binary columns like isMale and isFemale and keep both, it is incorrect. You need only one of them)
  - Features having a cardinality=1 are not useful in analytics and needs to be removed
  - Large number of cardinality (>50) can create issues for some ML algorithms
- Outliers
- Invalid data
- Duplicate data

Data Preparation

Format Data

Change the way data it is represented just to make it more compatible with certain machine learning algorithms

- Normalization
- Standardization
- Binning
- Sampling

Normalization

Normalization: change a continuous feature to fall within 0 and 1 while maintaining the relative differences between the values for the feature.

$$a_i' = \frac{a_i - min(a)}{max(a) - min(a)}$$

Range normalization: change a continuous feature to fall within a specified range [low, high]while maintaining the relative differences between the values for the feature.

$$a_i' = \frac{a_i - min(a)}{max(a) - min(a)} \times (high - low) + low$$

Standardization

Standardize data into standard scores

- Rescales data to have a mean of 0 and standard deviation of 1

$$a_i' = \frac{a_i - \bar{a}}{sd(a)}$$

Normalized vs Range-normalized vs Standardized

	Height	Range Normalized	Normalized	Standardized
	192	1.50	0.50	-0.07
	197	1.68	0.68	0.53
	192	1.50	0.50	-0.07
	182	1.14	0.14	-1.28
	206	2.00	1.00	1.62
	192	1.50	0.50	-0.07
	190	1.43	0.43	-0.31
	178	1.00	0.00	-1.77
	196	1.64	0.64	0.41
	201	1.82	0.82	1.02
Max	206			
Min	178			
Mean	192.6			
StdDev	8.26236447			
Low	1			
High	2			

Binning

Binning involves converting a continuous feature into a categorical feature.

To perform binning, we define a series of ranges (called **bins**) for the continuous feature that correspond to the levels of the new categorical feature we are creating.

Popular ways of defining bins:

equal-width binning - splits the range of the feature values into b bins each of size $\frac{range}{b}$.

equal-frequency binning - first sorts the continuous feature values into ascending order and then places an equal number of instances into each bin, starting with bin 1. The number of instances placed in each bin is simply the total number of instances divided by the number of bins b.

Sampling

- If the dataset is so large, we can take a sample (a smaller percentage)
- Sample should still be a representative of original data (otherwise, ???)

Common Sampling Techniques

- **Top sampling** - selects the top s% of instances from a dataset to create a sample. NOT recommended as it may introduce bias.
- **Random sampling** – randomly selects a proportion of s% of the instances from a large dataset (Recommended approach)
- **Stratified sampling** – instances are divided into groups (or strata), where each group contains only instances that have a particular level for the stratification feature. s% of the instances in each stratum are randomly selected. These selections are combined to give an overall sample of s% of the original dataset.

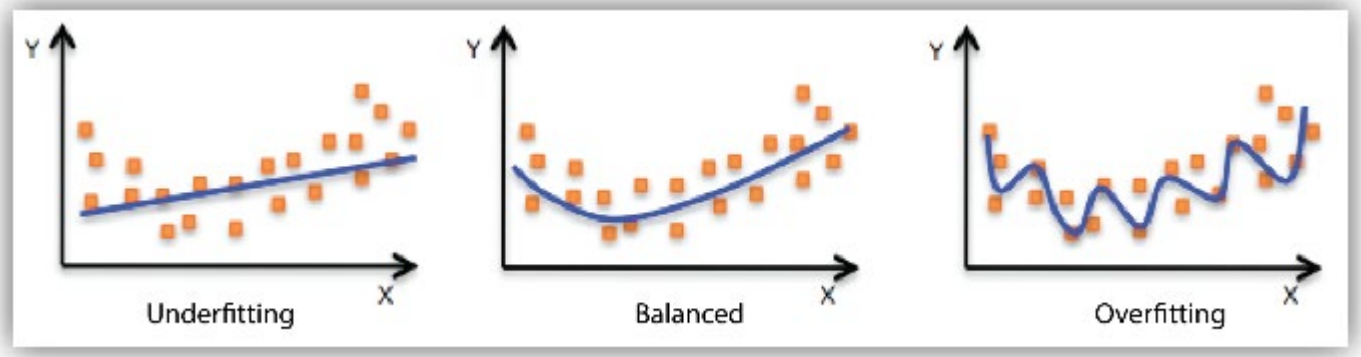
Train & Test set

- To perform learning, you need data.
- Learning will generate a classifier that can perform classification
- To test your classifier, you need data which is not used in learning process
- To test the effectiveness of your algorithm, you can split your data into two parts: a training set and a test set.
- The test set should be independent of the training set. It is required to verify the error rate of your algorithm.

Model Overfitting

- Overfitting is when you try to achieve 100% accuracy, even learning from the examples that are wrong (noise), leading to poor performance on new unseen data.
- Overfitted model performs well on the training data but does not perform well on the evaluation data. This is because the model is memorizing the data it has seen and is unable to generalize to unseen examples.
- **Be careful not to over fit.**

Underfitting vs Overfitting



Your model is *underfitting* the training data when the model performs poorly on the training data. This is because the model is unable to capture the relationship between the input and the target values (often called Y).

Refer to: https://docs.aws.amazon.com/machine-learning/latest/dg/model-fit-underfitting-vs-overfitting.html

Techniques for Evaluation

- **K-fold cross-validation**: partition into K equal groups. K-1 groups are for training, and the last remaining group is for testing. Repeat this K times, where K is usually 10. Take the average accuracy rate as the overall accuracy.
- **Percentage Split**: Split the data for train and test, majority of data for training and the remaining for testing (k% for training, (100-k)% for testing). Example: 70% for training, 30% for testing

Accuracy

- A confusion matrix is defined as the possible outcomes:

	Predicted +	Predicted -
Actually +	TP (actually +, predicted as +)	FN (actually +, but predicted as -)
Actually -	FP (actually -, predicted as +)	TN (actually -, predicted as -)

Terms

- *Accuracy*: $\frac{TP+TN}{TP+TN+FP+FN}$
  - Proportion of correct predictions
- *Precision*: $\frac{TP}{TP+FP}$
  - Of the predicted positives, how many are truly positive (ability to detect positives correctly)
- *Recall or Sensitivity*: $\frac{TP}{TP+FN}$.
  - These are the number of true cases you got right.
- *Specificity*: $\frac{TN}{TN+FP}$
  - These are the number of false cases you got right.
- F_1 Measure: $\frac{2 * precision * recall}{precision+recall}$
  - Harmonic mean of precision and recall

K-Nearest Neighbors

- One of the easiest classification algorithms
- For a new instance, it finds the k nearest neighbors (based on distance) in the train set.
- From the classes of k nearest neighbors, find the majority class.
- The class of the new instance will be the majority class

Demo in Excel

K3 =SQRT(POWER(F3-A3,2) + POWER(G3-B3,2) + POWER(H3-C3,2) + POWER(I3-D3,2))

	A	B	C	D	E	F	G	H	I	J	K	L
1	Train Instances					Test Instance						
2	SL	SW	PL	PW	FlowerType	SL	SW	PL	PW	FlowerType	Distance	
3	5.1	3.5	1.4	0.2	Iris-setosa	6.1	2.9	4.7	1.4	Iris-versicolor	3.7	
4	4.9	3	1.4	0.2	Iris-setosa	6.1	2.9	4.7	1.4	Iris-versicolor	3.712142	
5	4.7	3.2	1.3	0.2	Iris-setosa	6.1	2.9	4.7	1.4	Iris-versicolor	3.879433	
6	4.6	3.1	1.5	0.2	Iris-setosa	6.1	2.9	4.7	1.4	Iris-versicolor	3.737646	
7	5	3.6	1.4	0.2	Iris-setosa	6.1	2.9	4.7	1.4	Iris-versicolor	3.745664	
8	5.4	3.9	1.7	0.4	Iris-setosa	6.1	2.9	4.7	1.4	Iris-versicolor	3.38969	
9	4.6	3.4	1.4	0.3	Iris-setosa	6.1	2.9	4.7	1.4	Iris-versicolor	3.820995	
10	5	3.4	1.5	0.2	Iris-setosa	6.1	2.9	4.7	1.4	Iris-versicolor	3.624914	
11	4.4	2.9	1.4	0.2	Iris-setosa	6.1	2.9	4.7	1.4	Iris-versicolor	3.901282	
12	4.9	3.1	1.5	0.1	Iris-setosa	6.1	2.9	4.7	1.4	Iris-versicolor	3.661967	
13	5.4	3.7	1.5	0.2	Iris-setosa	6.1	2.9	4.7	1.4	Iris-versicolor	3.579106	
14	4.8	3.4	1.6	0.2	Iris-setosa	6.1	2.9	4.7	1.4	Iris-versicolor	3.604164	
15	4.8	3	1.4	0.1	Iris-setosa	6.1	2.9	4.7	1.4	Iris-versicolor	3.778889	
16	4.3	3	1.1	0.1	Iris-setosa	6.1	2.9	4.7	1.4	Iris-versicolor	4.230839	
17	5.8	4	1.2	0.2	Iris-setosa	6.1	2.9	4.7	1.4	Iris-versicolor	3.871692	
18	5.7	4.4	1.5	0.4	Iris-setosa	6.1	2.9	4.7	1.4	Iris-versicolor	3.694591	
19	5.4	3.9	1.3	0.4	Iris-setosa	6.1	2.9	4.7	1.4	Iris-versicolor	3.748333	
20	5.1	3.5	1.4	0.3	Iris-setosa	6.1	2.9	4.7	1.4	Iris-versicolor	3.668787	
21	5.7	3.8	1.7	0.3	Iris-setosa	6.1	2.9	4.7	1.4	Iris-versicolor	3.343651	
22	5.1	3.8	1.5	0.3	Iris-setosa	6.1	2.9	4.7	1.4	Iris-versicolor	3.641428	
23	5.4	3.4	1.7	0.2	Iris-setosa	6.1	2.9	4.7	1.4	Iris-versicolor	3.343651	
24	5.1	3.7	1.5	0.4	Iris-setosa	6.1	2.9	4.7	1.4	Iris-versicolor	3.588872	
25	4.6	3.6	1	0.2	Iris-setosa	6.1	2.9	4.7	1.4	Iris-versicolor	4.227292	
26	5.1	3.3	1.7	0.5	Iris-setosa	6.1	2.9	4.7	1.4	Iris-versicolor	3.312099	

Demo in Excel

	A	B	C	D	E	F	G	H	I	J	K	L	M
	Train Instances					Test Instance							
	SL	SW	PL	PW	FlowerType	SL	SW	PL	PW	FlowerType	Distance		Actual Class: Versicolor
	6.1	2.9	4.7	1.4	Iris-versicolor	6.1	2.9	4.7	1.4	Iris-versicolor	0		By 7NN
	6.1	3	4.6	1.4	Iris-versicolor	6.1	2.9	4.7	1.4	Iris-versicolor	0.141421		5 Versicolor
	6.1	2.8	4.7	1.2	Iris-versicolor	6.1	2.9	4.7	1.4	Iris-versicolor	0.223607		2 Virginica
	6	2.9	4.5	1.5	Iris-versicolor	6.1	2.9	4.7	1.4	Iris-versicolor	0.244949		Predicted Class: Versicolor
	6.2	2.9	4.3	1.3	Iris-versicolor	6.1	2.9	4.7	1.4	Iris-versicolor	0.424264		
	6	3	4.8	1.8	Iris-virginica	6.1	2.9	4.7	1.4	Iris-versicolor	0.43589		By 11NN
	6.2	2.8	4.8	1.8	Iris-virginica	6.1	2.9	4.7	1.4	Iris-versicolor	0.43589		7 Versicolor
	6.5	2.8	4.6	1.5	Iris-versicolor	6.1	2.9	4.7	1.4	Iris-versicolor	0.43589		4 Virginica
	6.1	3	4.9	1.8	Iris-virginica	6.1	2.9	4.7	1.4	Iris-versicolor	0.458258		Predicted Class: Versicolor
	5.7	2.8	4.5	1.3	Iris-versicolor	6.1	2.9	4.7	1.4	Iris-versicolor	0.469042		
	6.3	2.8	5.1	1.5	Iris-virginica	6.1	2.9	4.7	1.4	Iris-versicolor	0.469042		
	6.4	3.2	4.5	1.5	Iris-versicolor	6.1	2.9	4.7	1.4	Iris-versicolor	0.479583		
	6.3	3.3	4.7	1.6	Iris-versicolor	6.1	2.9	4.7	1.4	Iris-versicolor	0.489898		
	6	2.7	5.1	1.6	Iris-versicolor	6.1	2.9	4.7	1.4	Iris-versicolor	0.5		
	6.3	2.5	4.9	1.5	Iris-versicolor	6.1	2.9	4.7	1.4	Iris-versicolor	0.5		
	6.4	2.9	4.3	1.3	Iris-versicolor	6.1	2.9	4.7	1.4	Iris-versicolor	0.509902		
	6.6	2.9	4.6	1.3	Iris-versicolor	6.1	2.9	4.7	1.4	Iris-versicolor	0.519615		
	6.3	2.7	4.9	1.8	Iris-virginica	6.1	2.9	4.7	1.4	Iris-versicolor	0.52915		
	5.9	3.2	4.8	1.8	Iris-versicolor	6.1	2.9	4.7	1.4	Iris-versicolor	0.547723		
	5.9	3	4.2	1.5	Iris-versicolor	6.1	2.9	4.7	1.4	Iris-versicolor	0.556776		
	5.6	3	4.5	1.5	Iris-versicolor	6.1	2.9	4.7	1.4	Iris-versicolor	0.556776		
	6	3.4	4.5	1.6	Iris-versicolor	6.1	2.9	4.7	1.4	Iris-versicolor	0.583095		
	6.6	3	4.4	1.4	Iris-versicolor	6.1	2.9	4.7	1.4	Iris-versicolor	0.591608		