

# Module 6 - AI Ethics

- [Module 6 - AI Ethics](#)
- [General Notes](#)
- [Top Fears of AI](#)
  - [Privacy and Surveillance](#)
  - [Manipulation of Behavior](#)
  - [Opacity of AI](#)
  - [Bias in Decision Systems](#)
  - [Human Robot Interaction](#)
  - [Automation and Employment](#)
  - [Autonomous Systems](#)
  - [Machine Ethics](#)
  - [Responsibility for Robots](#)
  - [Singularity and Super-intelligence](#)
- [Australia Ethics Framework for AI \(2019\)](#)
- [More Info](#)
  - [Resources For Study](#)
- [Ethical Principles](#)
- [No Ethical AI](#)
- [Lab 4 - Letter Recognition](#)
  - [Support Vector Machines](#)
  - [Creating the SVM](#)
  - [Understanding The Model](#)
    - [Import Data Used](#)
    - [Two-Class Support Vector Machine and One vs All Multiclass](#)
  - [Evaluating The Model](#)

## General Notes

- [Google Slides](#)
- <https://wandb.ai/site/papers>
- **Convolutional Neural Network:** <https://poloclub.github.io/cnn-explainer/>

# Top Fears of AI

## 1. AI will produce biased outcomes

- An example was an AI that Amazon used for hiring that primarily hired only males.
- This is a valid fear, as people fear that the decisions made will not represent the minority.

## 2. We have no idea why AI does what it does

## 3. AI will make bad decisions

- A question to ask is why wasn't the model tested for the bad decision?

## 4. AI will lead to a loss of anonymity

- This is true as more and more data is collected, but it has been being collected long before AI.

## 5. AI will put me out of a job

# Privacy and Surveillance

- Our data that is being collected can be used for malicious purposes.
- There is also a positive side in that your collected data will provide relevant information and services to you.
- We live in a surveillance economy where there are cameras everywhere.
- *The Social Dilemma* is a good movie to watch that's related to AI ethics.

# Manipulation of Behavior

- Online manipulation and addiction.
  - One example is showing ads to a recovering alcoholic for alcohol. That's why there's now an option to filter / remove an ad.
- Deep fake text, photos, and video material.
  - Example deep fake video: <https://youtu.be/mUfJOQKdtAk>

# Opacity of AI

- Many AI systems rely on extracting patterns from given data sets, without "correct" solutions provided.

- Neither the programmer nor the end-user knows the patterns which the system has chosen.
  - How can we make fair decisions without knowing this process, if it's superior to humans?
    - I.e. Having to explain in court why this decision was made.

## Bias in Decision Systems

- How do we know if historical data has bias compared to new reformed data?
  - Part of the data engineer's job is to consider this.
  - You don't need historical data for things like autocorrect.
- Data sets can be made to focus on a single matter with no bias, but how do we account for the bias if it is used for another matter?

## Human Robot Interaction

- People may not want the human component taken away.
  - An example being sex robots. Do we want humans growing intimate deep attachments to objects that cannot have feelings or mean what it says?
    - Example video: [https://youtu.be/dtu4t\\_Zc3d4](https://youtu.be/dtu4t_Zc3d4)

## Automation and Employment

- Will the creation of new jobs and wealth keep up with the destruction of jobs?
- AI has the potential to further create the gap between high skill / high paid jobs and low skilled / low paid service jobs.

## Autonomous Systems

- [The Rise of the Machines – Why Automation is Different this Time](#)
- [How Insurance Will Work With Self-Driving Cars](#)

Some questions asked against autonomous systems:

- Is the autonomy responsible for its decision-making?
- Is it best to leave our lives to a system that will choose the common defined good over pursuing personal interest?
  - Usually an AI engineer in Silicon valley is deciding what the common good is.

- Is it best for an autonomous weapon to fight against other autonomous weapons to save human interaction in conflict?

## Machine Ethics

[Do Robots Deserve Rights? What if Machines Become Conscious?](#)

**Machine Ethics:** Ethics of machines as subjects, not objects in use.

- **First Law:** Robot may not injure a human being
- **Second Law:** Robot must obey orders given it by human except by conflict of first law.
- **Third Law:** Robot must protect its own existence as long as protection does not conflict with the first and second law.

## Responsibility for Robots

- Who is responsible, liable, or accountable? The owner or the manufacturer of the robot?

## Singularity and Super-intelligence

[What is a Singularity? | Eternally Curious #11](#)

- Does AI have the potential to surpass human intelligence and create its own systems that also surpass humanity?
- At what point do we define an AI to be super-intelligence and what is our plan if such technological advancement occurs?

## Australia Ethics Framework for AI (2019)

- Generates net-benefits > for people
- Do not harm or deceive
- Regulatory and legal compliance
- Privacy protection
- Fairness
  - No discrimination

- No bias in training data
- Transparency / Explain-ability
  - Explain impact and decision-making
- Contest-ability
  - Challenge use or output
- Accountability
  - Keep creators responsible for outcome

## More Info

- <https://plato.stanford.edu/entries/ethics-ai/>
- <https://www.nature.com/articles/s42256-019-0088-2#Sec15>
- <https://futureoflife.org/ai-policy>

## Resources For Study

- [Turing Institute \(UK\)](#)
- [AI Now](#)
- [Leverhulme Centre for the Future of Intelligence](#)
- [Future of Humanity Institute](#)
- [Future of Life Institute](#)
- [Stanford Center for Internet and Society](#)
- [Berkman Klein Center](#)
- [Digital Ethics Lab](#)
- [Open Roboethics Institute](#)

# Ethical Principles

## Ethical Principles Identified



- Transparency
- Justice and Fairness
- Non-Maleficence
- Responsibility
- Privacy
- Beneficence
- Freedom and Autonomy
- Trust
- Sustainability
- Dignity
- Solidarity

## No Ethical AI

The idea that there is no ethical AI: <https://onezero.medium.com/theres-no-such-thing-as-ethical-ai-38891899261d>

## Lab 4 - Letter Recognition

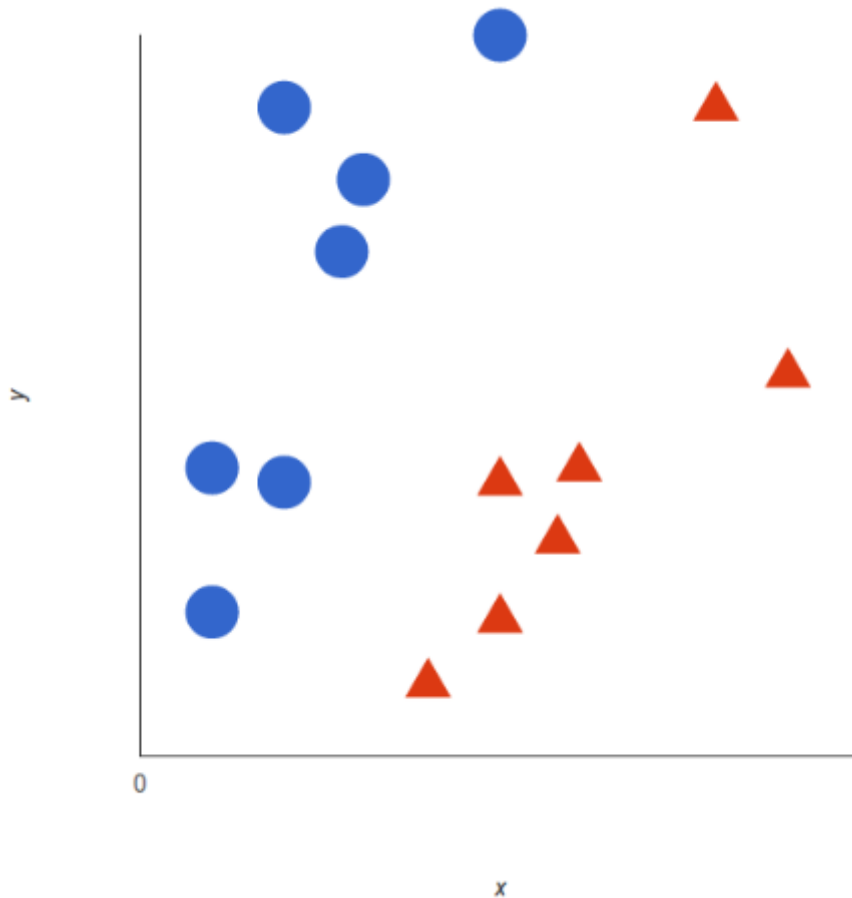
- Using [Microsoft Azure](#)

## Support Vector Machines

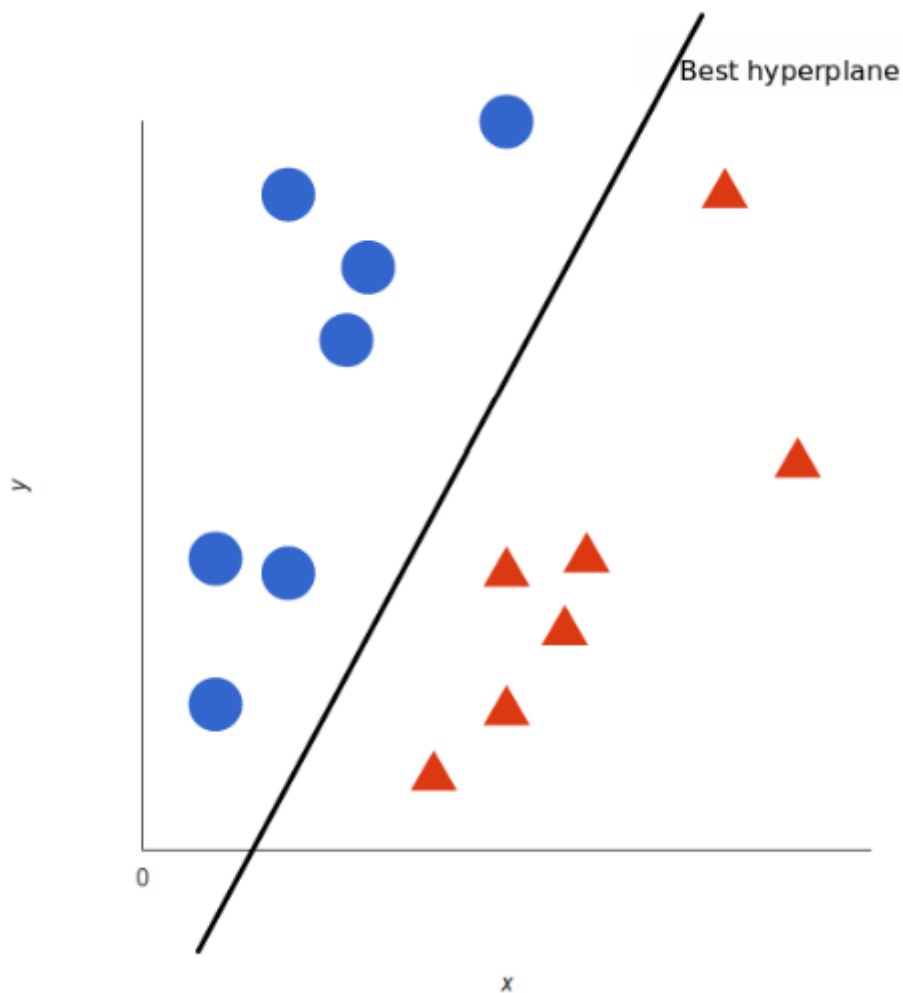
A **support vector machine (SVM)** is a machine learning model that is used to classify new data into two separate groups.

First, we give the model a set of data and what group each point of the data belongs to (training). The model then tells us what group the new data points belong to.

Let's imagine we have two groups: *red* and *blue*. We will also say that our data has two features:  $x$  and  $y$ . We want a classifier that given a pair of  $(x,y)$  coordinates, outputs if it's either *red* or *blue*.



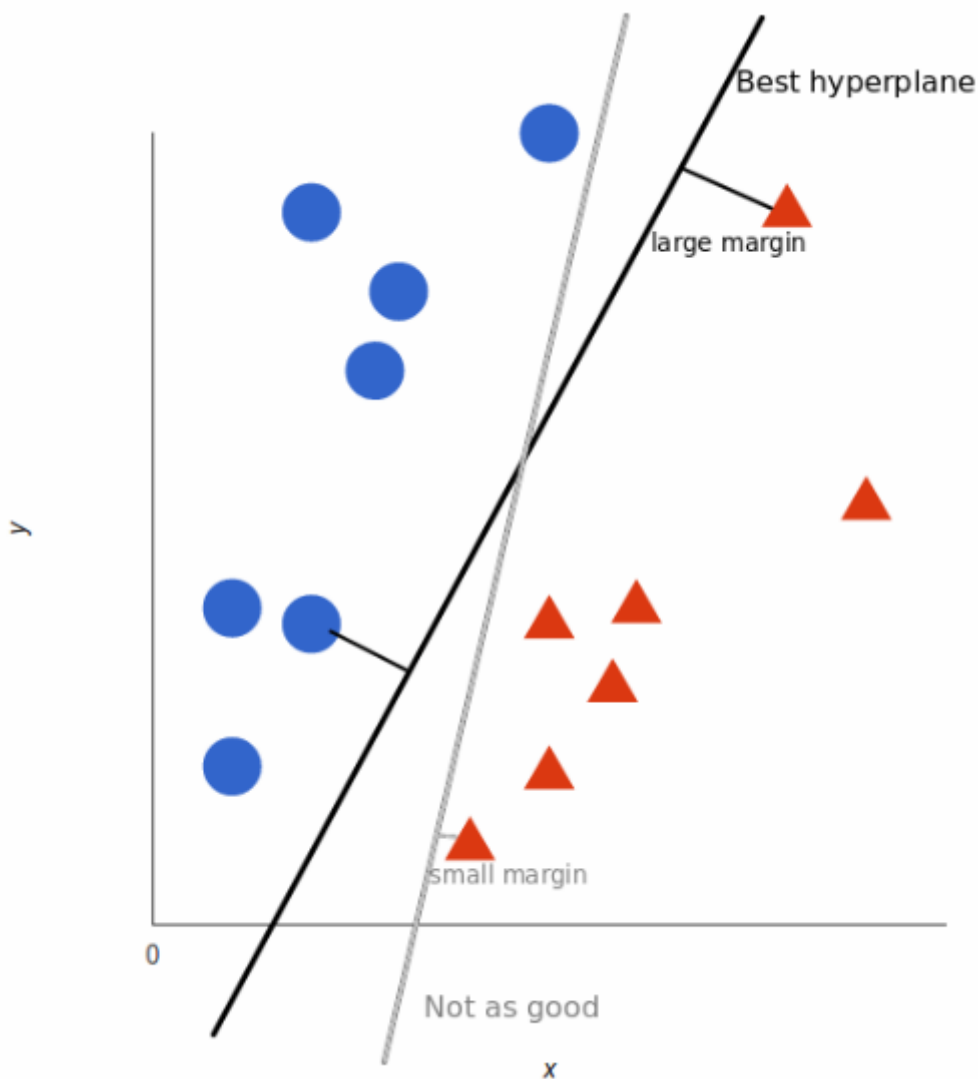
A support vector machine takes these data points and outputs the hyperplane (for two dimensions it's simply a line) that best separates the tags. This line is called the **decision boundary**.



- In the photo above, the decision boundary is the black line.
- Any new data point that falls left of the black line will be classified as blue.
- Any new data point falls right of the black line will be classified as red.

To find the best line to draw, there are a couple of ways to create this calculation but for the example below, it takes the largest margin between the line and the nearest point of each group to the line.






“source: <https://monkeylearn.com/blog/introduction-to-support-vector-machines-svm/>

## Creating the SVM

1. Go to <https://portal.azure.com/#home>
2. Create a new resource group
3. Create a new Azure Machine Learning Workspace
4. Launch the studio by going to the workspace and selecting **Launch Studio**
5. Go to Designer
6. Go to Settings
7. Create a new computer cluster using the cheapest options
  - Wait for it to finish initializing (It will go from blue to green)
8. Drag **Import Data** into the canvas


1.  Data Input and Output -> Import Data

2. Change **Data Source** to **URL via HTTP**


3. Use the following url for the data: <http://archive.ics.uci.edu/ml/machine-learning-databases/letter-recognition/letter-recognition.data>

9. Output **Import Data** to **Split Data** at 0.5 with randomized seed.

10. Drag **Two-Class Support Vector Machine** onto the canvas

◦  Machine Learning Algorithms -> Classification

11. Drag **One-vs-All Multiclass** onto canvas

◦  Machine Learning Algorithms -> Classification

12. Connect output of **Two-Class SVM** to input of **One-vs-All Multiclass**

13. Drag **Train Model** onto the canvas

14. Connect output of **One-vs-All Multiclass** into **Train Model input 1**

15. Connect **output 1** of **Split Data** into **Train Model input 2**

- For **Train Model** properties, launch the column select and under **With Rules**, use **Include -> Column Index** and enter the value **1**.

16. Drag **Score Model** onto the canvas.

17. Connect **Train Model** output into **Score Model input 1**

18. Connect **Split Data output 2** into **Score Model input 2**

19. Drag **Evaluate Model** onto the canvas

20. Connect **Score Model** output into **Evaluate Model input 1**

21. Hit Run and create a **Computer Cluster**

## Understanding The Model

### Import Data Used

Here is a description of the dataset: <https://archive.ics.uci.edu/ml/datasets/Letter+Recognition>

Column Meanings:

1. lettr capital letter (26 values from A to Z)

2. x-box horizontal position of box (integer)
3. y-box vertical position of box (integer)
4. width width of box (integer)
5. high height of box (integer)
6. onpix total # on pixels (integer)
7. x-bar mean x of on pixels in box (integer)
8. y-bar mean y of on pixels in box (integer)
9. x2bar mean x variance (integer)
10. y2bar mean y variance (integer)
11. xybar mean x y correlation (integer)
12. x2ybr mean of  $x * x * y$  (integer)
13. xy2br mean of  $x * y * y$  (integer)
14. x-egge mean edge count left to right (integer)
15. xegvy correlation of x-egge with y (integer)
16. y-egge mean edge count bottom to top (integer)
17. yegvx correlation of y-egge with x (integer)

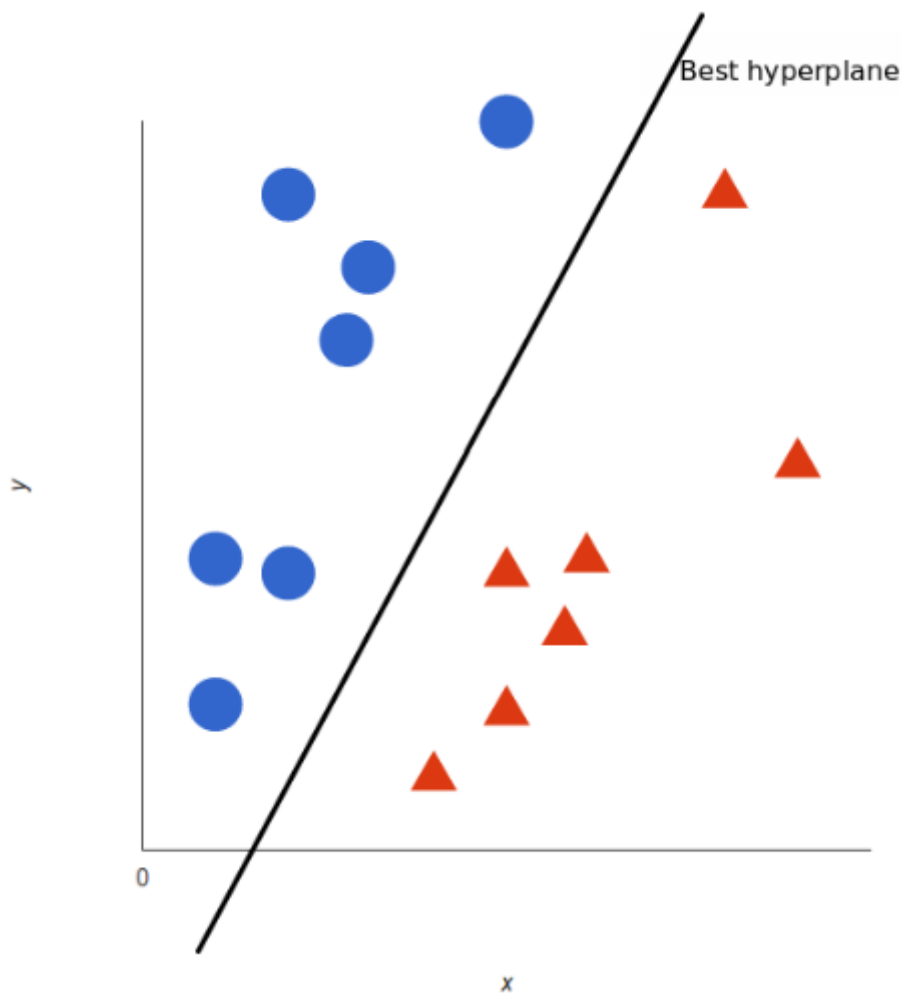
**Summary:** The objective is to identify each of a large number of black-and-white rectangular pixel displays as one of the 26 capital letters in the English alphabet.

The character images were based on 20 different fonts and each letter within these 20 fonts was randomly distorted to produce a file of 20,000 unique samples.

The data itself is a CSV (excel file) that has 17 columns in total. The first column is the capital letter that we describe and the rest of the columns are statistics that characterize the letter.

## Two-Class Support Vector Machine and One vs All Multiclass

There's two classes (or groups) that the support vector machine is drawing a line (hyperplane) to classify between one or the other on a given data point (row, column).

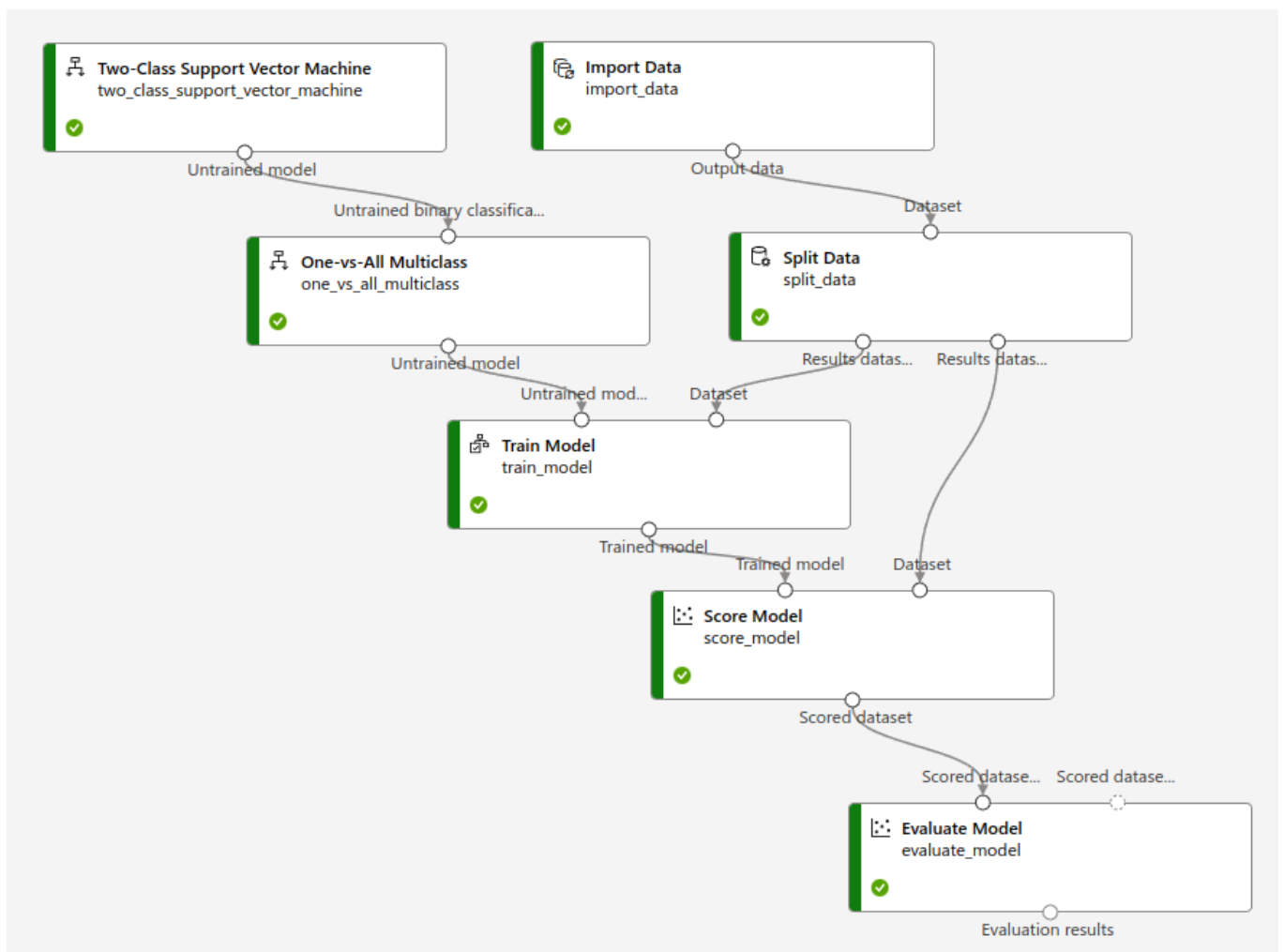


If you right-click **Import Data** -> **Preview Data**, you will see the data pulled.

Column1	Column2	Column3	Column4	Column5	Column6	Column7	Column8	Column9	Column10	Column11	Column12	Column13	Column14	Column15
T	2	8	3	5	1	8	13	0	6	6	10	8	0	8
I	5	12	3	7	2	10	5	5	4	13	3	9	2	8
D	4	11	6	8	6	10	6	2	6	10	3	7	3	7
N	7	11	6	6	3	5	9	4	6	4	4	10	6	10
G	2	1	3	1	1	8	6	6	6	6	5	9	1	7
S	4	11	5	8	3	8	8	6	9	5	6	6	0	8
B	4	2	5	4	4	8	7	6	6	7	6	6	2	8
A	1	1	3	2	1	8	2	2	2	8	2	8	1	6
J	2	2	4	4	2	10	6	2	6	12	4	8	1	6
M	11	15	13	9	7	13	2	6	2	12	1	9	8	1
X	3	9	5	7	4	8	7	3	8	5	6	8	2	8
O	6	13	4	7	4	6	7	6	3	10	7	9	5	9
G	4	9	6	7	6	7	8	6	2	6	5	11	4	8
M	6	9	8	6	9	7	8	6	5	7	5	8	8	9
R	5	9	5	7	6	6	11	7	3	7	3	9	2	7
F	6	9	5	4	3	10	6	3	5	10	5	7	3	9
O	3	4	4	3	2	8	7	7	5	7	6	8	2	8
C	7	10	5	5	2	6	8	6	8	11	7	11	2	8
T	6	11	6	8	5	6	11	5	6	11	9	4	3	12
J	2	2	3	3	1	10	6	3	6	12	4	9	0	7
J	1	3	2	2	1	8	8	2	5	14	5	8	0	7
H	4	5	5	4	4	7	7	6	6	7	6	8	3	8
S	3	2	3	3	2	8	8	7	5	7	5	7	2	8
O	6	11	7	8	5	7	6	9	6	7	5	9	4	8
J	3	6	4	4	2	6	6	4	4	14	8	12	1	6
C	6	11	7	8	3	7	8	7	11	4	7	14	1	7
M	7	11	11	8	9	3	8	4	5	10	11	10	10	9
W	12	14	12	8	5	9	10	4	3	5	10	7	10	12

- Col 1 is the letter and ever other column is a value that represents a feature of that letter.
- The SVM is creating a hyperplane for each of these columns to tell us whether it is that specific letter.
- Using the image above, the blue could be data points that **are not T**, while the red are data points that **are T**.
- The Support Vector Machine is doing this for every letter.

The hyperplane is not a line in a 2-D space, it's actually a *line (or plane)* in a multidimensional space with each dimension being a column that is in the data.



- The **Two-Class support Vector Machine** is outputting an estimation of its confidence of a letter.
- The **One-vs-All Multiclass** is doing a comparison between the confidence of all the letters and choosing the letter with the highest confidence to output.

## Evaluating The Model

Go to **Score Model -> Preview Data**, and scroll to the right to **Scored Probabilities** section.

- You can see the real letter in *col1* and the probability assigned to the likelihood that the letter is one of the letters in the alphabet.

Go to **Evaluate Model -> Preview Data** to see the overall & average accuracy, precision, recall, and a confusion matrix that shows on the x-axis the Predicted Class and on the y-axis the Actual Class.