SFU SURGE

Intro to Data Science and ML

We acknowledge that all SFU Surge initiatives, including the opportunity to host this workshop today, takes place on xʷməθkʷəy̓əm (Musqueam), səlilwətaɬ (Tsleil-Waututh), and Sḵwx̱wú7mesh (Squamish) nations.

As build the tech community of tomorrow, its important to understand and respect indigenous histories!

Access This File.

It is also available online in the StormHacks discord Server!

Hi 👋👋

I’m Matt!

👨‍⚖️ I’ve judged at over 10 hackathons

🏆 Won over 20 hackathons globally

💉 Worked in the healthcare industry

🚀 Organized 6 hackathons

@ermergesh

/in/MatthewWong1129

What we'll cover

Introduction to Data Science
What it is and why it matters

Data Types & Structure
Understanding the different kinds of data

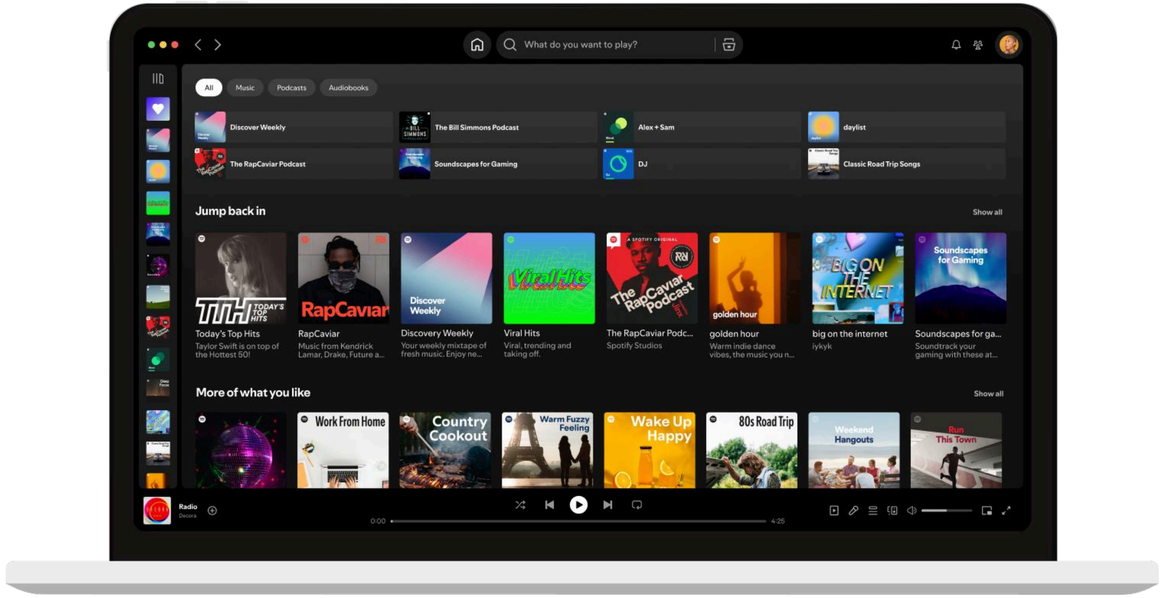
Data Cleaning & Preprocessing
Making messy data ready for use

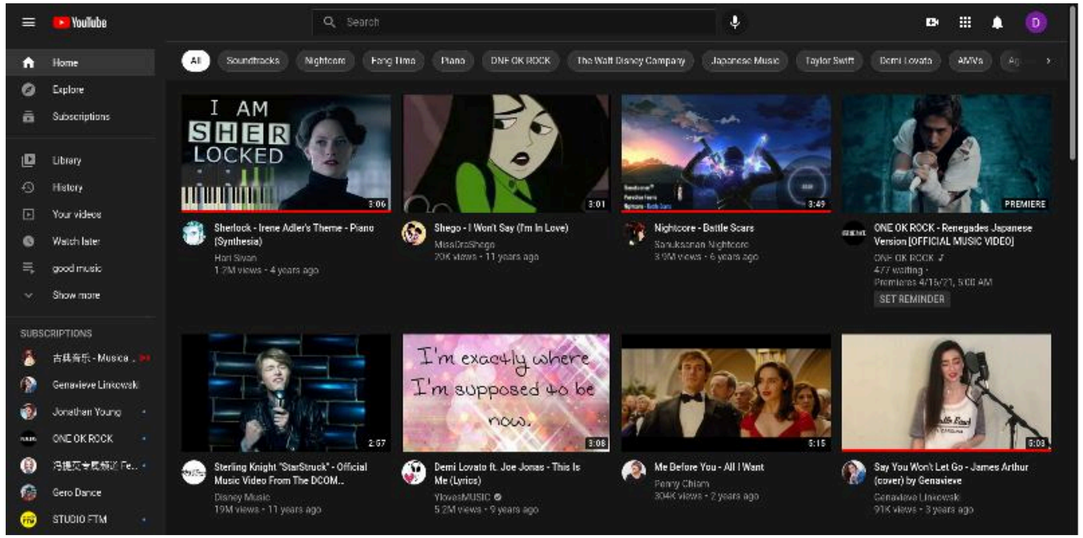
Exploratory Data Analysis
Finding insights through visualization

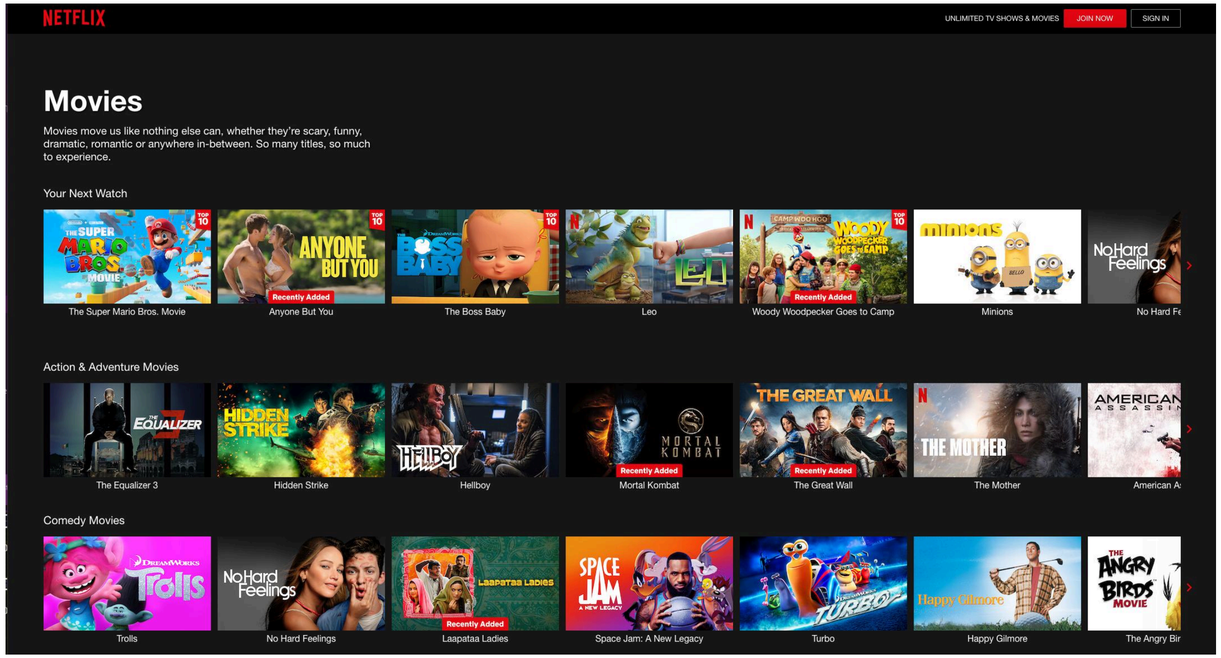
Machine Learning Basics
Core ML concepts for a hackathon

Demo: Kaggle Titanic Dataset
Applying what we learned

Why is it so important?


What do you want to play?
All
Music
Podcasts
Audiobooks
Discover Weekly
The Bill Simmons Podcast
Alex + Sam
daylist
The RapCaviar Podcast
Soundscapes for Gaming
DJ
Classic Road Trip Songs
Jump back in
Show all
Today's Top Hits
Taylor Swift is on top of the Hottest 50!
RapCaviar
Music from Kendrick Lamar, Drake, Future a...
Discover Weekly
Your weekly mixtape of fresh music. Enjoy ne...
Viral Hits
Viral, trending and taking off.
The RapCaviar Podc...
Spotify Studios
golden hour
Warm indie dance vibes, the music you n...
big on the internet
iykyk
Soundscapes for ga...
Soundtrack your gaming with these at...
More of what you like
Show all
Work From Home
Country Cookout
Warm Fuzzy Feeling
Wake Up Happy
80s Road Trip
Weekend Hangouts
Run This Town
Radio


YouTube
Search
Home
Explore
Subscriptions
Library
History
Your videos
Watch later
good music
Show more
SUBSCRIPTIONS
Genevieve Linkowski
Jonathen Young
ONE OK ROCK
Gero Dance
STUDIO FTM
All
Soundtracks
Nightcore
Pang Hina
Piano
ONE OK ROCK
The Walt Disney Company
Japanese Music
Taylor Swift
Demi Lovato
AMVs
I AM SHERLOCKED
Sherlock - Irene Adler's Theme - Piano (Synthesia)
Hari Sivan
1.2M views • 4 years ago
Shego - I Won't Say (I'm In Love)
MisaDrsShego
20K views • 11 years ago
Nightcore - Battle Scars
Sanukaenan Nightcore
3.9M views • 6 years ago
PREMIERE
ONE OK ROCK - Renegades Japanese Version [OFFICIAL MUSIC VIDEO]
ONE OK ROCK
477 waiting
Premieres 4/15/21, 5:00 AM
SET REMINDER
Sterling Knight "StarStruck" - Official Music Video From The DCOM...
Disney Music
19M views • 11 years ago
I'm exactly where I'm supposed to be now.
Demi Lovato ft. Joe Jonas - This Is Me (Lyrics)
YlovatoMUSIC
5.2M views • 9 years ago
Me Before You - All I Want
Penny Chiam
304K views • 2 years ago
Say You Won't Let Go - James Arthur (cover) by Genavieve
Genavieve Linkowski
91K views • 3 years ago


NETFLIX
UNLIMITED TV SHOWS & MOVIES
JOIN NOW
SIGN IN
Movies
Movies move us like nothing else can, whether they're scary, funny, dramatic, romantic or anywhere in-between. So many titles, so much to experience.
Your Next Watch
The Super Mario Bros. Movie
Anyone But You
The Boss Baby
Leo
Woody Woodpecker Goes to Camp
Minions
No Hard Fe
Action & Adventure Movies
The Equalizer 3
Hidden Strike
Hellboy
Mortal Kombat
The Great Wall
The Mother
American A
Comedy Movies
Trolls
No Hard Feelings
Laapataa Ladies
Space Jam: A New Legacy
Turbo
Happy Gilmore
The Angry Bir


WAYMO

Data Science

Turning raw data into actionable insights
using math, coding and storytelling

Data Cleaning

Removing duplicate/missing values and standardizing information

Exploratory Data Analysis

Summarizing and visualizing data

Feature Engineering

Creating useful new variables for models

Modeling

Building mathematical representations of data to make predictions

Types of Data

Structured

- Tables/arrays

Numerical

- numbers
  - stock prices

Unstructured

- Raw text or images

Categorical

- types
  - product categories (clothing, tech, etc)

Data Science Tools

Pandas

- Data manipulation

Scikit-Learn, TensorFlow, Pytorch

- Machine learning models

Seaborn + Matplot

- Data Visualization

NumPy

- Numerical Calculations

Kaggle

- Database of datasets

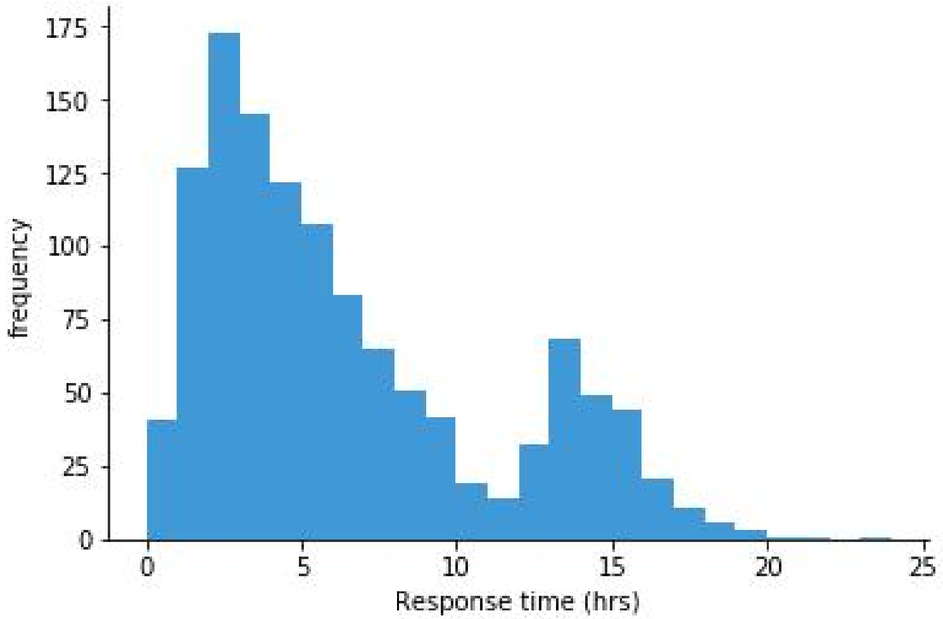
Jupyter Notebook

- IDE

Data Visualization Basics

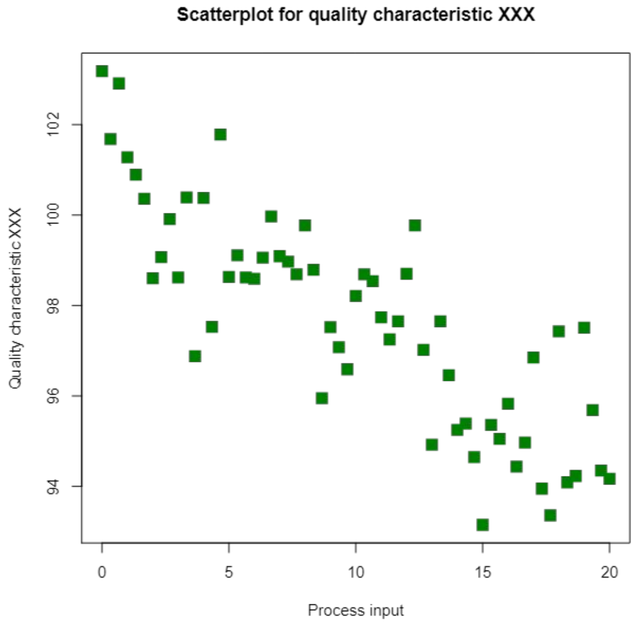
Histogram

- Shows data distribution



Scatterplot

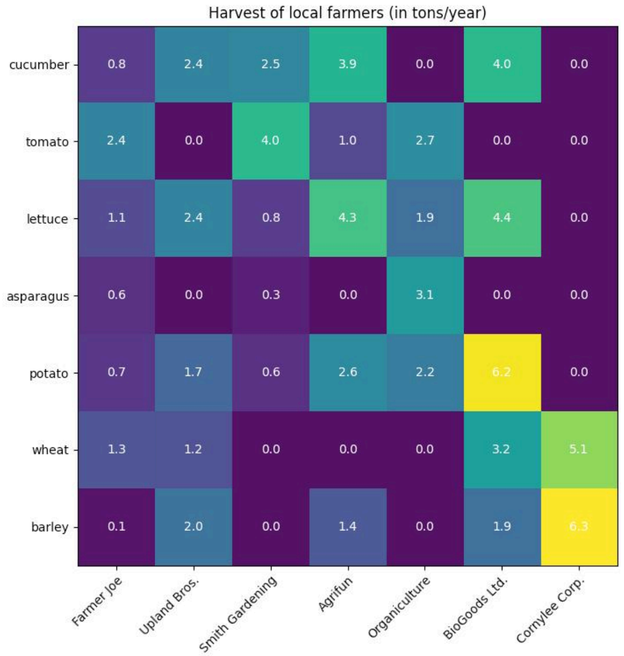
- Identifies relationships



Data Visualization Basics

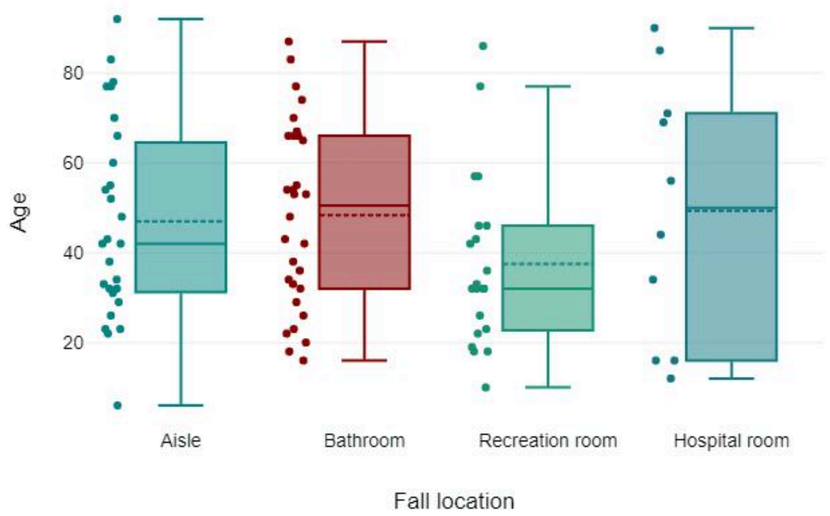
Heatmap

- Shows correlations

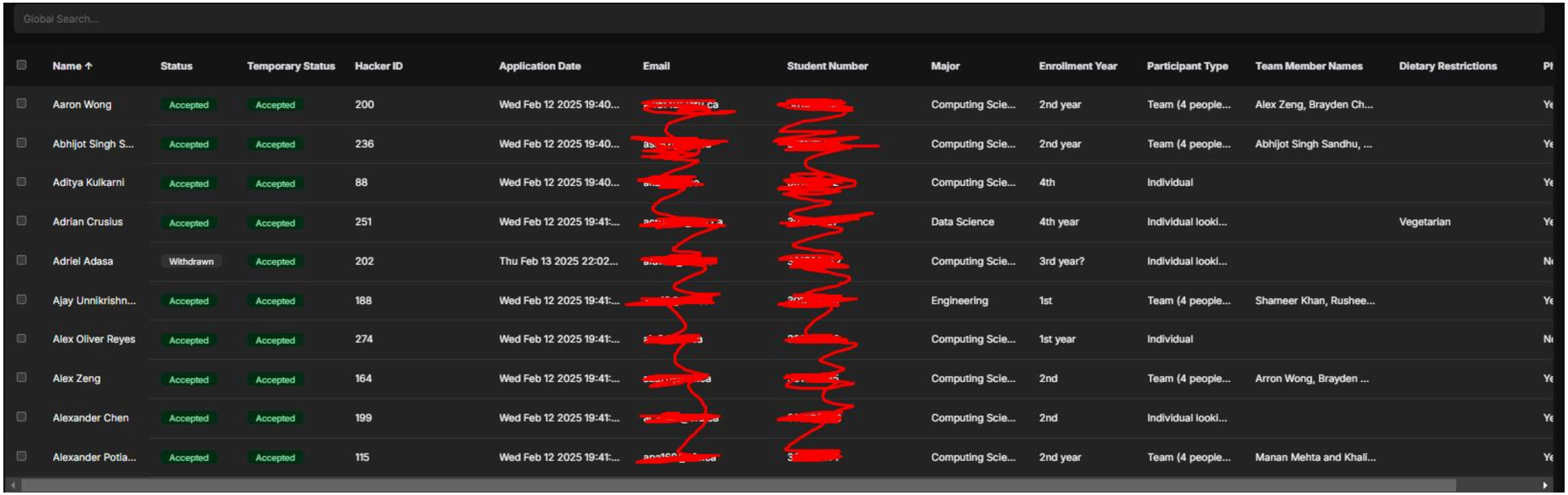


Box Plot

- Highlights outliers



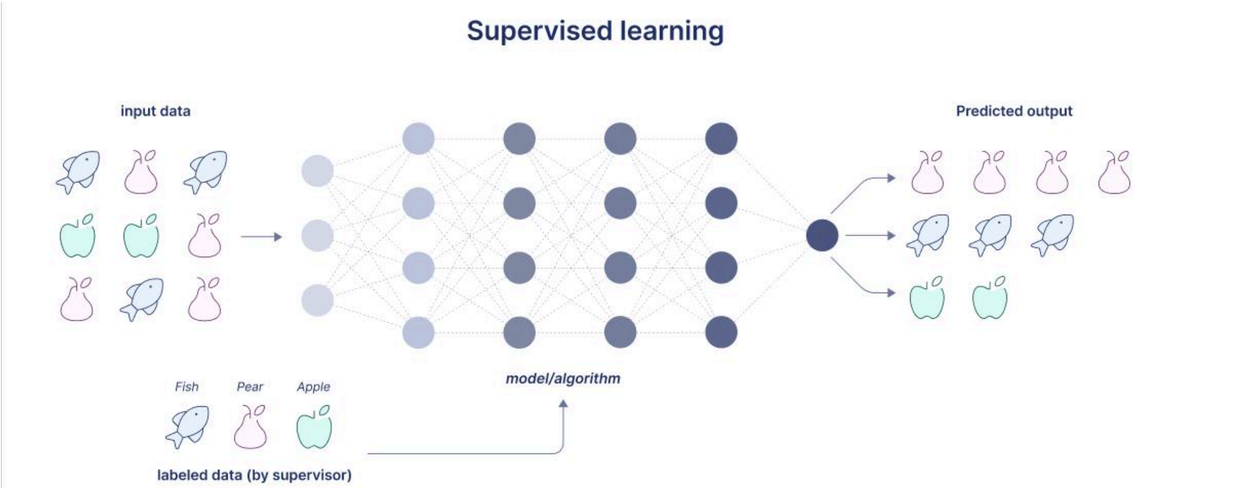
What's wrong?

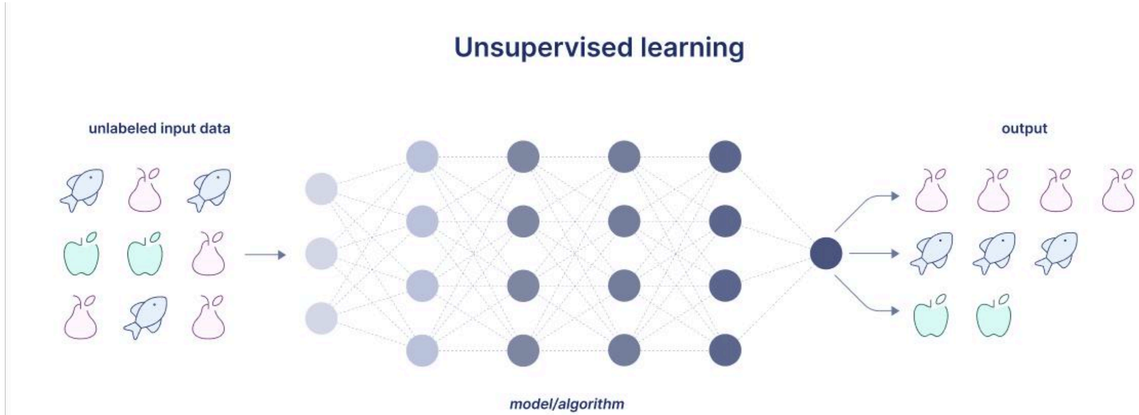


Global Search...

Name	Status	Temporary Status	Hacker ID	Application Date	Email	Student Number	Major	Enrollment Year	Participant Type	Team Member Names	Dietary Restrictions	Pl
Aaron Wong	Accepted	Accepted	200	Wed Feb 12 2025 19:40...	[redacted]	[redacted]	Computing Scie...	2nd year	Team (4 people...	Alex Zeng, Brayden Ch...		Ye
Abhijot Singh S...	Accepted	Accepted	236	Wed Feb 12 2025 19:40...	[redacted]	[redacted]	Computing Scie...	2nd year	Team (4 people...	Abhijot Singh Sandhu, ...		Ye
Aditya Kulkarni	Accepted	Accepted	88	Wed Feb 12 2025 19:40...	[redacted]	[redacted]	Computing Scie...	4th	Individual			Ye
Adrian Crusius	Accepted	Accepted	251	Wed Feb 12 2025 19:41:...	[redacted]	[redacted]	Data Science	4th year	Individual looki...		Vegetarian	Ye
Adriel Adasa	Withdrawn	Accepted	202	Thu Feb 13 2025 22:02...	[redacted]	[redacted]	Computing Scie...	3rd year?	Individual looki...			N
Ajay Unnikrishn...	Accepted	Accepted	188	Wed Feb 12 2025 19:41:...	[redacted]	[redacted]	Engineering	1st	Team (4 people...	Shameer Khan, Rushee...		Ye
Alex Oliver Reyes	Accepted	Accepted	274	Wed Feb 12 2025 19:41:...	[redacted]	[redacted]	Computing Scie...	1st year	Individual			N
Alex Zeng	Accepted	Accepted	164	Wed Feb 12 2025 19:41:...	[redacted]	[redacted]	Computing Scie...	2nd	Team (4 people...	Arron Wong, Brayden ...		Ye
Alexander Chen	Accepted	Accepted	199	Wed Feb 12 2025 19:41:...	[redacted]	[redacted]	Computing Scie...	2nd	Individual looki...			Ye
Alexander Potia...	Accepted	Accepted	115	Wed Feb 12 2025 19:41:...	[redacted]	[redacted]	Computing Scie...	2nd year	Team (4 people...	Manan Mehta and Khali...		Ye

Machine Learning in 5 minutes

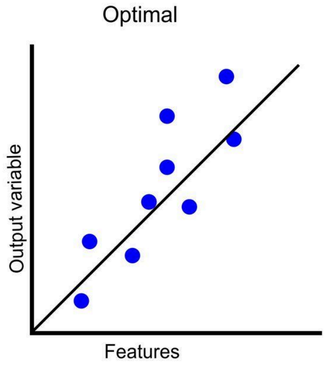


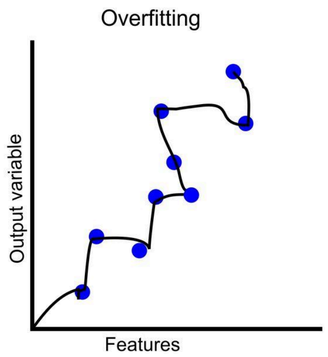


Machine Learning in 5 minutes

Overfitting

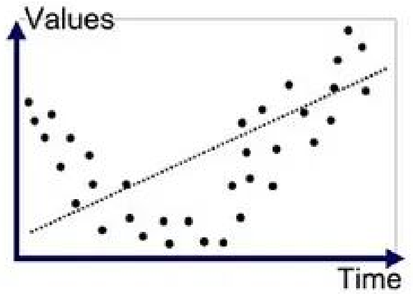
Model memorizes the data instead of generalizing it



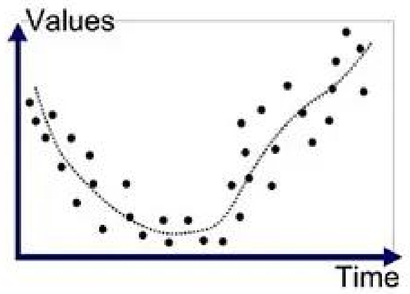


Underfitting

Model is too simple and misses out on key patterns



Underfitted



Good Fit/Robust

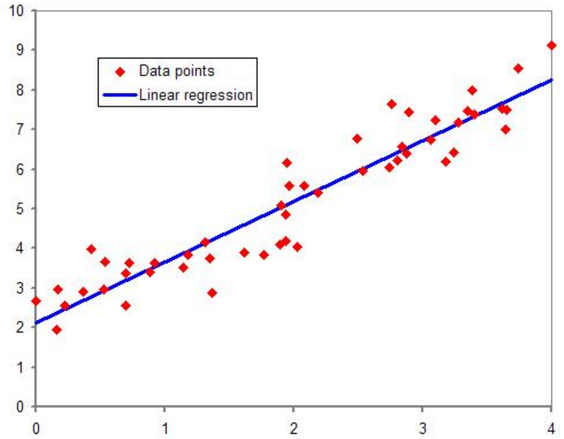
Pro hacker tip:

Garbage in, garbage out

Machine Learning in 5 minutes

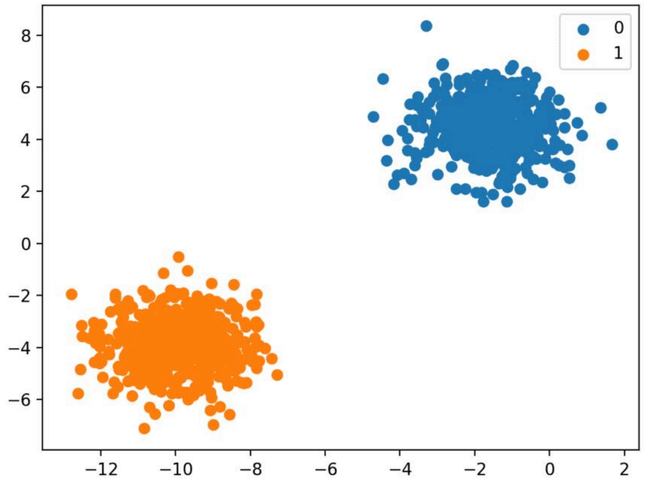
Regression

Predicts recurring values like housing prices



Classification

Predicts categories such as spam emails



Regression Models

Linear Regression

Line of best fit

$y = mx + b \dashrightarrow y=\beta_0+\beta_1 x+\varepsilon$

$\beta_0 = m$, $\beta_1 = b$, ε = error

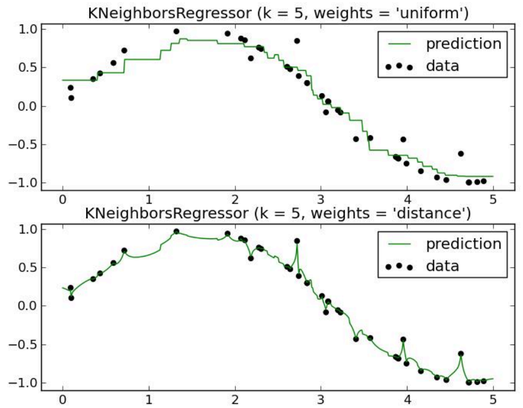
kNN (regression vers.)

Predict value by average of ‘k’ nearest neighbours

Facebook Prophet

A regression model for time series

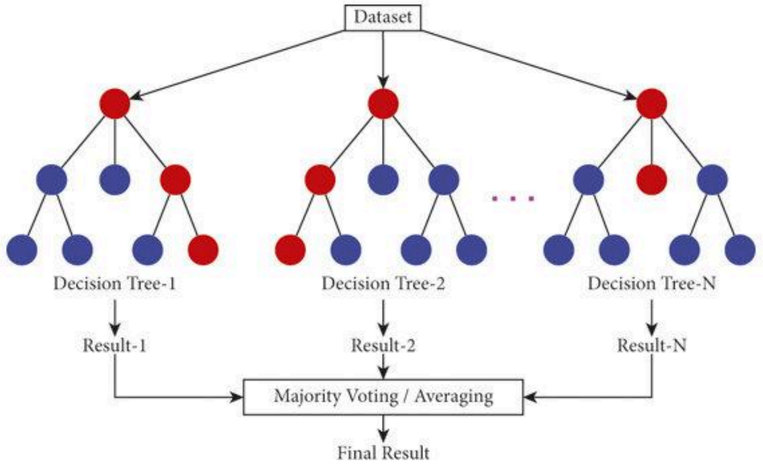
Breaks data into: Trend + seasonality + special events + error



Classification Models

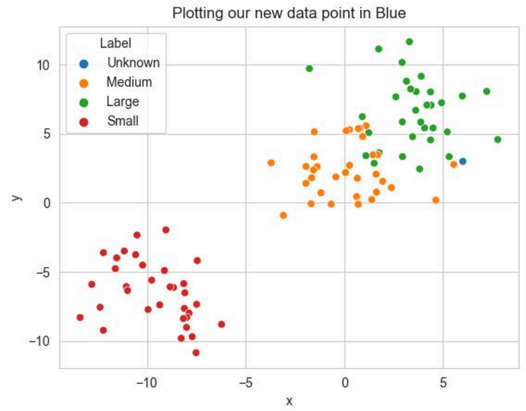
Random Forests

A bunch of decision trees



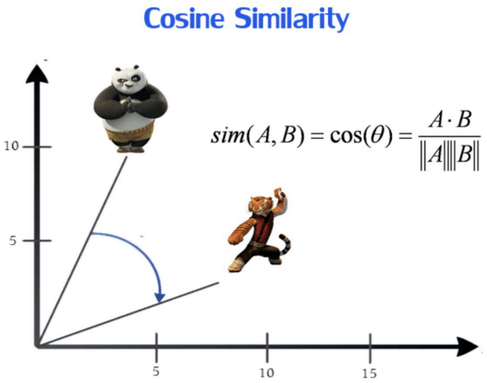
TF-IDF + Naive Bayes (Gaussian)

Classification of word weights



kNN (classification vers.)

Labels all closest neighbors.



Demo Time!