SPAI

An AI Singapore Student Chapter

ML Bootcamp

Day 3

Scan to mark attendance

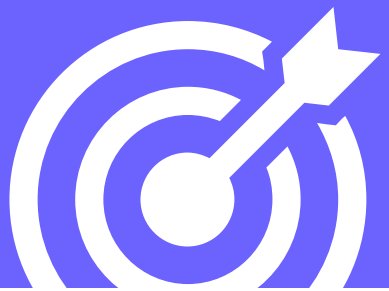
Scan the QR code to mark your attendance

Attendance

Machine Learning Hands-On

Hoo-ray!

SPAI



Objectives

- [x] Quick Recap
- [x] Introduction to Machine Learning Problems and Models
- [x] Model Building and Evaluation

1

Introduction to ML Problems and Models

Classification VS Regression

Classification

- Supervised Machine Learning
- Predicts category (E.g. “Cat” vs “Dog”)

Regression

- Supervised Machine Learning
- Predicts number (E.g. Price of house)

Knowledge Check

Which of the following scenario represents a Supervised Classification Problem?

A. Using labelled financial data to predict whether the value of a stock will go up or go down next week.

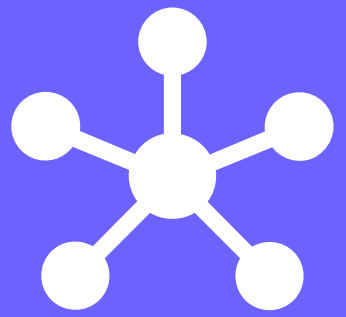
B. Using labelled housing price data to predict the price of a new house based on various features.

C. Using unlabelled data to cluster the students of an online education company into different categories based on their learning styles.

D. Using labelled financial data to predict what the value of a stock will be next week.

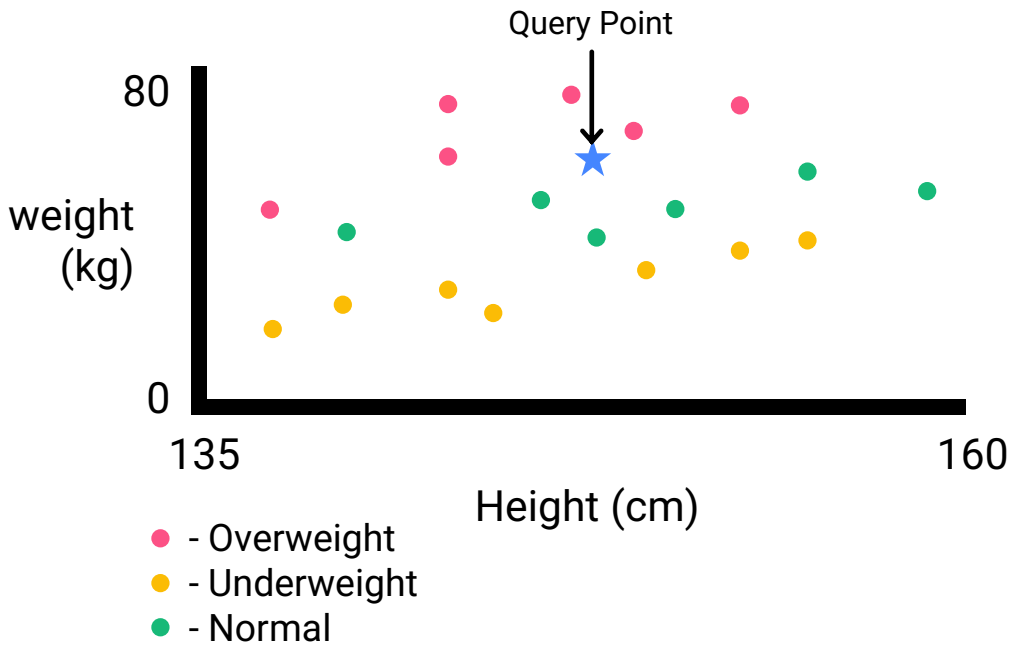
Source : DataCamp supervised-learning-with-scikit-learn/classification?ex=2

Classification Model



KNN

- K-Nearest Neighbours
- Classifies new data points based on similarity measures of stored data points
- Can be used for both classification and regression



Given the height of the weight, predict the category of person (Overweight/Underweight/Normal)

If K is set to 3, we only use 3 nearest neighbours to predict the category.

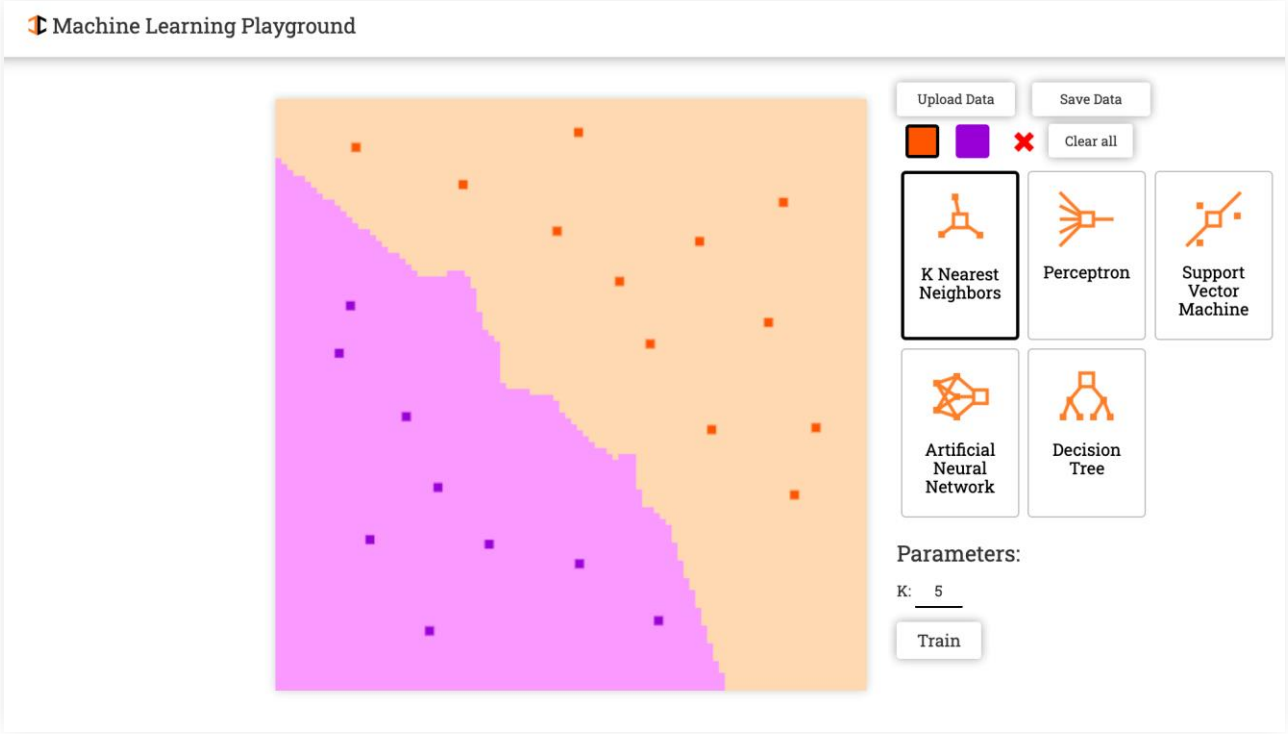
Out of the 3 nearest neighbours, 2 of them belong to the “Overweight” category.

Hence, that person belongs to the “Overweight” category

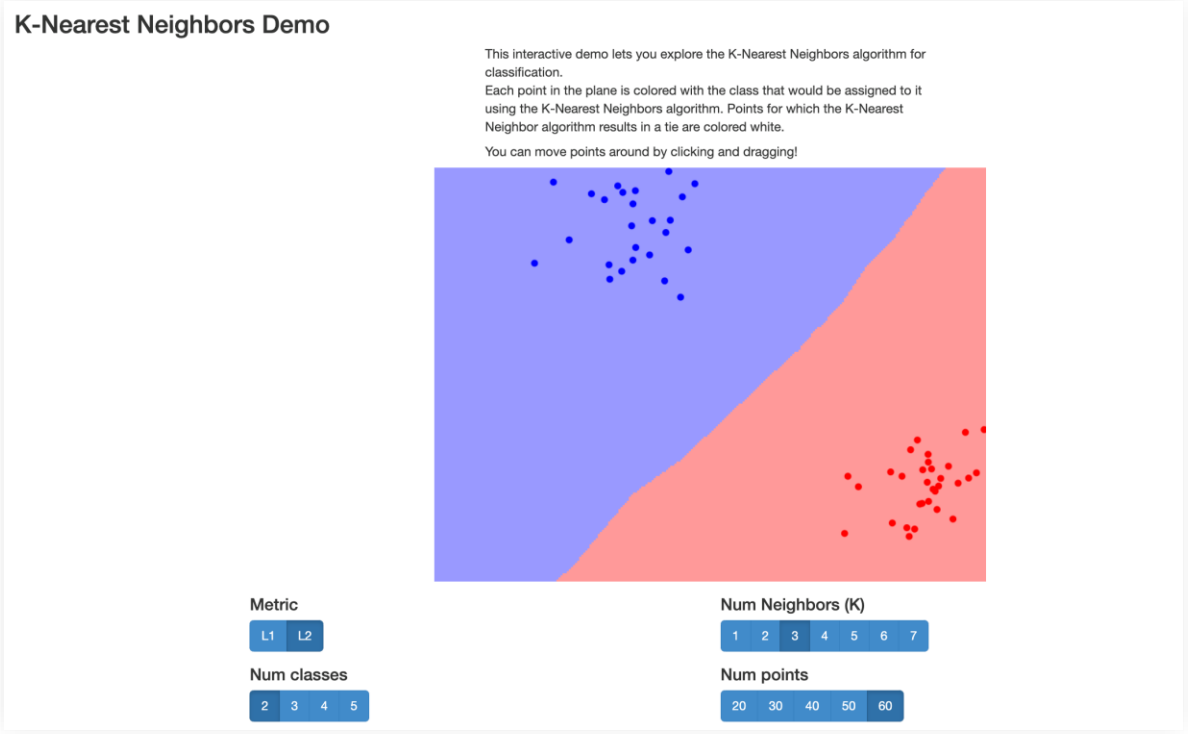
KNN

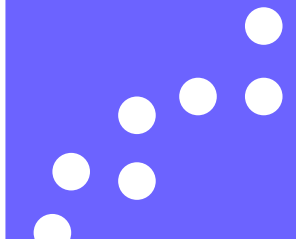
Interactive Examples:

https://peterleong.github.io/ML-Playground/



http://vision.stanford.edu/teaching/cs231n-demos/knn/





KNN

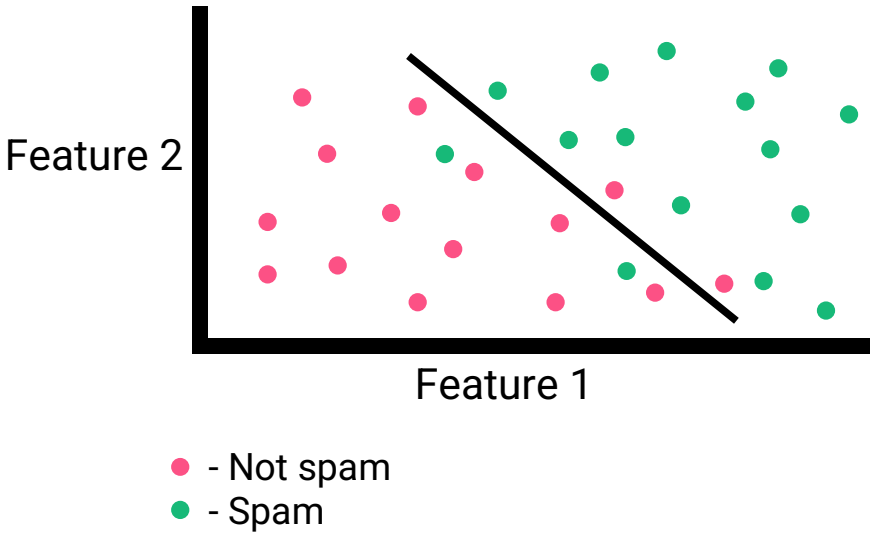
Advantages

- Fast training speed as no training period involved

Disadvantages

- Requires explicitly testing against much training data
- Requires feature scaling to perform better

Logistic Regression



Outputs probability of 0 to 1

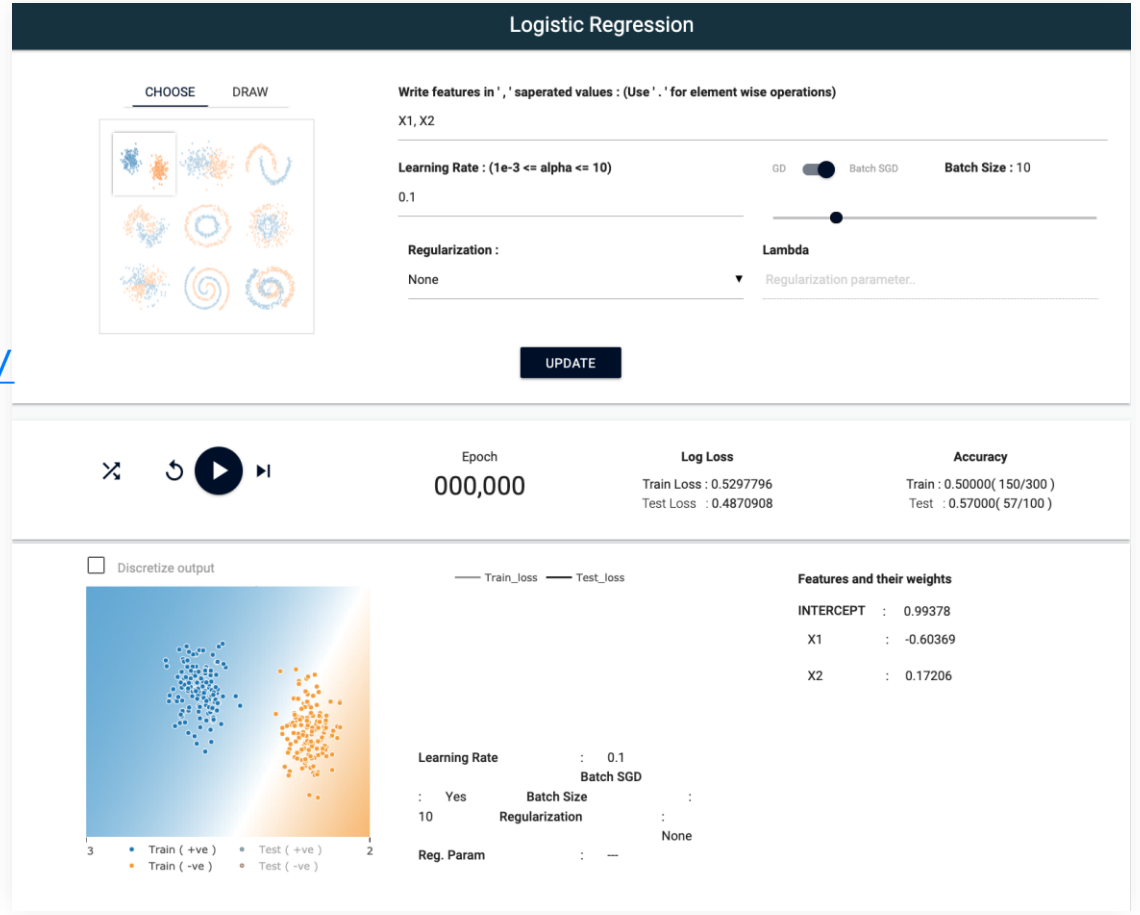
Predicts binary outcome (True / False) based on a set of features

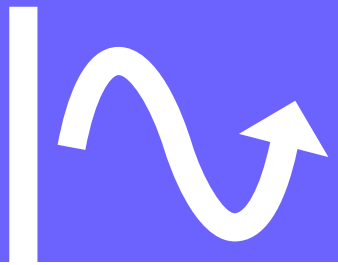
Forms decision boundary that performs classification

Logistic Regression

Interactive Examples:

https://ramsane.github.io/ml-playground/logistic_regression/





Logistic Regression

Advantages

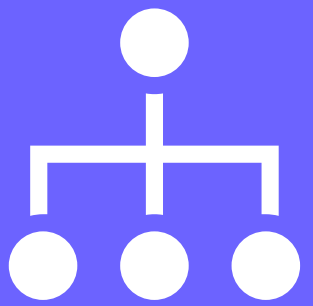
- Interprets coefficient to visualise decision boundaries and as indicators of feature importance
- Performs well when dataset is linearly separable

Disadvantages

- Sensitive to outliers
- Assumes that dataset is linearly separable

4

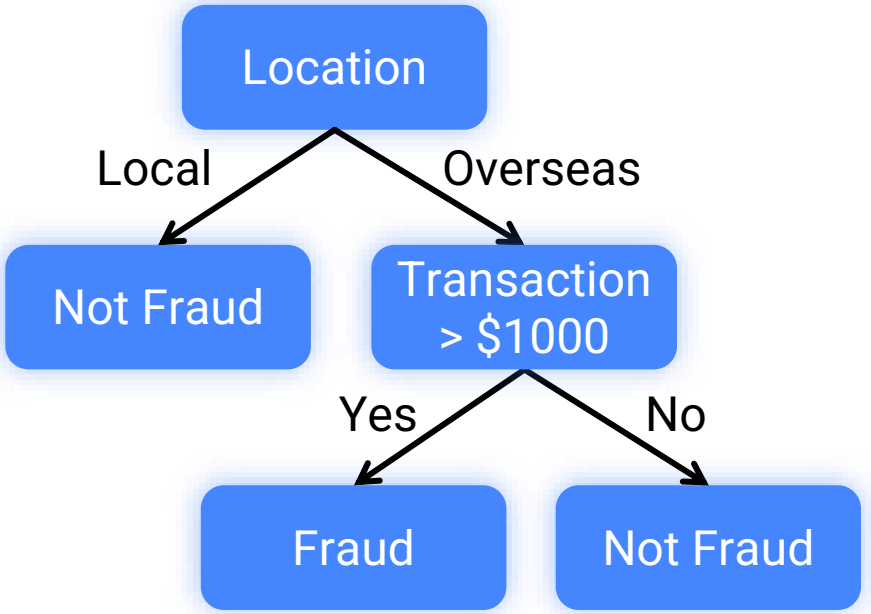
Decision Tree

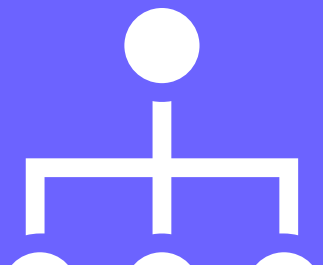


Decision Tree

- Made up of a series of questions to predict target variable
- Can be used for both classification and regression problems

Credit Card Fraud Classification





Decision Tree

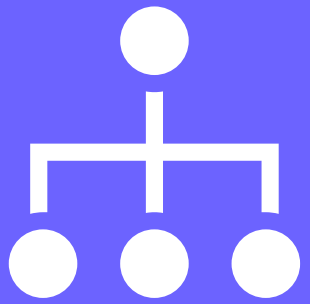
Advantages

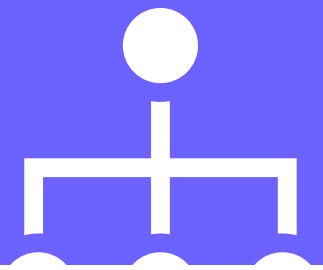
- Requires little data processing
- High Interpretability

Disadvantages

- Prone to overfitting

Random Forest

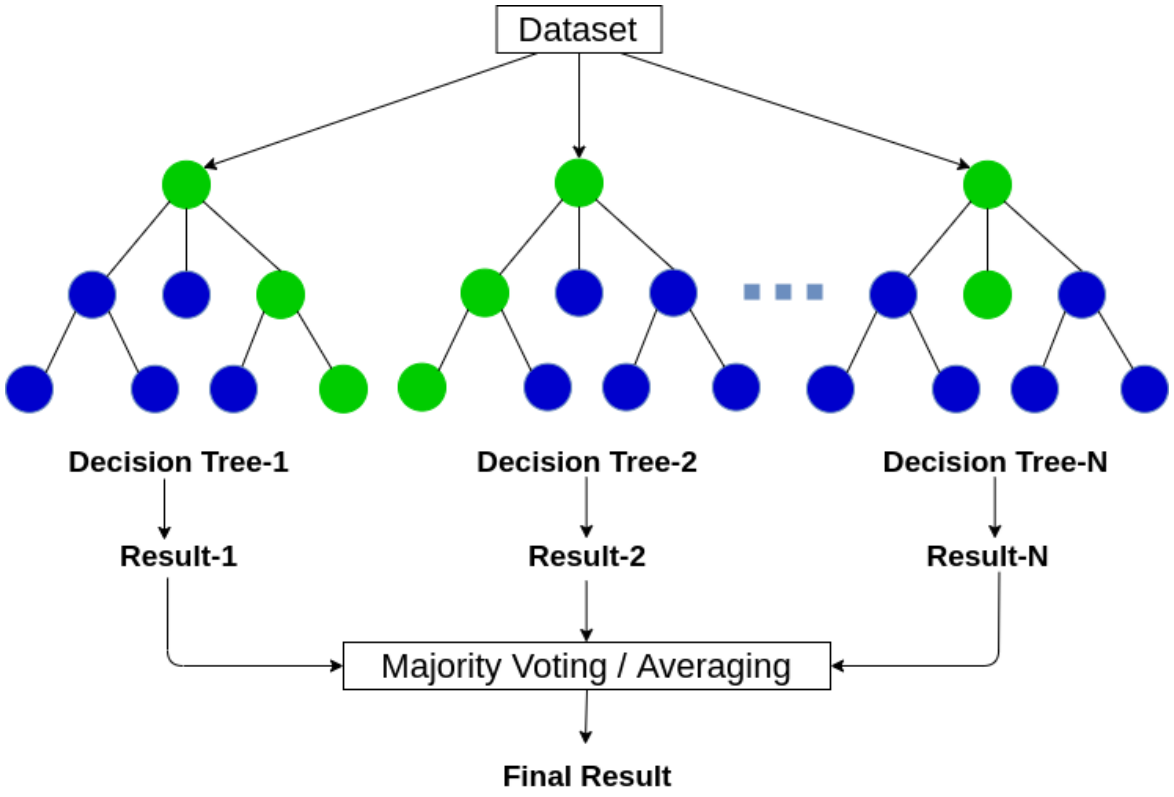


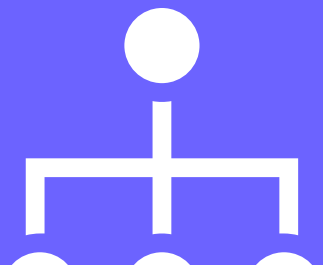


Random Forest

- Consists of multiple decision trees
- Each tree produces their own prediction
- Final prediction made based on average prediction of all decision trees
- Can be used for both classification and regression problems

Random Forest





Random Forest

Advantages

- Retains the high accuracy of a decision tree
- Less prone to overfitting as compared to decision tree

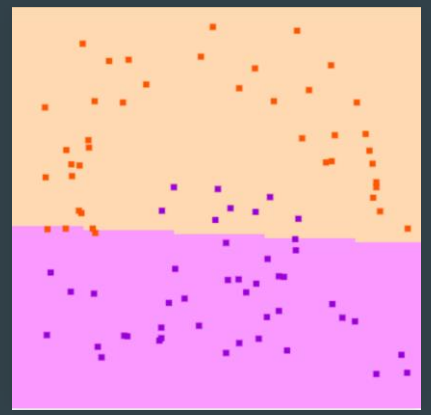
Disadvantages

- Slow training process
- Low interpretability

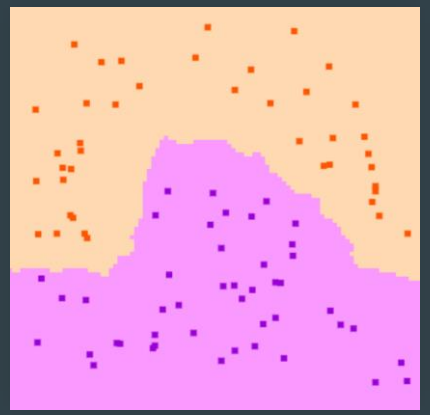
Knowledge Check

Given the following dataset, which Machine Learning model is the best fit to the dataset?

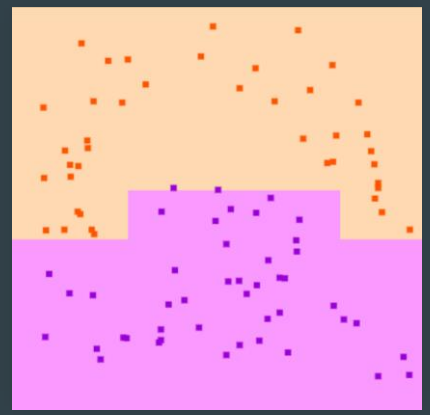
A. Logistic Regression

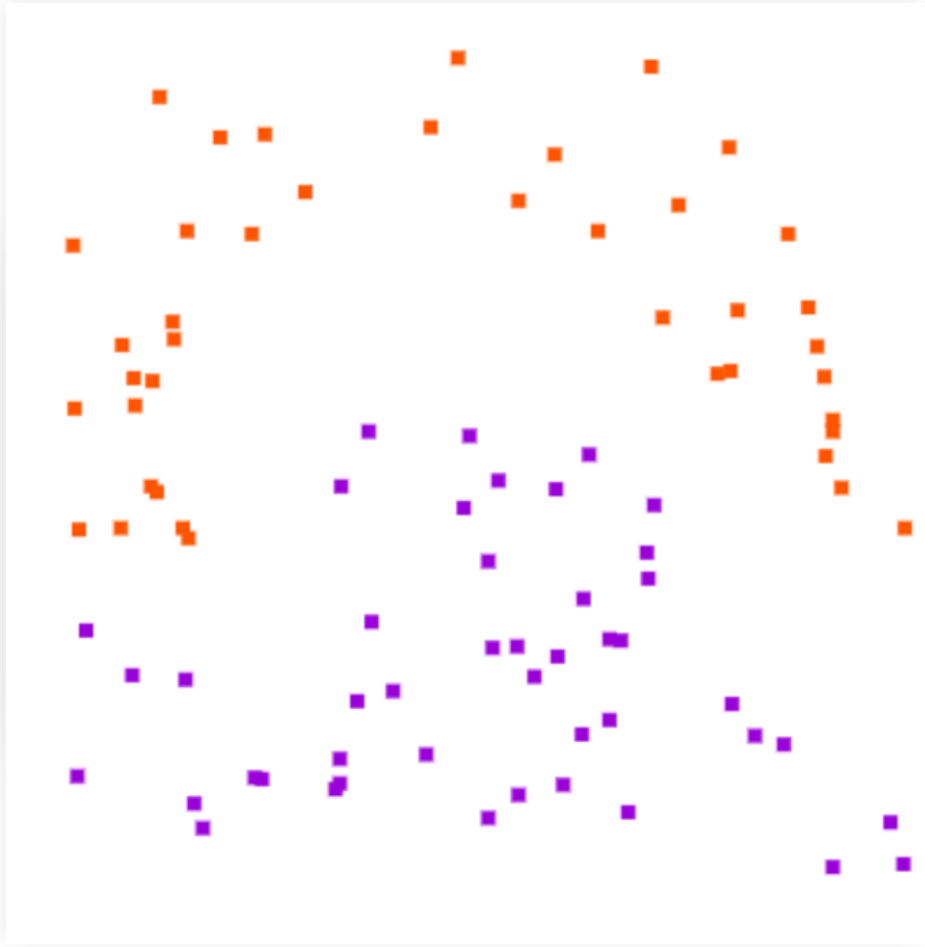


B. K-Nearest Neighbours



C. Decision Tree





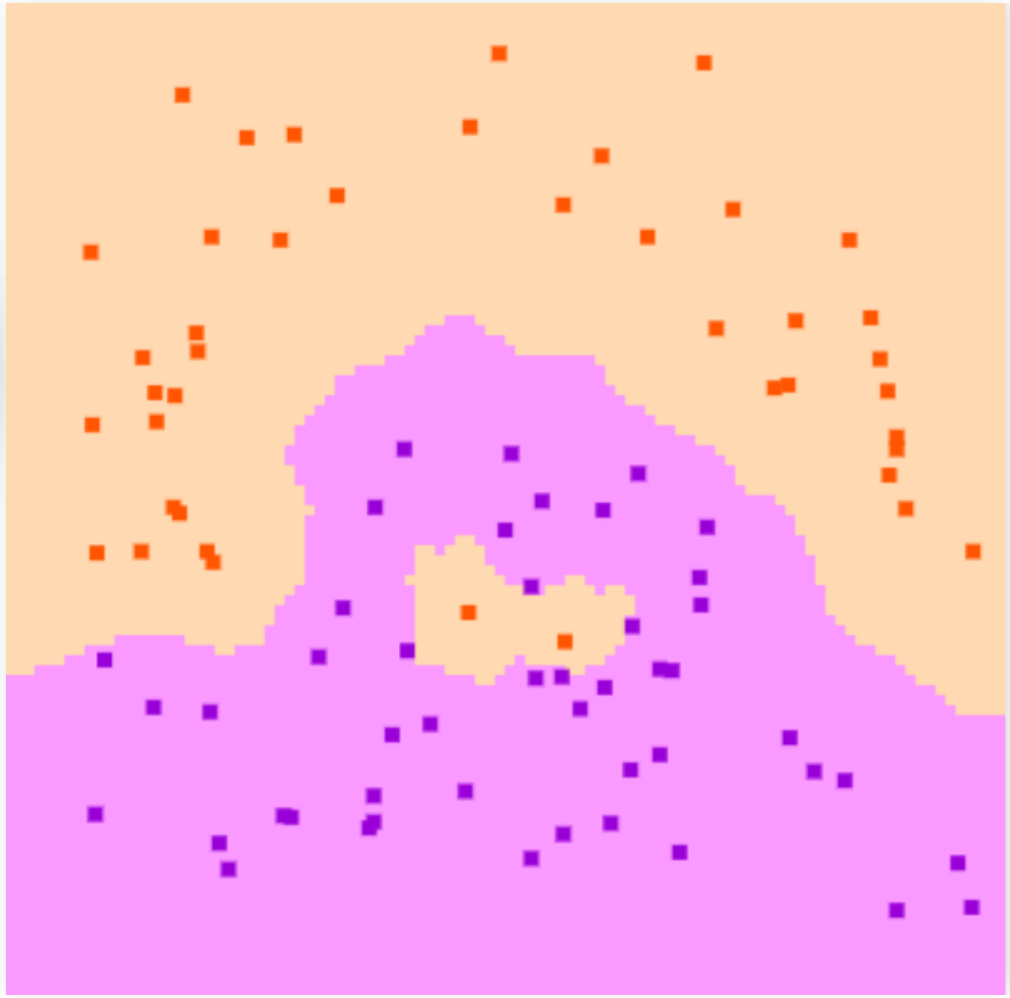
Source: ML-Playground

Knowledge Check

What problem is the model suffering from if I obtain the following decision boundary when I visualise it?

A. Overfitting

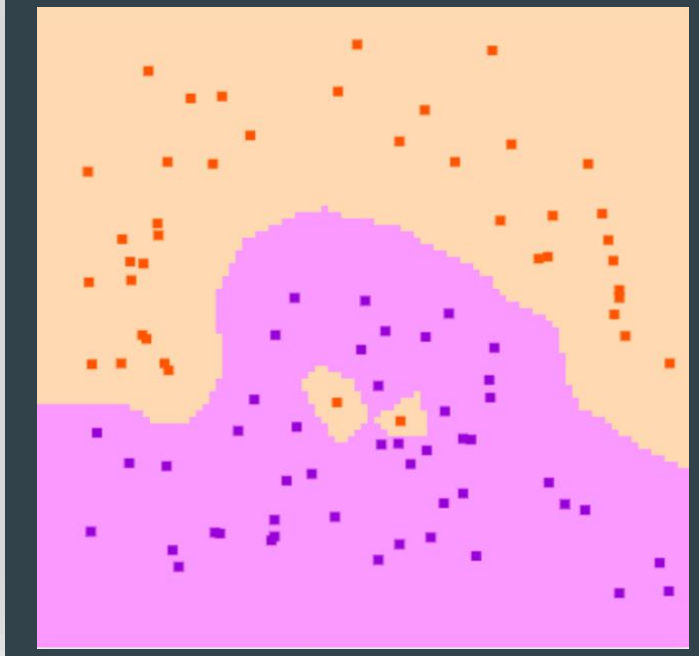
B. Underfitting



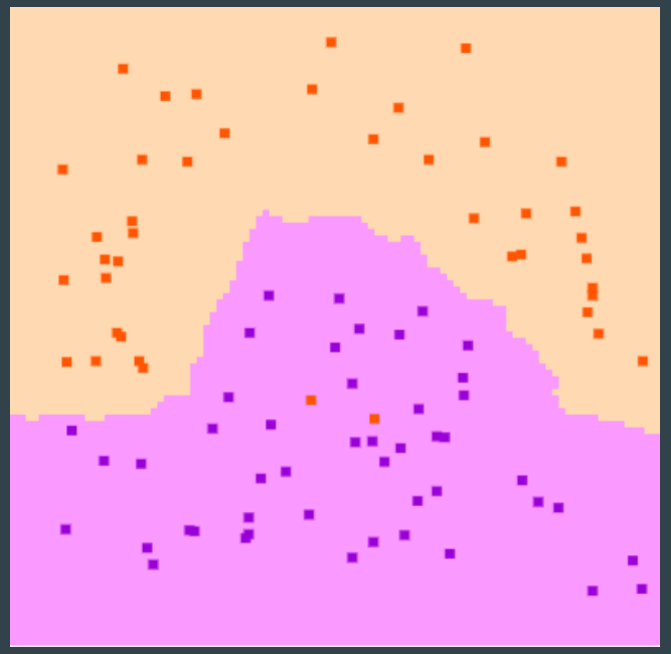
Knowledge Check

What value should I set for hyperparameter K to reduce overfitting?

A. K = 1



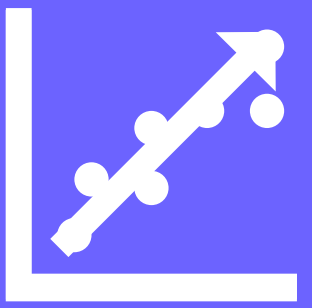
B. K = 5



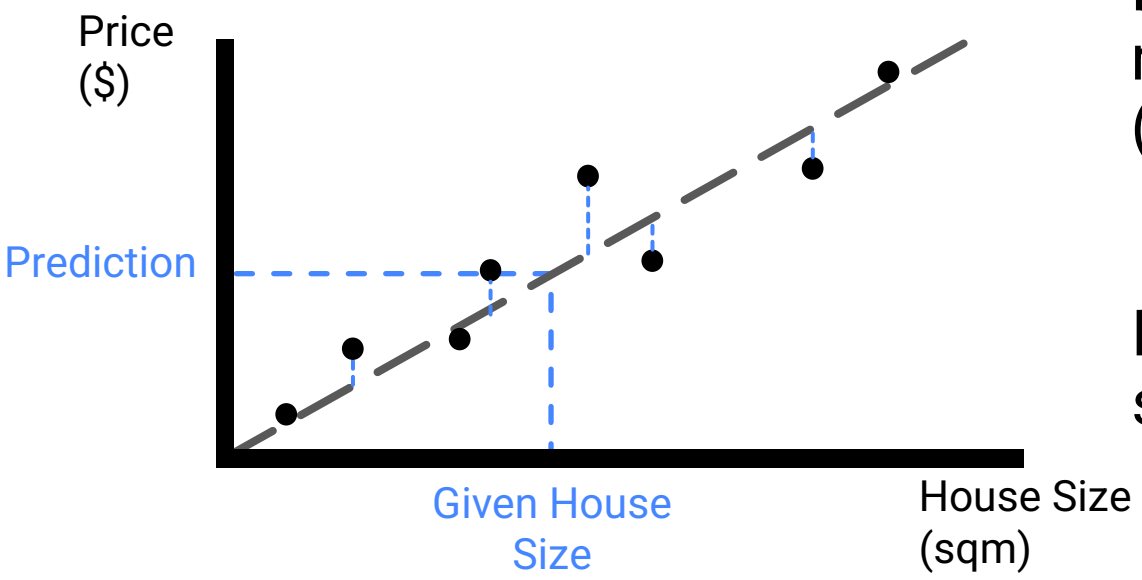
Break and Q&A

SPAI

Regression Model



Linear Regression



Linear approach to model relationships between response (target) and explanatory variables (features)

Fitting a straight line while minimizing the square of errors (Best Fit Line)

Linear Regression

Generate intercept value and list of coefficient for all explanatory variable.

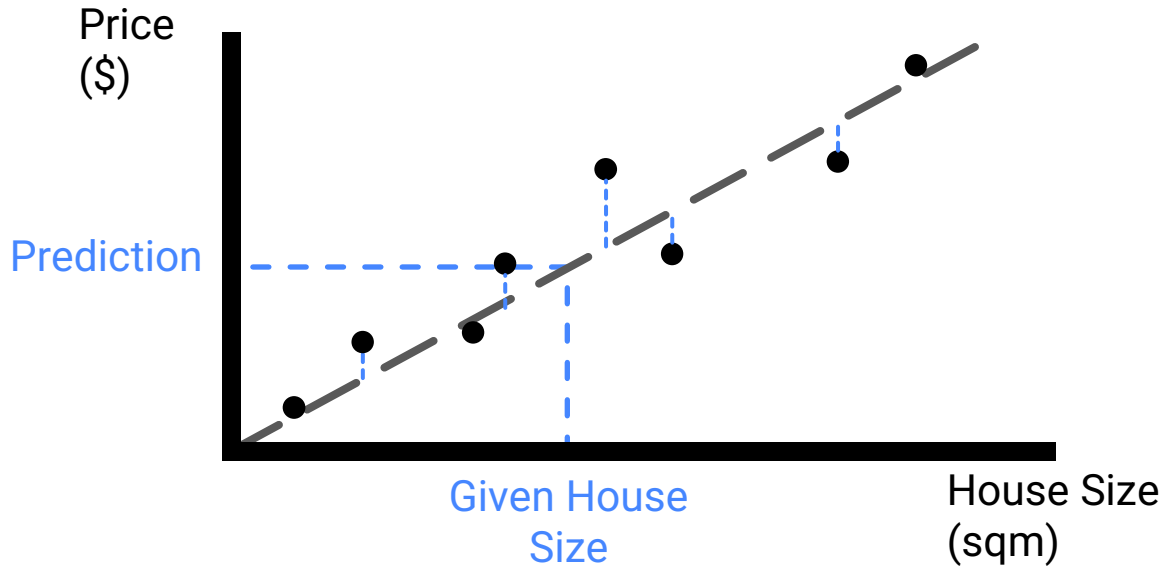
Predicting house by knowing house size: Get prediction by projecting data point over line.

Intercept Value

$$y = \theta_0 + \theta_1 x$$

Response variable

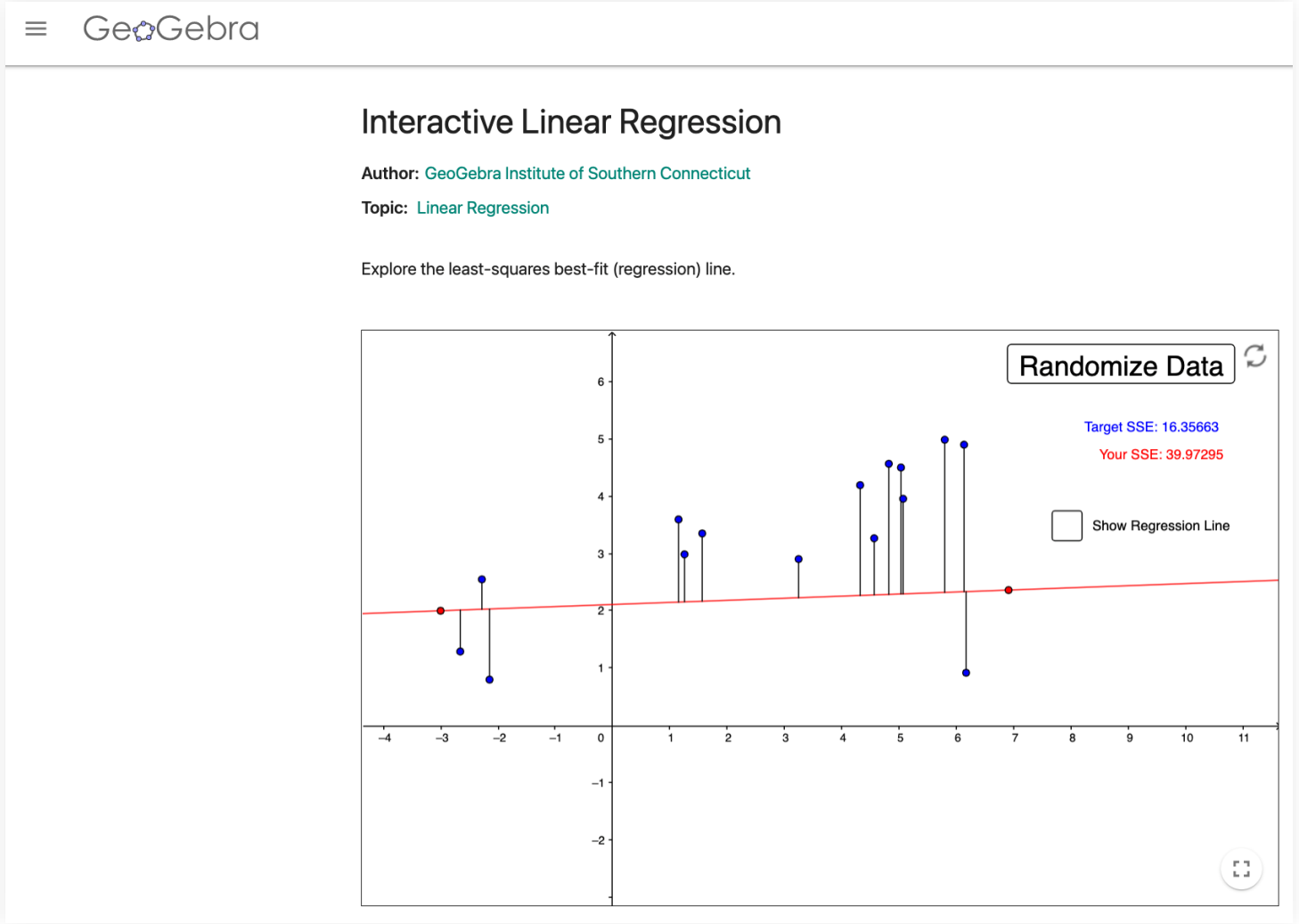
Coefficient(s)



Linear Regression

Interactive Examples:

https://www.geogebra.org/m/xC6zq7Zv



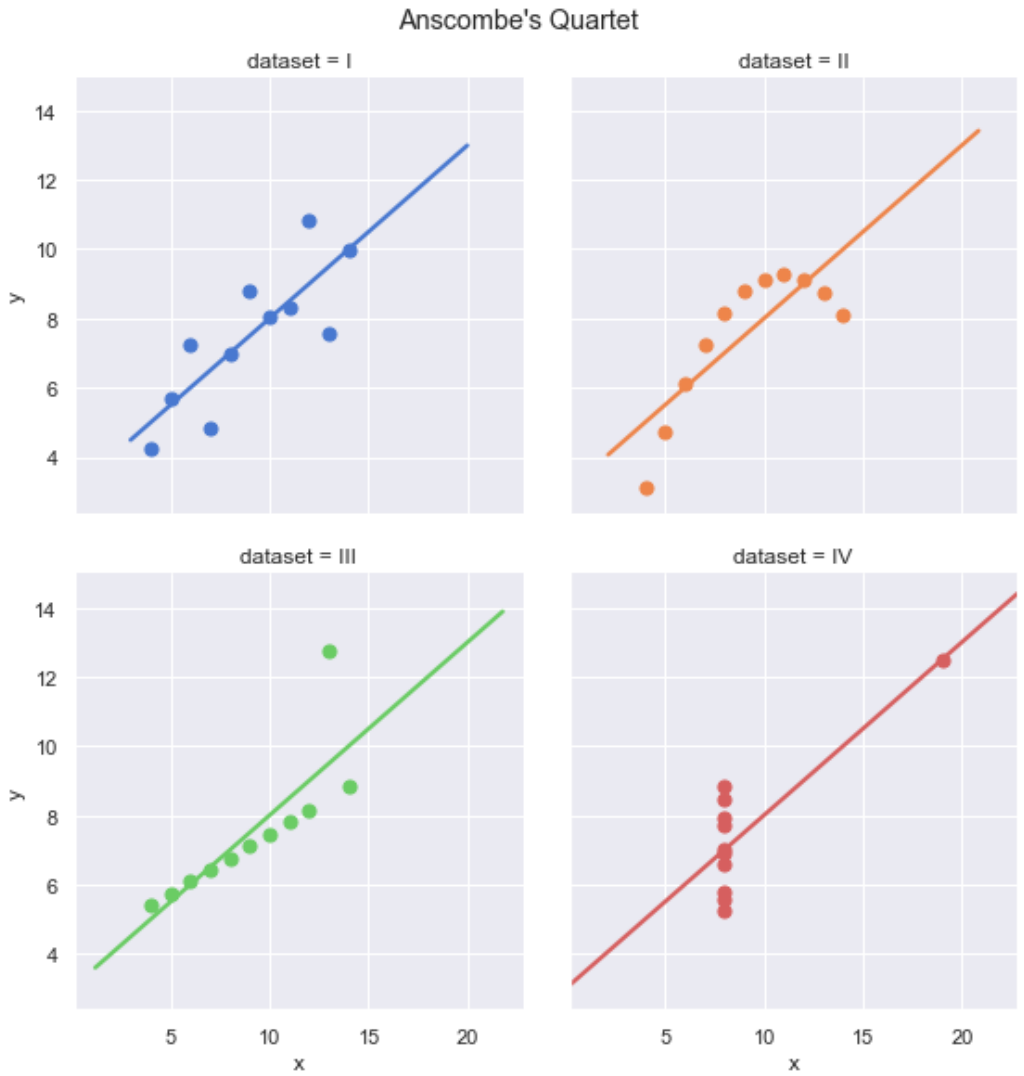
Linear Regression

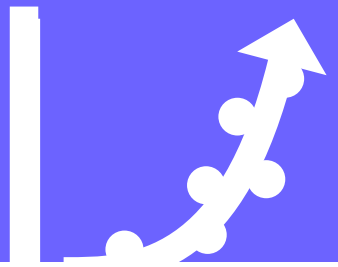
Advantages

- High Interpretability

Disadvantages

- Prone to outliers
- Linearity Assumption





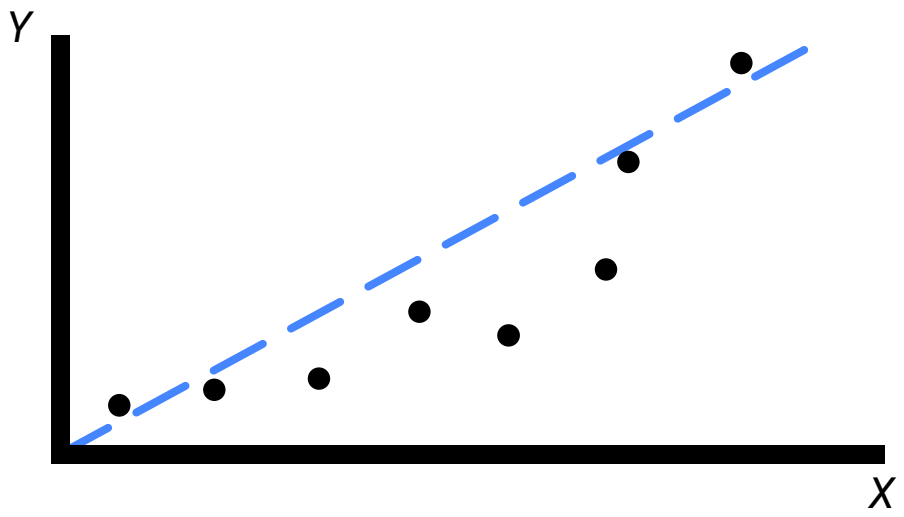
Polynomial Regression

- Relationship is modelled as an (n)th degree
- Better fit than Linear Regression (degree = 1) when relationship between features and target variable is non-linear

Polynomial Regression

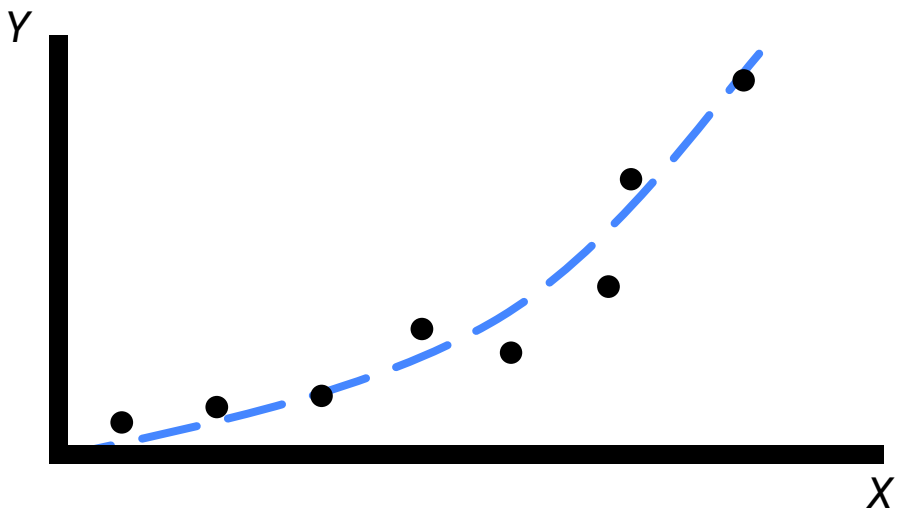
Linear Model

$$y = \theta_0 + \theta_1 x$$



Polynomial Model

$$y = \theta_0 + \theta_1 x + \theta_2 x^2$$

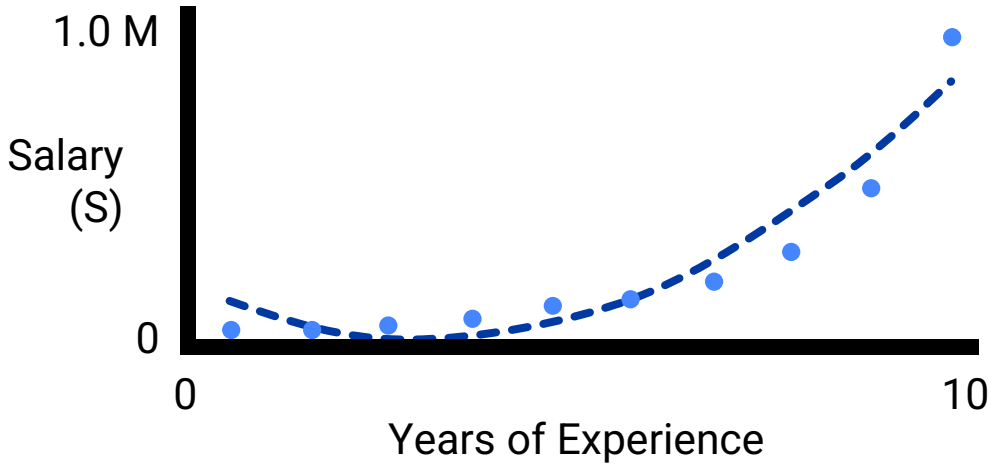


SPAI

Polynomial Regression

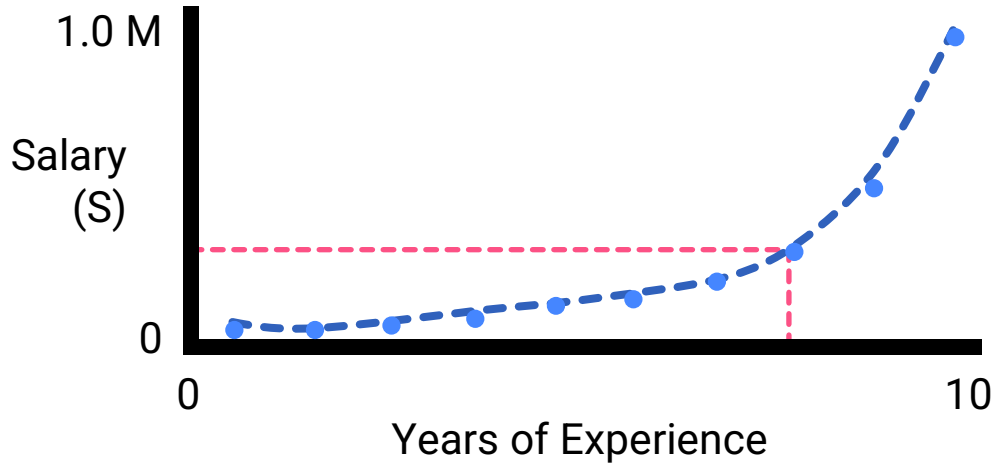
Degree = 2

$$y = \theta_0 + \theta_1 x + \theta_2 x^2$$



Degree = 4

$$y = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$$



SPAI

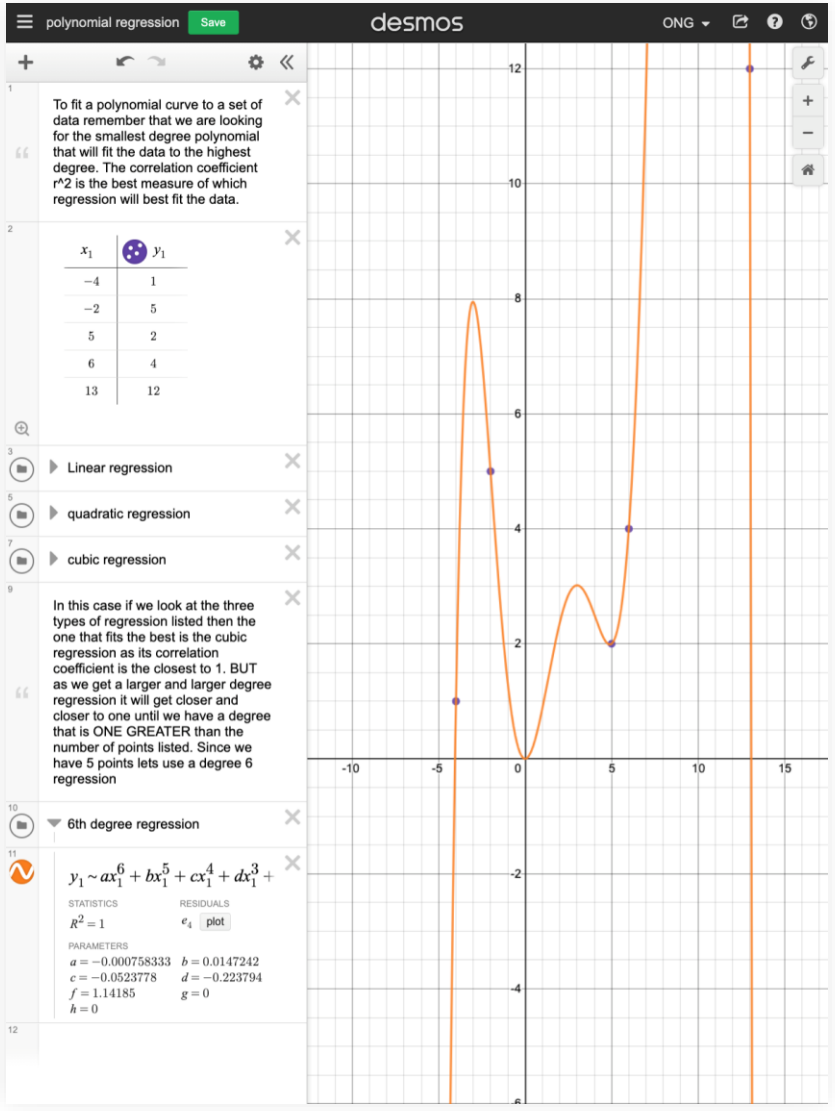
Polynomial Regression

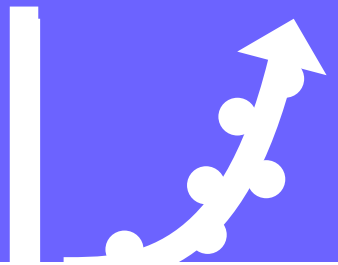
- The plot shows prediction of salary based on “Years of experience”
- In this case, 4^{th} degree is a better fit than 2^{nd} degree, as more of the points fall near regression line
- If we want to predict the salary of worker in their 8^{th} year, the predicted salary would be 289,994 dollars per year

Polynomial Regression

Interactive Examples:

https://www.desmos.com/calculator/wdb45brrj8





Polynomial Regression

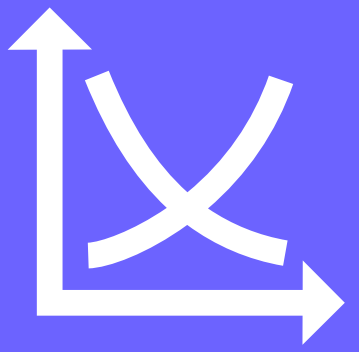
Advantages

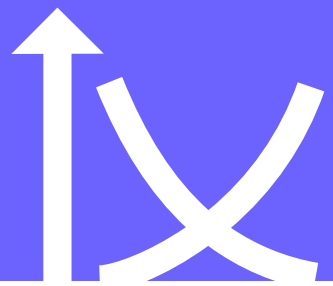
- Can represent Non-linear relationship between features and target variable by changing degree

Disadvantages

- Prone to overfitting when degree is too high

Overfitting VS Underfitting

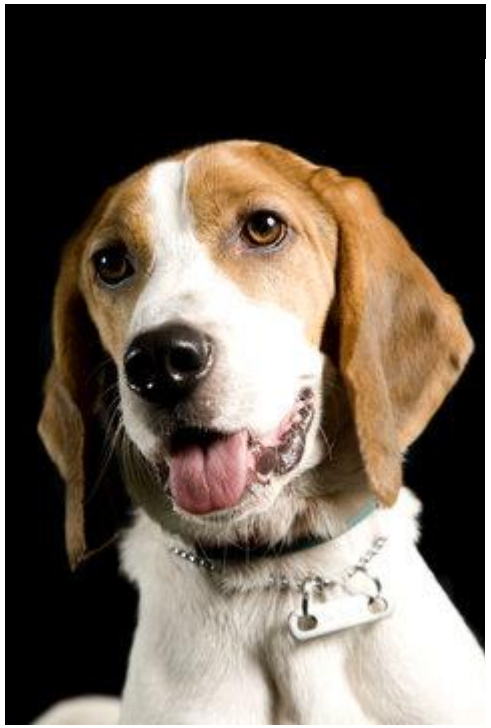




Overfitting

- Model that fits to training data too well instead of learning general distribution
- Example: If model is trained with dog images, overfitting happens when model learns meaningless details and noise in training data
- Results in model very good at predicting pictures it has seen only
- Not able to classify new pictures

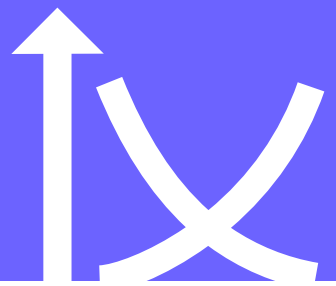
Example



Train Data

Model

Test Data



Underfitting

- Underfitting occurs when model performs poorly on training data
- Example: If one only studies addition mathematical operation, he/she would only be able to do questions related to addition
- Additionally, he/she would fail to answer questions related to other mathematical operations.
- Model is underfitting when it is not complex enough to accurately capture relationships between dataset’s features and target variable

“No Free Lunch”

- ⚠ Best machine learning algorithm does not exist
- ☑ Different models have individual assumptions that are better or worse for different data.
- ☑ Try different models and evaluate performance of model to decide which model to use.

Lunch Time

SPAI

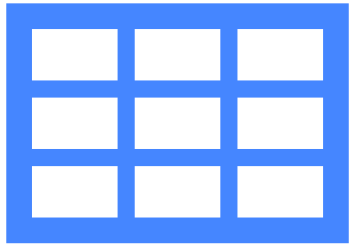
Model Building and Evaluation

Problem Recap

- Titanic Classification; Titanic shipwreck
- Women, Children and the Upper-class were more likely to survive the shipwreck

Problem Recap

- Task: let machine learn and classify whether a passenger did survive the shipwreck (target variable, *y*) with information given (features, *X*)
- We have performed basic Exploratory Data Analysis (EDA) and Data Cleaning towards raw dataset.
- We will continue to work towards building and evaluation of machine learning models with Scikit-Learn library in Python



Titanic Variables Table

Variable	Definition	Key
Survival	Survival	0 = No, 1 = Yes
Pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
Sex	Sex	
Age	Age in years	
SibSp	# of siblings / spouses aboard the Titanic	
Parch	# of parents / children aboard the Titanic	
Ticket	Ticket number	
Fare	Passenger fare	
Cabin	Cabin number	
Embarked	Port of Embarkation	C = Cherbourg, Q = Queenstown, S = Southampton

Extract Features and Target Variable

In Supervised Learning, our task is to predict target variable (*y*) based on features (*X*)

	pclass	age	sibsp	parch	fare	sex_male	embarked_Q	embarked_S
0	1.0	-0.039005	-0.479087	-0.445000	3.442584	0	0	1
1	1.0	-2.215952	0.481288	1.866526	2.286639	1	0	1
2	1.0	-2.131977	0.481288	1.866526	2.286639	0	0	1
3	1.0	0.038512	0.481288	1.866526	2.286639	1	0	1
4	1.0	-0.349075	0.481288	1.866526	2.286639	0	0	1
...	...	...	...	...	...	...	...	...
1304	3.0	-1.163009	0.481288	-0.445000	-0.364003	0	0	0
1305	3.0	-0.116523	0.481288	-0.445000	-0.364003	0	0	0
1306	3.0	-0.232799	-0.479087	-0.445000	-0.503774	1	0	0
1307	3.0	-0.194040	-0.479087	-0.445000	-0.503774	1	0	0
1308	3.0	-0.039005	-0.479087	-0.445000	-0.491207	1	0	1

1309 rows × 8 columns

Feature, *X*

Model

	survived
0	1.0
1	1.0
2	0.0
3	0.0
4	0.0
...	...
1304	0.0
1305	0.0
1306	0.0
1307	0.0
1308	0.0

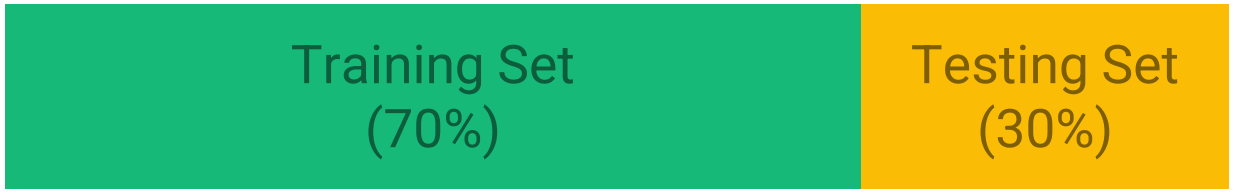
1309 rows × 1 columns

Target Variable, *y*

Train Test Split

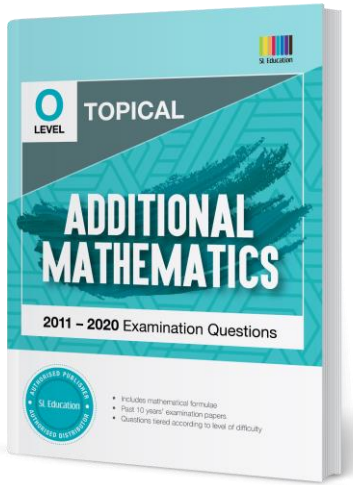
Motive:

- Train your model with training set and test performance of it with testing set which has data that hasn’t been seen by model
- Reason: to evaluate how well your model generalizes to data that hasn’t been seen by model.
- Common Ratio of splitting data for train set and test set is 7:3 and 8:2



Train Test Split

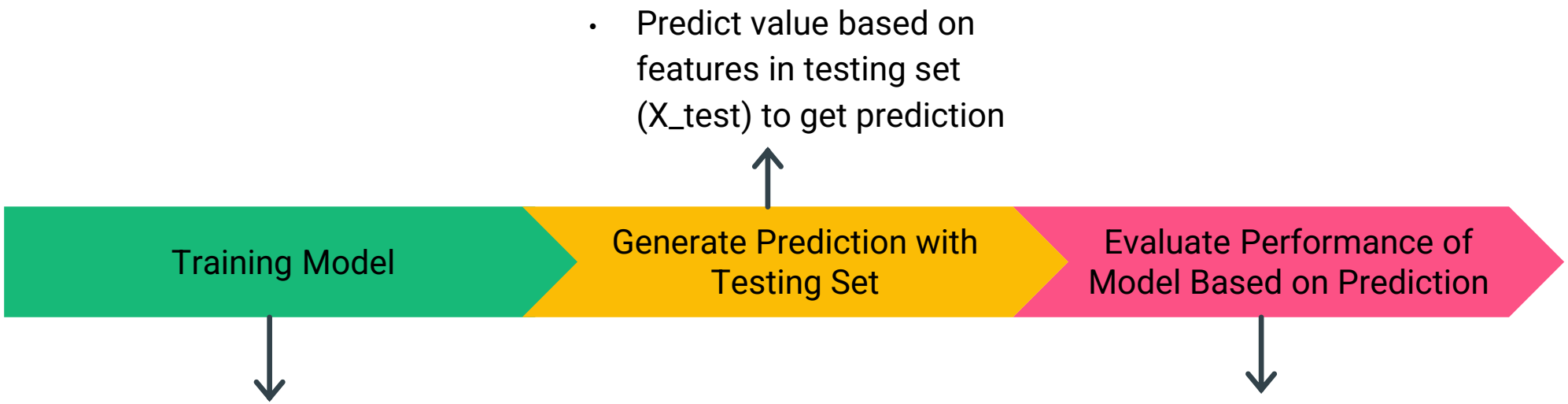
Examination Analogy:



- [x] Aim: To know performance of student in certain subject
- [x] Model: Student's brain
- [x] Training Set: Past Year Paper to train the student
- [x] Testing Set: Exam Paper to test performance and understanding of model (Student)
- [x] Testing Set separated from Training Set as we wouldn't want to leak exam paper questions to exercise paper.

Training and Evaluation

After obtaining training and testing set from train-test-split, begin to train model based on training set and evaluate evaluate performance



Training and Evaluation

5 Steps of Model Fitting and prediction with Scikit-Learn Library:

1. Import model class from Sklearn library
2. Initiate class with hyperparameters and store the instance at a variable
3. Call .fit() method to train the model by parsing X_train and y_train (training features and training answers)
4. Call .predict() method with the trained model to make prediction on testing features, X_test
5. Evaluate the performance of model by comparing the prediction with ground truth, y_test

Knowledge Check

Which of the following is the reason that we would want to have a testing set?

A. You want to maximize the amount of training data used.

B. You want to absolutely be certain about your model's ability to generalize to unseen data.

C. You want to tune the hyperparameters of your model.

Source : DataCamp supervised-learning-with-scikit-learn/fine-tuning-your-model?ex=13

Knowledge Check

Which of the following is NOT within the 5 step of Model Fitting and Prediction with Scikit-Learn Library?

A. Import the desired Model Classes from Sklearn library

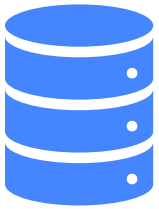
B. Perform cross validation with the model

C. Use .fit() to train the model by parsing in the X_train and y_train

Problem with imbalanced Classes

Tumour Analogy:

Problem statement: evaluating if a tumour is harmful or subsiding



Dataset: 99% Benign(Good) Tumour, 1% Malignant(Bad) Tumour

Problem with imbalanced Classes

If model is trained with such dataset, the model might not have enough harmful examples to learn from

Hence model might predict benign all the time

Although model still achieves high accuracy of 99% as 99% of data are benign examples

Model would still be bad at predicting harmful tumours

Therefore it is important to have suitable dataset to know true performance of model

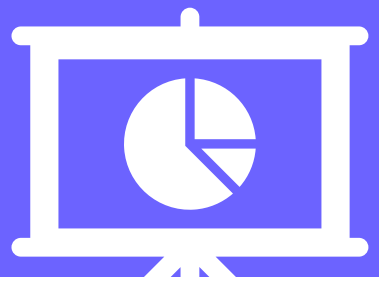
Problem with imbalanced Classes

Unequal distribution of target variable

```
y.values_count()
>>
0 : 9900
1 : 100
Name : Is_cancer, dtype: int64
```

Problem with imbalanced Classes

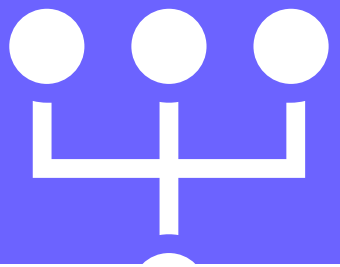
- Other Methods/Advance Algorithms to deal with imbalance classes:
  - Random Downsample : Randomly removing records from majority class
  - Random Upsample : Randomly duplicating observation from minority class
  - Advance Algorithm from imbalanced-learn library (e.g. SMOTE, ClusterCentroids, TomekLink)



Model Evaluation

Motive:

- [x] Understand performance of model
- [x] Creating single value metric for decision making purposes

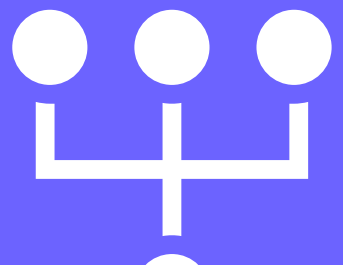


Evaluation Metric for Classification problem

Accuracy Score:

- Proportion of correct prediction rows divided by total number of rows in dataset
- Prone to class imbalance problem in dataset

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$



Evaluation Metric for Classification problem

Confusion Matrix:

- Illustrate classifier performance based on matrix comprises 4 elements:

- True Positive(TP) : Label is 1 and predicted value is 1
  True Negative(TN) : Label is 0 and predicted value is 0
  False Positive(FP) : Label is 0 but predicted value is 1
  False Negative(FN) : Label is 1 but predicted value is 0

Evaluation Metric for Classification problem

		Actual values: Positive	Actual values: Negative
Predicted Values	Positive	TP	FP
Predicted Values	Negative	FN	TN

		Actual values: Cancer	Actual values: No Cancer
Predicted Values	Cancer	45	18
Predicted Values	No Cancer	12	25

Evaluation Metric for Classification problem

- F1-Score: Harmonic mean of precision and recall

$$\text{f1score} = \frac{1 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}}$$

- Precision: Percent of true positive predictions

$$\text{precision} = \frac{TP}{TP + FP}$$

- Recall: Percentage of correctly classified positive values

$$\text{recall} = \frac{TP}{TP + FN}$$

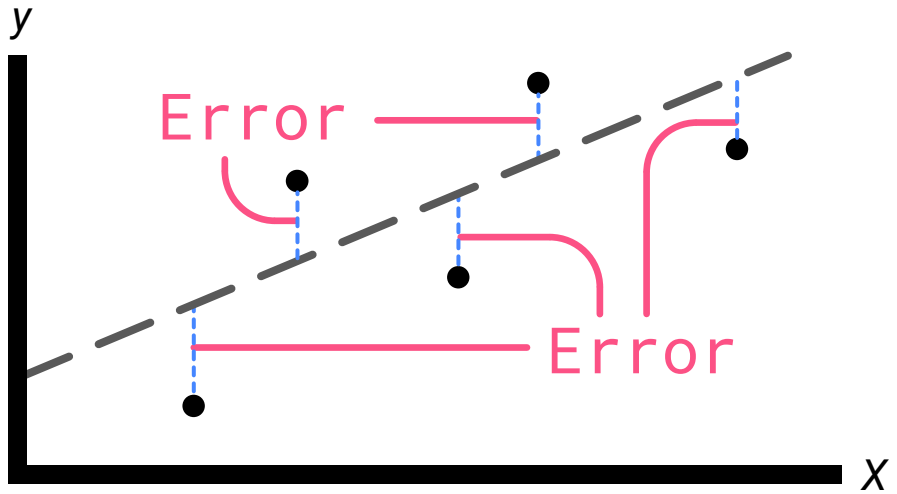
Evaluation Metric for Regression Problem

Mean Squared Error (MSE)

- Squared difference between predicted values and actual values
- Lower error means better fitted model

$$\text{Squared Error} = (y - \hat{y})^2$$

$$\text{Absolute Error} = |y - \hat{y}|$$



Knowledge Check

Why should we not always use Accuracy Score as evaluation metrics for every classification problem?

A. Accuracy Score can only be used in two-class classification problem (Predicting YES or NO only)

B. Accuracy Score might subject to biased number when we are facing class imbalance problem (99% of data belongs to a single class)

Knowledge Check

Which evaluation metrics should we use when we are dealing with imbalanced class problem?

A. Accuracy Score

B. Precision

C. F1 Score

D. Mean Squared Error

Scan to mark attendance

Scan the QR code to check out

Check Out

SPAI