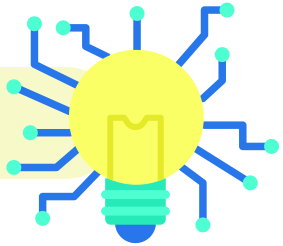
Case Study 2

Era Kalaja

My Process

The Objective



Classification Task - Predict Location

ML Model Performance

- Feature engineering
- Find best performing model

More Data or Less Data

- Split training set
- Does high accuracy = more data?

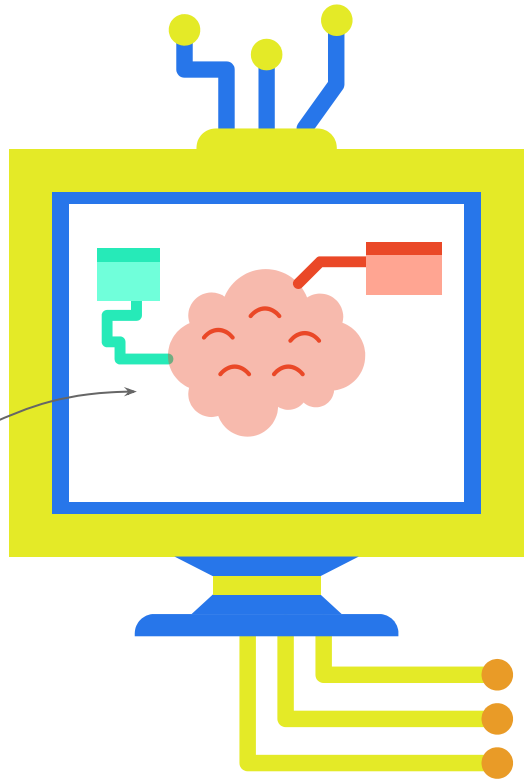
The Data

Feature Engineering

100’s changed to - 105

- **Features:**

  WAP 1-520



Target Manipulation

Combined Building & Floor

- **Target:**

  BuildingID_Floor

3 Machine Learning Models

01

Kth Nearest Neighbor

- Flexible data types
- Based on local structure & similarity

02

Random Forest Classifier

- Manages overfitting
- Not every feature is considered in trees

03

Support Vector Classifier

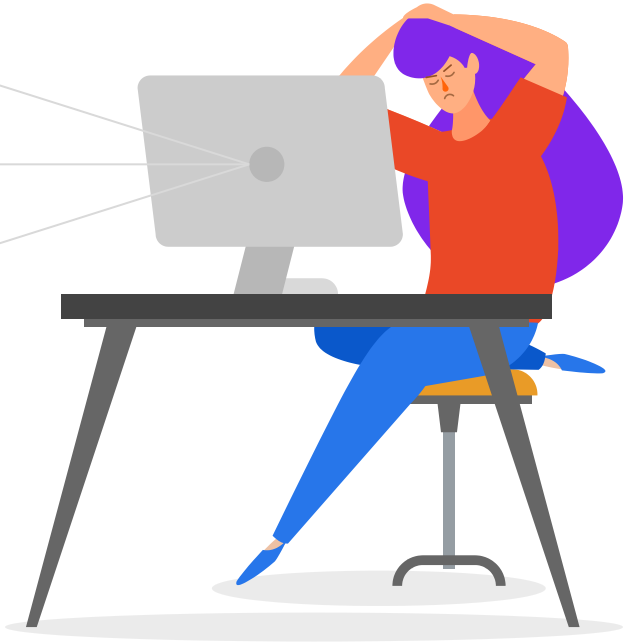
- Linear separation
- Manages high dimension

Challenges

Working with large dataset 01

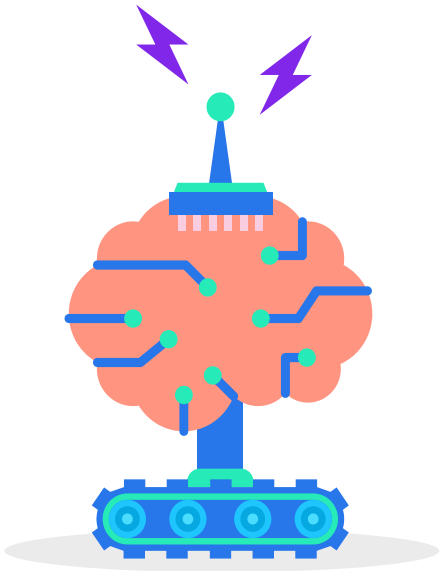
Scikit Learn Documentation 02

Cross Validation 03



My Findings

Best Model

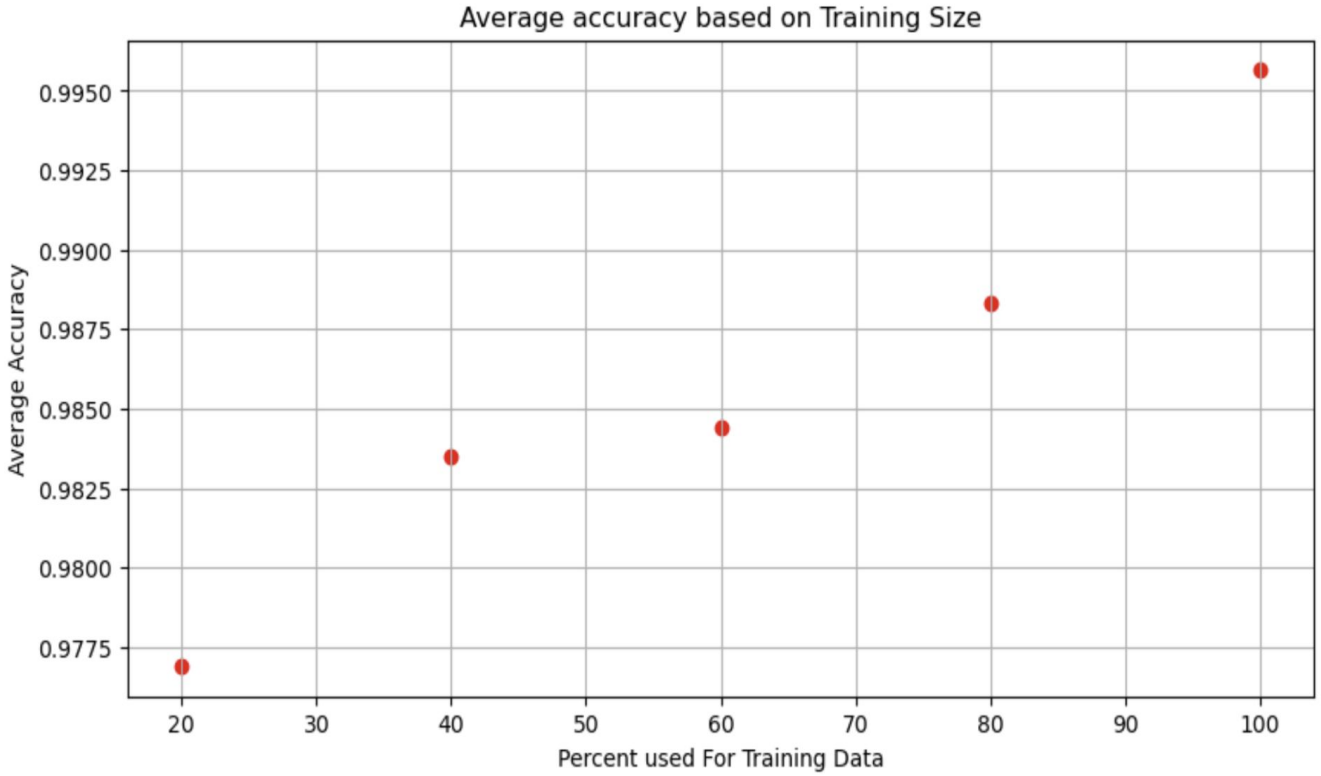


Support Vector Classifier

Highest Accuracy - 0.99

Easiest Implementation

More Data Is Better ... Sometimes



- Higher accuracy
- Overfitting with 100%
- 80% = Best Option

Thank you.
Questions?