Feature Engineering

Week 7

Table of Contents

Lifecycle


01
BUSINESS UNDERSTANDING
Ask relevant questions and define objectives for the problem that needs to be tackled.
02
DATA MINING
Gather and scrape the data necessary for the project.
03
DATA CLEANING
Fix the inconsistencies within the data and handle the missing values.
04
DATA EXPLORATION
Form hypotheses about your defined problem by visually analyzing the data.
05
FEATURE ENGINEERING
Select important features and construct more meaningful ones using the raw data that you have.
06
PREDICTIVE MODELING
Train machine learning models, evaluate their performance, and use them to make predic-tions.
07
DATA VISUALIZATION
Communicate the findings with key stakeholders using plots and interactive visualizations.
DATA SCIENCE LIFECYCLE
sudeep.co

Feature Engineering

Different algorithms have distinct requirements in order for us to fit them to our data. Some require:

- Only numerical values
- Only categorical values
- No null values

And more...

We just need to make the correct transformations in order to make our data suitable for specific/different algorithms.

Feature Engineering

- So far we have studied **KNN Classifier** and **Regressor** as a predictive model.
- Because KNN is a distance based algorithm, it requires all entries in our data to be **numerical**.
- On the other hand, some models, like Naive-Bayes (yet to study) requires all entries to be **categorical**.

Let's explore some techniques to transform all data into numerical or categorical.

Feature Engineering - One Hot Encoding

Recap - Nominal Data

- One-hot encoding converts **categorical** variables into **numerical** (binary vectors) where each category is represented by a single bit, indicating its presence.

Original Data

Team	Points
A	25
A	12
B	15
B	14
B	19
B	23
C	25
C	29

→

One-Hot Encoded Data

Team_A	Team_B	Team_C	Points
1	0	0	25
1	0	0	12
0	1	0	15
0	1	0	14
0	1	0	19
0	1	0	23
0	0	1	25
0	0	1	29

Feature Engineering - Label Encoding

Recap - Ordinal Data

- **Label encoding** for ordinal data assigns numerical values to categories **based on their order or ranking**.

Height
Tall
Medium
Short

Height
0
1
2

Feature Engineering - Binning

Recap

- Grouping a **continuous** variable into intervals can make analysis more intuitive and can highlight patterns better in some cases.

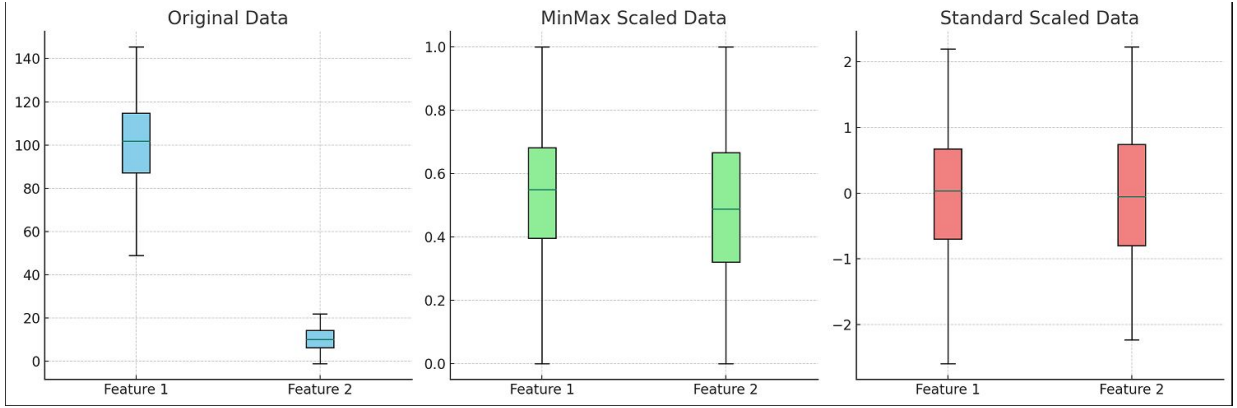
Sex	Age
male	22
female	38
female	26
female	35
male	35
male	80
male	54
male	2
female	27
female	14
female	4
female	58

Sex	Age
male	Adult
female	Adult
female	Adult
female	Adult
male	Adult
male	Elderly
male	Adult
male	Toddler/baby
female	Adult
female	Child
female	Toddler/baby
female	Adult

Feature Scaling

Feature Scaling

- **Feature scaling** is the process of adjusting the range of features in a dataset to make sure they're all on a **similar scale**, which might help improve machine learning algorithms.



Feature Scaling

Normalization - MinMaxScaler

- **MinMaxScaler** is a **normalization** technique that scales features to a specified range, between 0 and 1, preserving the relationship between data points.

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Feature Scaling

Standardization - Z-score

- **Z-score** is a way to standardize all your data in a way that tells you how many standard deviations each point is from the mean.

Specific value

Mean

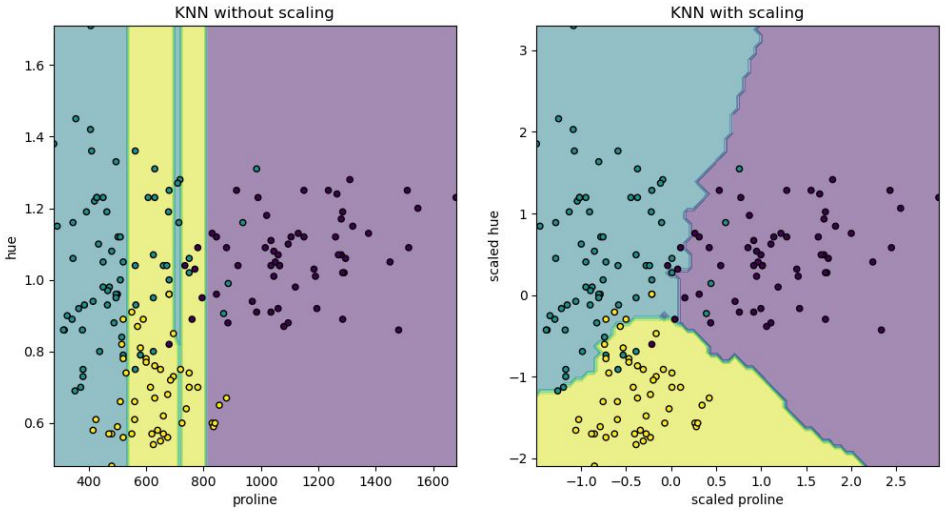
$$x_{std}^{(i)} = \frac{x^i - \mu_x}{\sigma_x}$$

Standard deviation

Feature Scaling

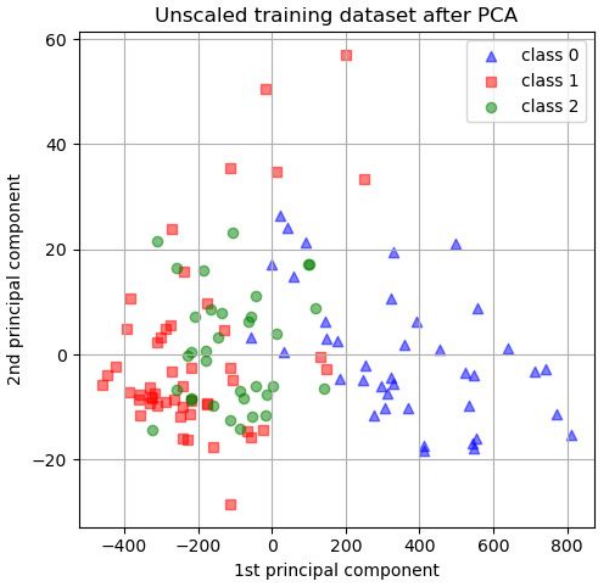
Examples

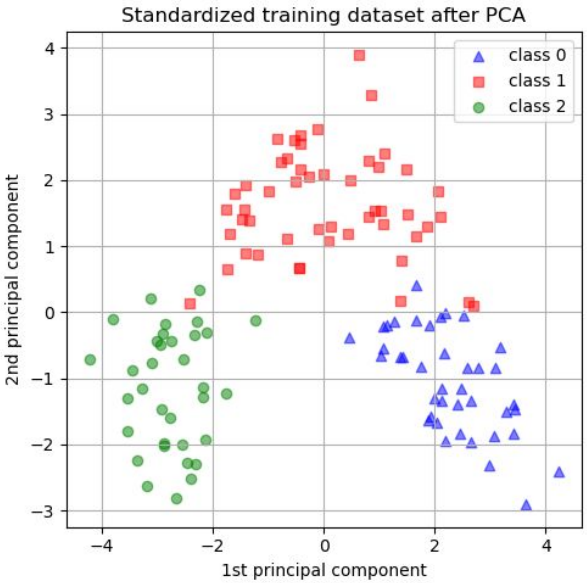
- **Normalizing** or **Standardizing** your data can improve your models by a lot, specially when those models are distance based.



Feature Scaling

Examples





Feature Selection

Feature Selection

- In ML we want **features** to be **highly correlated** with the **target**, but not between themselves.
- High correlation among features themselves, can lead to **redundancy** and instability in models, potentially degrading performance.

Feature Selection

Recap

- When dealing with **categorical** data, we can check if two features are correlated with Chi-square test.

- For **numerical** data, computing correlation matrix will allow you to see the relationship between **features between themselves** and also with the **target**.

	Apple	Samsung	Google
Youth	40	25	5
Middle-Aged	15	30	5
Seniors	10	20	10