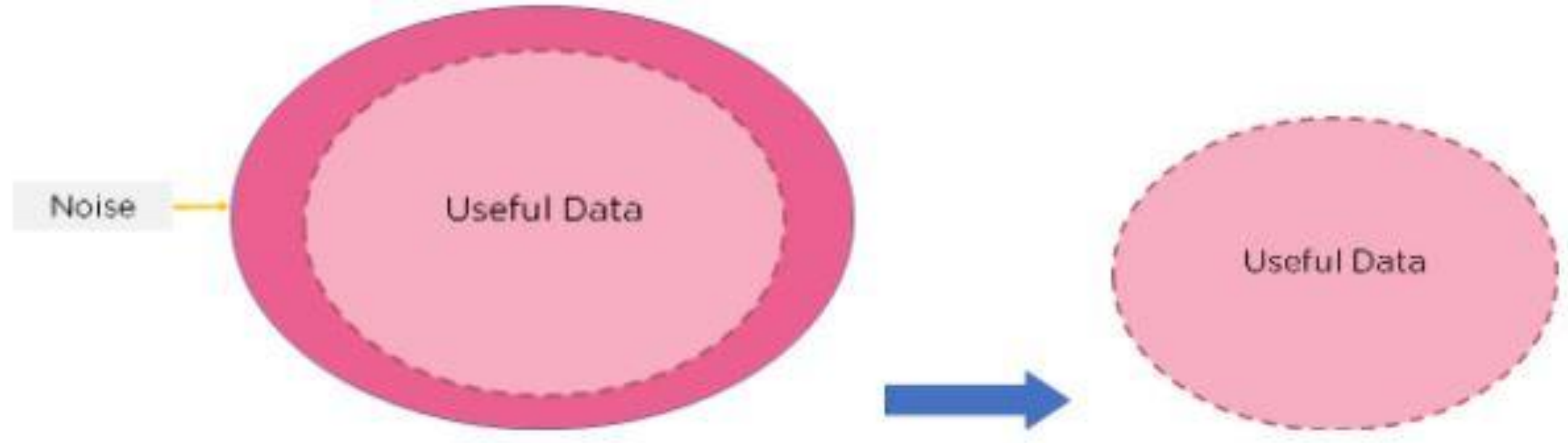
Ways of reducing feature selection with reference to the LEGO Use-Case

Problem: Karpaga Priyaa Velayutham, Patrick Rodriguez Granda
Solution: Juan Pablo Vargas Rodríguez, Diego Arnoldo Azuela Rosas
Reflection: Nina Sepúlveda Conde, René Francisco Basañez Córdoba

FHWS University of Applied Sciences Würzburg-Schweinfurt | Faculty Business and Engineering | https://fwi.fhws.de
TECNOLÓGICO DE MONTERREY | Faculty Industrial and Systems Engineering

PROBLEM

Feature selection is the process of identifying and selecting relevant features for your sample.It is the process of automatically choosing relevant features for your machine learning model based on the type of problem you are trying to solve. We do this by including or excluding important features without changing them. It helps in cutting down the noise in our data and reducing the size of our input data.



Feature importance refers to techniques that assign a score to input features based on how useful they are at predicting a target variable.

Models with fewer features have:

- Better interpretability
- Reduced redundancy & noise
- Shorter training times
- Simpler production code
- Less storage space because of reduced dataset
- Model accuracy improves

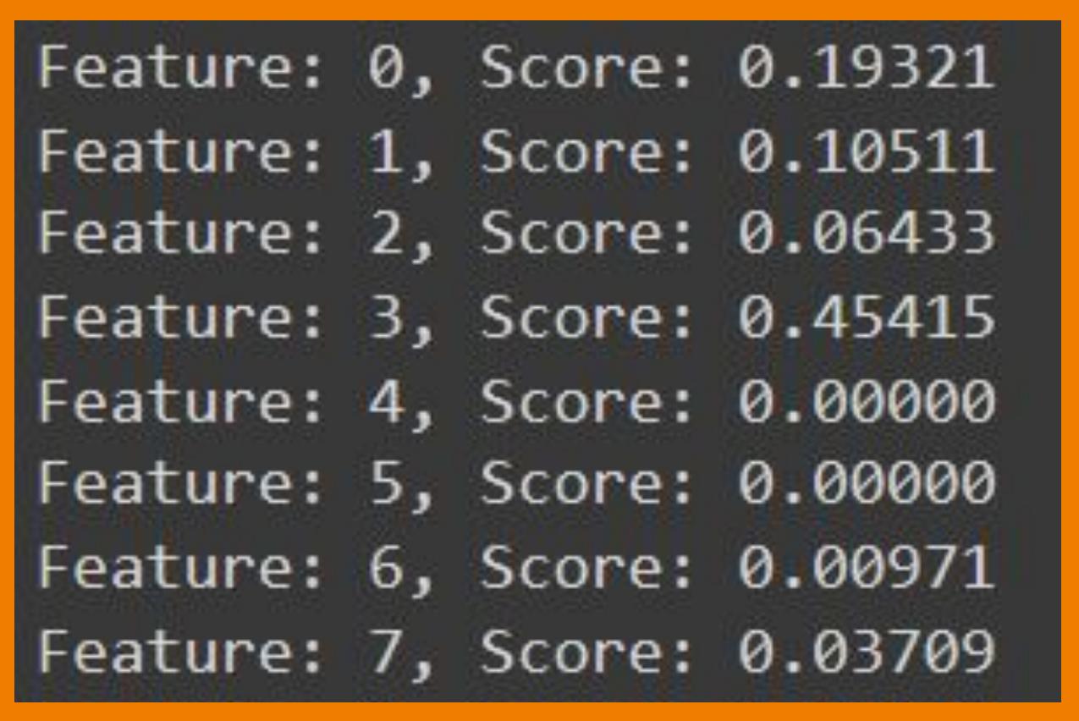
RELEVANCE TO LEGO USE-CASE

In the LEGO Use-Case , out of 12 features for the Decision tree modeling , 5 were dropped ie the 'fork_light_barrier', 'acc_sensor', 'width_Lego', 'horizontal_distance', 'length_Lego' and a reduced dataset was created to generate the training and test datasets for higher accuracy modeling of the decision tree model. There are certain techniques involved to filter out the dataset.

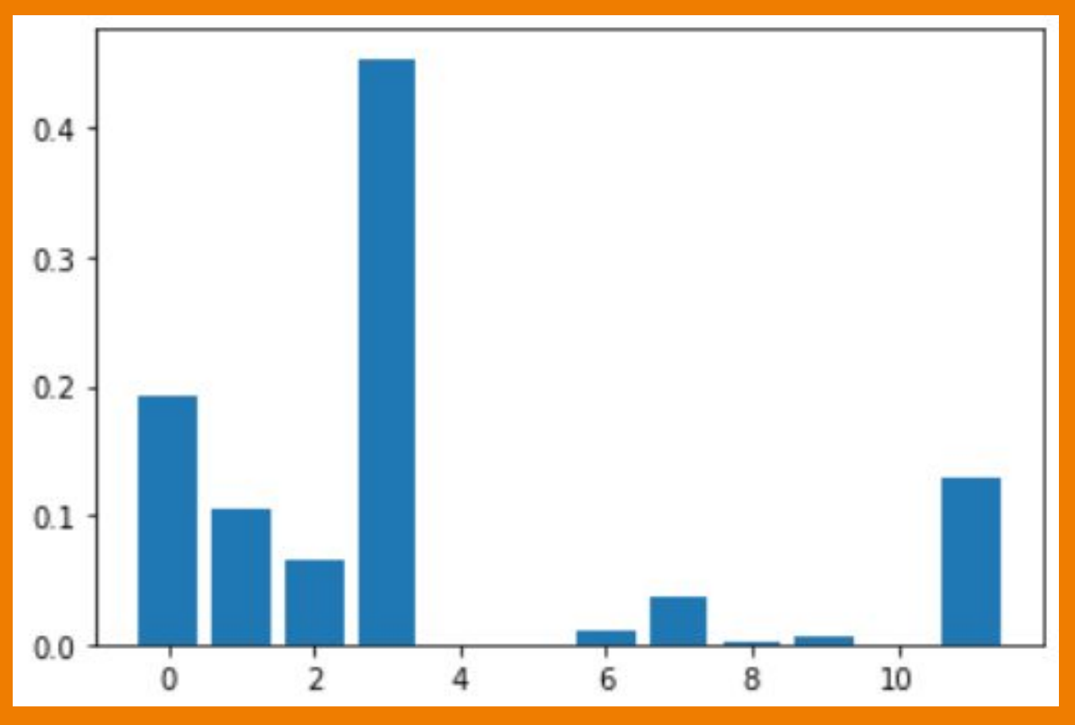
CLASS SOLUTION

Library → SciKit Library

- Assigns a "score" according to the "purity" of feature.
  - This means how much a specific feature helps us get to a correct prediction.



```
Feature: 0, Score: 0.19321
Feature: 1, Score: 0.10511
Feature: 2, Score: 0.06433
Feature: 3, Score: 0.45415
Feature: 4, Score: 0.00000
Feature: 5, Score: 0.00000
Feature: 6, Score: 0.00971
Feature: 7, Score: 0.03709
```



Advantages

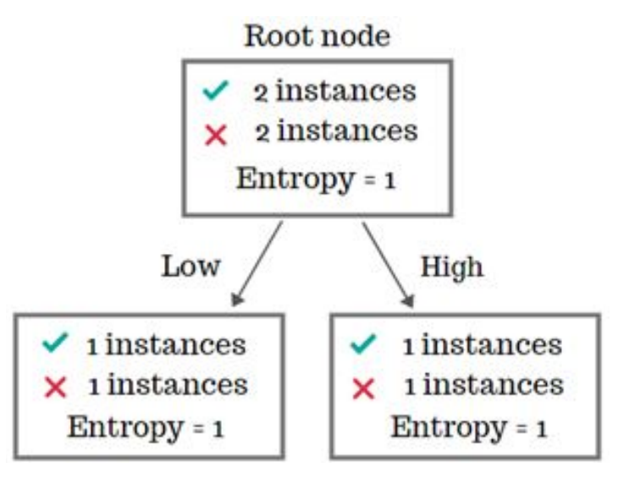
- Easy to apply and use, requiring minimum manipulation.

Disadvantages

- Errors when dataset has high cardinality.
  - *Permutation Feature Importance*

ALTERNATE SOLUTIONS

Information Gain



Entropy used to find the information gain of a variable. If a variable has a high entropy, it means it's probably more important and therefore should not be eliminated.

Confusion Matrix

TP	FP
FN	TN

Displaying the discrepancies between a true and predicted value and the classes that cause the most noise within the results.

Correlation Matrix

1.00	0.64	0.71	0.65	0.77
0.64	1.00	0.58	0.64	0.62
0.71	0.58	1.00	0.66	0.72
0.65	0.64	0.66	1.00	0.60
0.77	0.62	0.72	0.60	1.00

The correlation value usually goes from 0 to 1 or -1.

We read it as <column> is related to <row> and take the corresponding value as the correlation.

DALEX Library

Calculates the level of importance of features using loss functions such as: 'Root Mean Squared Error' and others to see how it fits.

¿What should you choose for your project?
It depends, IT ***ALWAYS*** DEPENDS on the project

REFLECTIONS

58% Reduction of selections

Developers have come up with these feature selection techniques in order to help make data preparation faster, but it is important to note that we cannot rely on these techniques 100%. While these tools do allow you to work faster, they have their downsides, such as not being able to work on large data sets, or the alignment of the information on certain types of training models. We have to understand that these instruments help us, but aren't a hands-free solution.

Tecnológico de Monterrey

FH-W-S University of Applied Sciences Würzburg-Schweinfurt