

WEKA for Data Analysis

Premanand S

Assistant Professor,
School of Electronics and Engineering,
Vellore Institute of Technology, Chennai

premanand.s@vit.ac.in

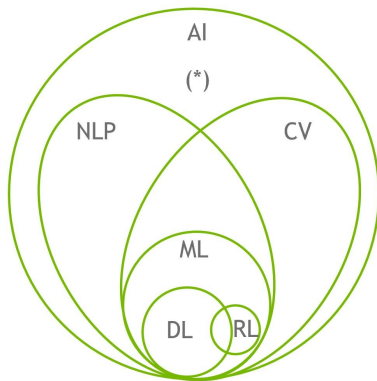
September 18, 2024

- Technology applies scientific knowledge, tools, and techniques to solve problems, improve processes, and create new solutions.
- It includes everything from the wheel to computers to medicines to zippers and buttons on clothes.

Trending Technologies

- Artificial Intelligence
- Artificial Intelligence of Things
- Quantum Computing
- Blockchain
- Cybersecurity
- Augmented Reality and Virtual Reality
- Robotics and many more...

Why do we need to know about these technologies?



AI = Artificial Intelligence
NLP=Natural Language Processing
CV=Computer Vision
ML=Machine Learning
DL=Deep Learning
RL=Reinforcement Learning

(*)=We would have more ellipses there (similar to NLP or CV) representing Robotics, Expert Systems, Speech, and Planning, Scheduling & Optimization systems. But it would look very messy. So, go ahead and imagine they are there too.

Understanding Machine Learning - MEME

Albert Einstein: Insanity Is Doing
the Same Thing Over and Over Again
and Expecting Different Results

Machine learning:



- **General Intro** Machine Learning, means it can access the data and use it to learn for itself without any programming.
Machine Learning is the field of study that gives computers, the ability to learn without being explicitly programmed. — **Arthur Samuel, 1959**
- **Engineering Intro** - A computer program is said to learn from experience E with respect to some task T and some performance measure P , if its performance on T , as measured by P , improves with experience E . — **Tom Mitchell, 1997**
- A baby learns to crawl, walk and then run. We are in the crawling stage when it comes to applying machine learning. - Dave Waters
- Mathematical toolbox used to solve the problems with data

- Data mining is the process of extracting valuable insights, patterns, and knowledge from large datasets.
- It involves applying various techniques and algorithms to discover hidden relationships, trends, and anomalies within the data.
- The main goals of data mining are to gain insights, make predictions, and support decision-making.

- Different programming languages like Python, R, Julia, JavaScript, C/C++, MATLAB, Scala and many more
- WEKA (No code)

- WEKA Download Page

- WEKA - Waikato Environment for Knowledge Analysis
- 1997, Agriculture data analysis
- Objective - Data Mining and Machine Learning tasks
- GUI - Plug and Play
- No code required
- No libraries
- Comprehensive collection of algorithms and tools for various data analysis applications.

Features of WEKA

- Machine Learning Algorithms - Classification, Regression, Clustering, and Association Rule
- Graphical User Interface (GUI) - User-friendly GUI
- Data Preprocessing - Data Cleaning, Normalization, Attribute Selection
- Extensibility - Python and R
- Data Format Support - CSV and ARFF

File format supported

- ARFF (Attribute-Relation File Format)
- ARFF.gz
- CSV (Comma-Separated Values)
- JSON (JavaScript Object Notation)
- JSON.gz
- C4.5 files
- LibSVM
- Matlab ASCII files (.m)
- SVM Light files (.dat)
- Binary Serialized Instances (.bsi)
- XRFF (Extended ARFF) (XRFF.gz)

WEKA - Steps

- Data Preprocessing
- Data Splitting
- Regression
- Classification

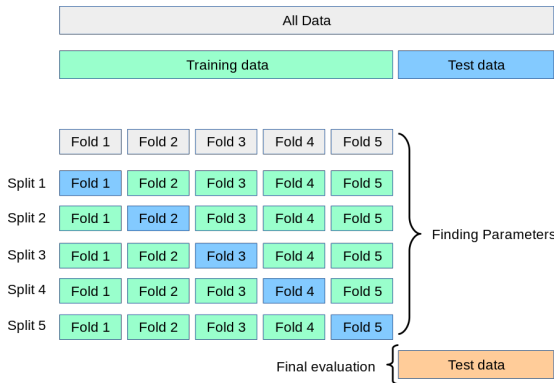
Correlation Coefficient

- Measures the strength and direction of a linear relationship between two variables.
- Pearson's Correlation Coefficient is the most commonly used measure.
- Range: $[-1, 1]$
- Interpretation:
 - 1: Perfect positive linear relationship.
 - -1: Perfect negative linear relationship.
 - 0: No linear relationship.

$$r = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}}$$

Cross Validation

- Instead of a simple train-test split, cross-validation divides the data into several folds (or subsets). The model is trained and tested multiple times, each time using different folds for training and testing.



Mean Absolute Error (MAE)

- Calculates the average absolute difference between the actual and predicted values.
- It gives an idea of how far the predictions are from the actual values, without considering their direction (i.e., no negative or positive errors).
- Range: $[0, \infty]$
- Interpretation: A smaller MAE indicates better model performance.

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

Root Mean Squared Error (RMSE)

- RMSE is the square root of the Mean Squared Error (MSE).
- It is more interpretable because it's in the same unit as the target variable.
- Like MSE, RMSE is sensitive to large errors.
- Range: $[0, \infty]$, where lower values are better.
- Interpretation: A smaller RMSE indicates better model performance and is commonly used to measure the accuracy of predictions.

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

Relative Absolute Error (RAE)

- Measures the relative error of the model's predictions compared to a baseline model (typically the mean of the actual values).
- Provides insight into how much better (or worse) the model performs compared to simply predicting the mean.
- Range: $[0, \infty]$
- Interpretation: RAE less than 1 indicates the model performs better than the baseline; RAE greater than 1 suggests the baseline is better than the model.





$$RAE = \frac{\sum_{i=1}^n |y_i - \hat{y}_i|}{\sum_{i=1}^n |y_i - \bar{y}|}$$

Root Relative Squared Error (RRSE)

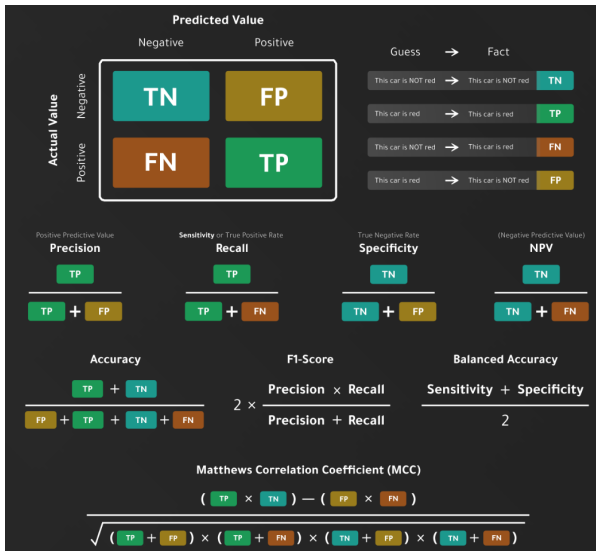
- Measures the relative squared error of the model's predictions compared to a baseline model (typically the mean of the actual values).
- Takes the square root of the average squared differences to give the error in the original units of the target variable.
- Range: $[0, \infty]$
- Interpretation: RRSE less than 1 indicates the model performs better than the baseline; RRSE greater than 1 suggests the baseline is better than the model.

$$RRSE = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

Metrics - Classification - Confusion matrix 1

		Actual Values	
		1	0
Predicted Values	1	TRUE POSITIVE 	FALSE POSITIVE  TYPE 1 ERROR
	0	FALSE NEGATIVE  TYPE 2 ERROR	TRUE NEGATIVE 

Metrics - Classification - Confusion matrix 2



- Confusion Matrix: A table layout for visualizing the performance of a classification model.
- Precision: Indicates how many of the predicted positive cases are actually positive.
- Recall or Sensitivity: Indicates how many of the actual positive cases were correctly predicted.
- Specificity: Indicates how many of the actual negative cases were correctly predicted.
- ROC Curve: Plots the True Positive Rate (Recall) against the False Positive Rate at various threshold settings.
- MCC: Provides a comprehensive measure of classification performance by considering true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN).

mail me: er.anandprem@gmail.com / premanand.s@vit.ac.in
ring me: +91 73586 79961
follow me: LinkedIn
author at Analytics Vidhya: premanand17
author at Medium: Premanand S

Predicting the future isn't magic, it's artificial intelligence!