Experiment No.8
Pre-processing using WEKA tool
Date of Performance:
Date of Submission:

Aim: To implement Pre-processing using WEKA tool

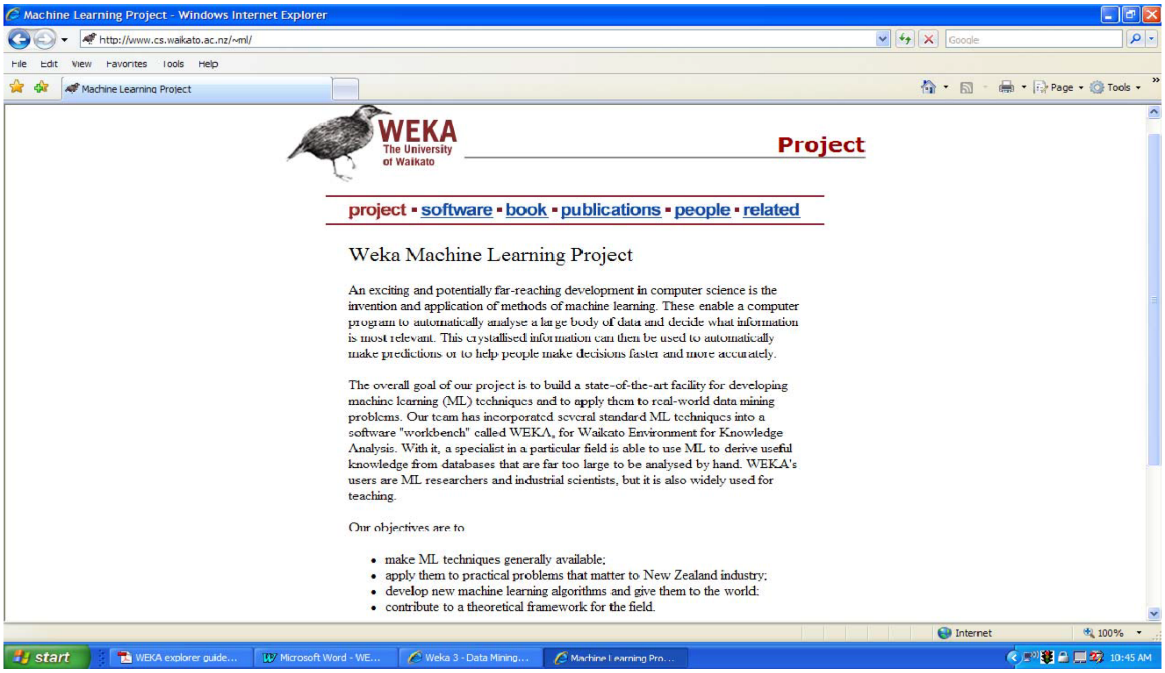
Objective: Pre-processing using WEKA tool

Theory:

WEKA, formally called Waikato Environment for Knowledge Learning, is a computer program that was developed at the University of Waikato in New Zealand for the purpose of identifying information from raw data gathered from agricultural domains. WEKA supports many different standard data mining tasks such as data preprocessing, classification, clustering, regression, visualization and feature selection. The basic premise of the application is to utilize a computer application that can be trained to perform machine learning capabilities and derive useful information in the form of trends and patterns. WEKA is an open source application that is freely available under the GNU general public license agreement. Originally written in C the WEKA application has been completely rewritten in Java and is compatible with almost every computing platform. It is user friendly with a graphical interface that allows for quick set up and operation. WEKA operates on the predication that the user data is available as a flat file or relation, this means that each data object is described by a fixed number of attributes that usually are of a specific type, normal alpha-numeric or numeric values. The WEKA application allows novice users a tool to identify hidden information from database and file systems with simple to use options and visual interfaces.

Installation

The program information can be found by conducting a search on the Web for WEKA Data Mining . The site has a very large amount of useful information on the program's benefits and background. New users might find some benefit from investigating the user manual for the program. The main WEKA site has links to this information as well as past experiments for new users to refine the potential uses that might be of particular interest to them. When prepared to download the software it is best to select the latest application from the selection offered on the site. The format for downloading the application is offered in a self installation package and is a simple procedure that provides the complete program on the end users machine that is ready to use when extracted.


Machine Learning Project - Windows Internet Explorer
http://www.cs.waikato.ac.nz/~ml/
Google
File Edit View Favorites Tools Help
Machine Learning Project
Page
Tools
WEKA
The University of Waikato
Project
project · software · book · publications · people · related
Internet
100%
start
WEKA explorer guide...
Microsoft Word - WE...
Weka 3 - Data Mining...
Machine Learning Pro...
10:45 AM

Weka Machine Learning Project

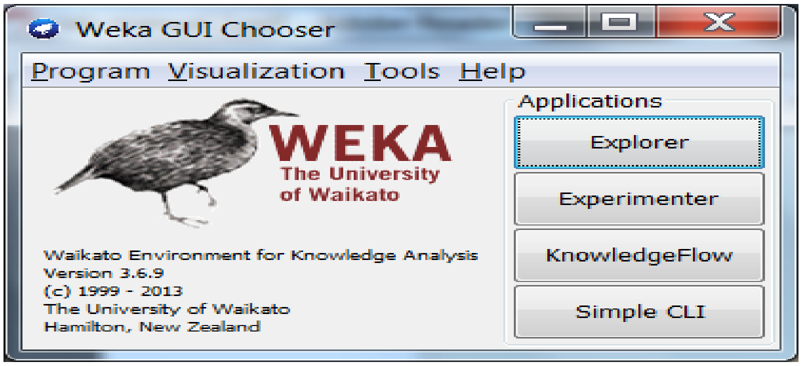
An exciting and potentially far-reaching development in computer science is the invention and application of methods of machine learning. These enable a computer program to automatically analyse a large body of data and decide what information is most relevant. This crystallised information can then be used to automatically make predictions or to help people make decisions faster and more accurately.

The overall goal of our project is to build a state-of-the-art facility for developing machine learning (ML) techniques and to apply them to real-world data mining problems. Our team has incorporated several standard ML techniques into a software "workbench" called WEKA, for Waikato Environment for Knowledge Analysis. With it, a specialist in a particular field is able to use ML to derive useful knowledge from databases that are far too large to be analysed by hand. WEKA's users are ML researchers and industrial scientists, but it is also widely used for teaching.

Our objectives are to

- make ML techniques generally available;
- apply them to practical problems that matter to New Zealand industry;
- develop new machine learning algorithms and give them to the world;
- contribute to a theoretical framework for the field.

The Weka GUI Chooser (class weka.gui.GUIChooser) provides a starting point for launching Weka's main GUI applications and supporting tools. If one prefers a MDI ("multiple document interface") appearance, then this is provided by an alternative launcher called "Main" (class weka.gui.Main).
The GUI Chooser consists of four buttons—one for each of the four major Weka applications—and four menus.



Waikato Environment for Knowledge Analysis
Version 3.6.9
(c) 1999 - 2013
The University of Waikato
Hamilton, New Zealand

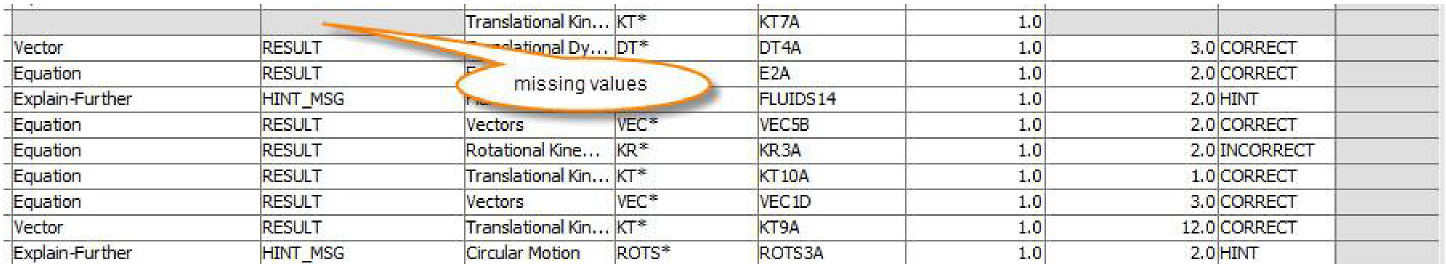
The buttons can be used to start the following applications:

- Explorer An environment for exploring data with WEKA (the rest of this documentation deals with its application in more detail).
- Experimenter An environment for performing experiments and conducting statistical tests between learning schemes.
- KnowledgeFlow This environment supports essentially the same functions as the Explorer but with a drag-and-drop interface. One advantage is that it supports incremental learning.
- SimpleCLI Provides a simple command-line interface that allows direct execution of WEKA commands for operating systems that do not provide their own command line interface.

Step 1: Data Pre Processing or Cleaning

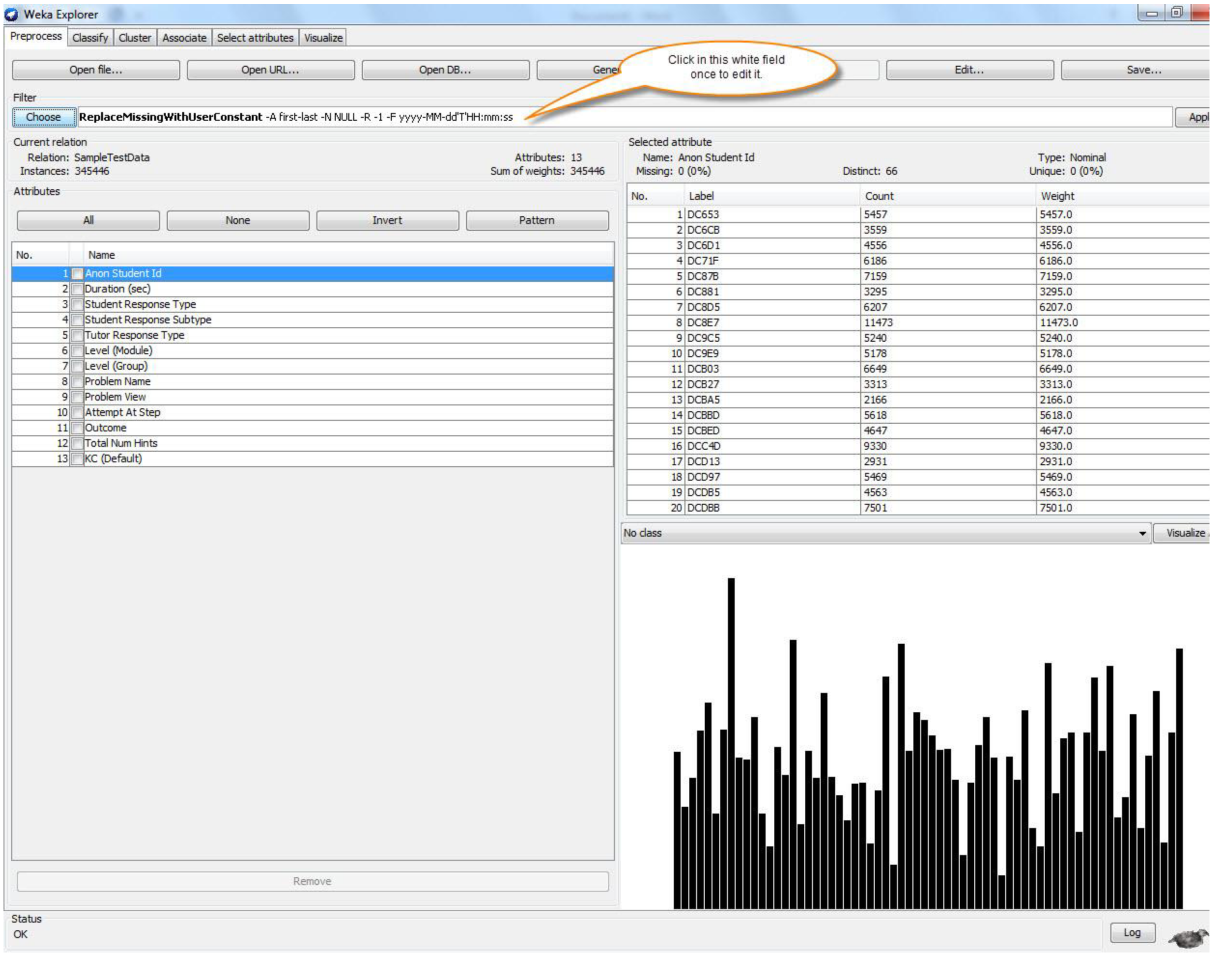
1. Launch Weka-> click on the tab Explorer
2. Load a dataset. (Click on "Open File" & locate the datafile)
3. Click on PreProcess tab & then look at your lower R.H.S. bottom window click on drop down arrow and choose "No Class"
4. Click on "Edit" tab, a new window opens up that will show you the loaded datafile. By looking at your dataset you can also find out if there are missing values in it or not. Also please note the attribute types on the column header. It would either be 'nominal' or 'numeric'.

a. If your data has missing values then its best to clean it first before you apply any forms of mining algorithm to it. In the below figure , you will see the highlighted fields are blank that means the data at hand is dirty and it first needs to be cleaned.



		Translational Kin…	KT*	KT7A	1.0			
Vector	RESULT	[illegible]slational Dy…	DT*	DT4A	1.0	3.0	CORRECT	
Equation	RESULT	[illegible]	[illegible]	E2A	1.0	2.0	CORRECT	
Explain-Further	HINT_MSG	[illegible]	[illegible]	FLUIDS14	1.0	2.0	HINT	
Equation	RESULT	Vectors	VEC*	VEC5B	1.0	2.0	CORRECT	
Equation	RESULT	Rotational Kine…	KR*	KR3A	1.0	2.0	INCORRECT	
Equation	RESULT	Translational Kin…	KT*	KT10A	1.0	1.0	CORRECT	
Equation	RESULT	Vectors	VEC*	VEC1D	1.0	3.0	CORRECT	
Vector	RESULT	Translational Kin…	KT*	KT9A	1.0	12.0	CORRECT	
Explain-Further	HINT_MSG	Circular Motion	ROTS*	ROTS3A	1.0	2.0	HINT	

b. Data Cleaning: To clean the data, you apply "Filters" to it. Generally the data will be missing with values, so the filter to apply is "ReplaceMissingWithUserConstant" (the filter choice may vary according to your need, for more information on it please consult the resources).Click on Choose button below Filters-> Unsupervised->attribute—————> ReplaceMissingWithUserConstant



Click in this white field once to edit it.

ReplaceMissingWithUserConstant -A first-last -N NULL -R -1 -F yyyy-MM-dd'T'HH:mm:ss

Current relation
Relation: SampleTestData
Instances: 345446
Attributes: 13
Sum of weights: 345446

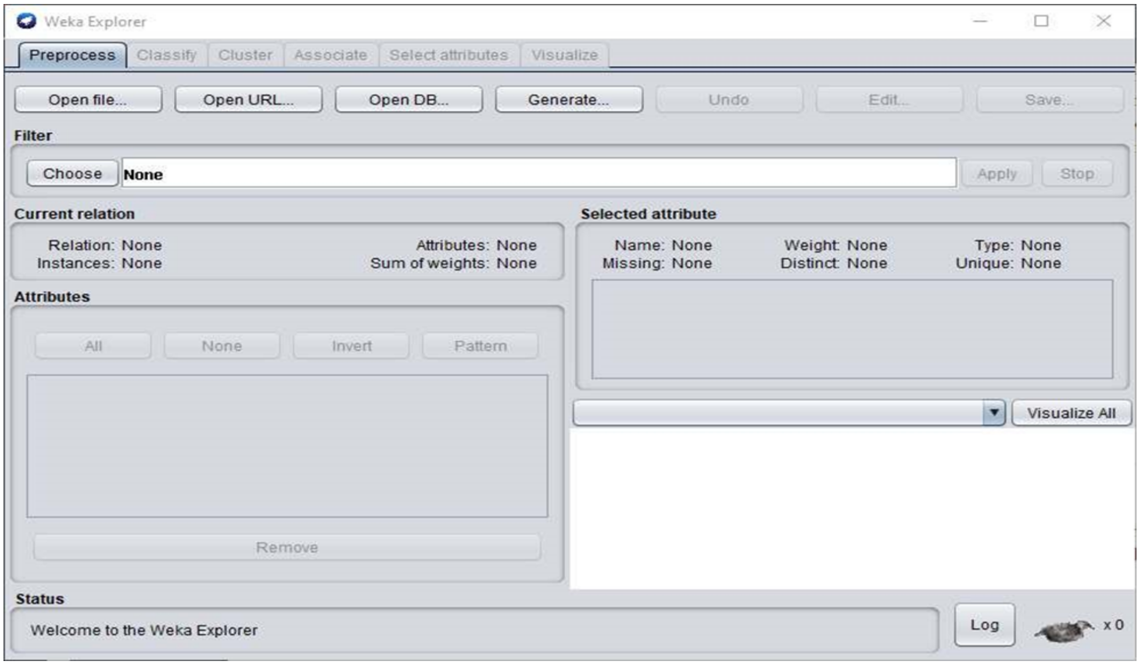
No.	Name
1	Anon Student Id
2	Duration (sec)
3	Student Response Type
4	Student Response Subtype
5	Tutor Response Type
6	Level (Module)
7	Level (Group)
8	Problem Name
9	Problem View
10	Attempt At Step
11	Outcome
12	Total Num Hints
13	KC (Default)

Selected attribute
Name: Anon Student Id
Missing: 0 (0%)
Distinct: 66
Type: Nominal
Unique: 0 (0%)

No.	Label	Count	Weight
1	DC653	5457	5457.0
2	DC6CB	3559	3559.0
3	DC6D1	4556	4556.0
4	DC71F	6186	6186.0
5	DC87B	7159	7159.0
6	DC881	3295	3295.0
7	DC8D5	6207	6207.0
8	DC8E7	11473	11473.0
9	DC9C5	5240	5240.0
10	DC9E9	5178	5178.0
11	DCB03	6649	6649.0
12	DCB27	3313	3313.0
13	DCBA5	2166	2166.0
14	DCBBD	5618	5618.0
15	DCBED	4647	4647.0
16	DCC4D	9330	9330.0
17	DCD13	2931	2931.0
18	DCD97	5469	5469.0
19	DCDB5	4563	4563.0
20	DCDBB	7501	7501.0

Status
OK

Code and output:



Filter

None

Current relation

Relation: None
Instances: None
Attributes: None
Sum of weights: None

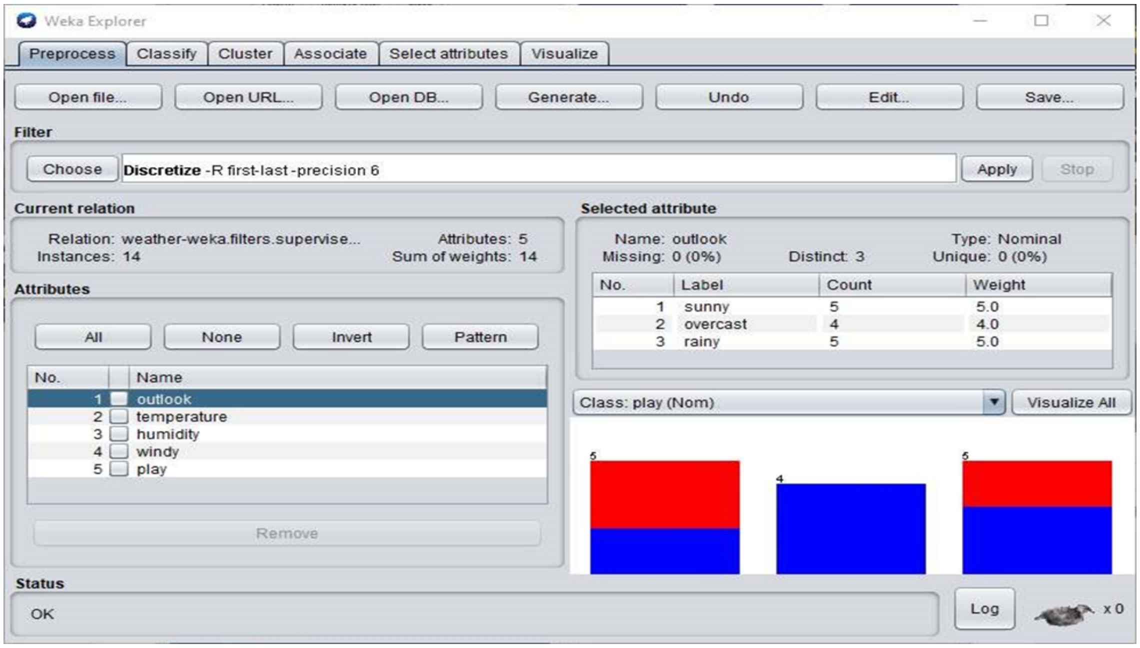
Selected attribute

Name: None
Missing: None
Weight: None
Distinct: None
Type: None
Unique: None

Attributes

Status

Welcome to the Weka Explorer



Filter

Discretize -R first-last -precision 6

Current relation

Relation: weather-weka.filters.supervise...
Instances: 14
Attributes: 5
Sum of weights: 14

Selected attribute

Name: outlook
Missing: 0 (0%)
Distinct: 3
Type: Nominal
Unique: 0 (0%)

No.	Label	Count	Weight
1	sunny	5	5.0
2	overcast	4	4.0
3	rainy	5	5.0

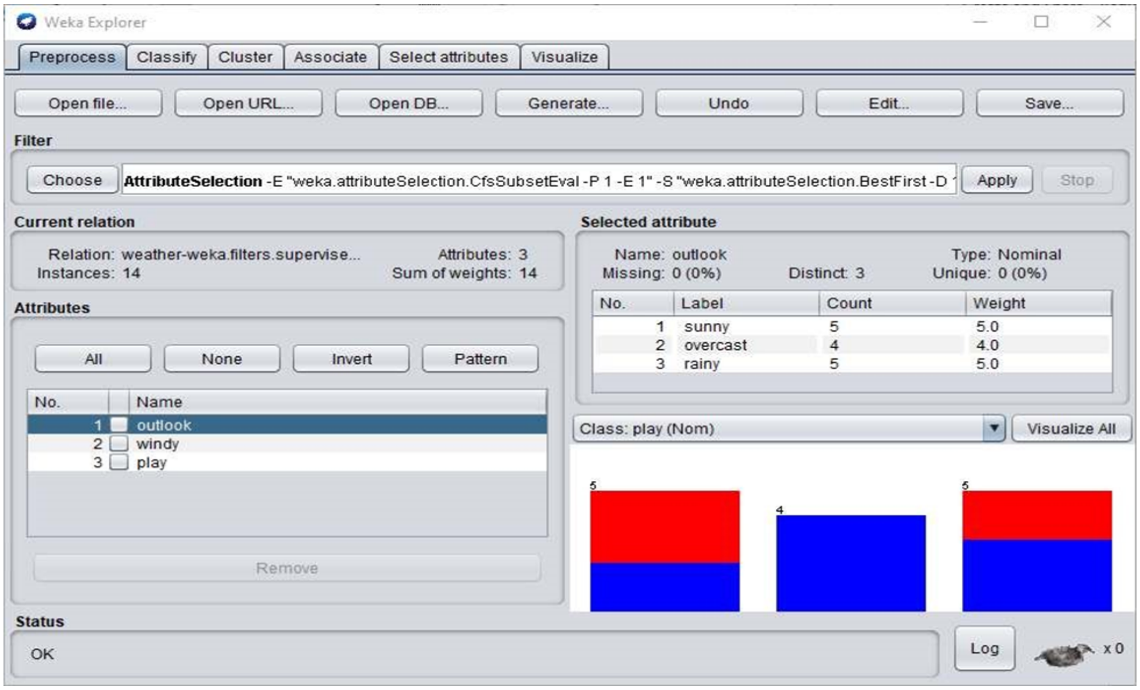
Attributes

No.		Name
1	□	outlook
2	□	temperature
3	□	humidity
4	□	windy
5	□	play

Class: play (Nom)

Status

OK



Filter

AttributeSelection -E "weka.attributeSelection.CfsSubsetEval -P 1 -E 1" -S "weka.attributeSelection.BestFirst -D

Current relation

Relation: weather-weka.filters.supervise... Attributes: 3
Instances: 14 Sum of weights: 14

Selected attribute

Name: outlook Type: Nominal
Missing: 0 (0%) Distinct: 3 Unique: 0 (0%)

No.	Label	Count	Weight
1	sunny	5	5.0
2	overcast	4	4.0
3	rainy	5	5.0

Attributes

No.		Name
1	□	outlook
2	□	windy
3	□	play

Class: play (Nom)

Status

OK

Conclusion: In conclusion, WEKA's usefulness lies in its ability to simplify various aspects of the machine learning workflow, from data preprocessing to model evaluation and deployment. The outputs it generates, including performance metrics, visualizations, and model files, empower users to make data-driven decisions and effectively communicate their results, making it a valuable tool for data scientists, researchers, and analysts.