**Empirical Software Engineering (SE 302)**

**Practical File**

**(2023- 2024)**

**Submitted By**

**Sanoj (2K21/SE/159)**

**Under the Supervision of**

**Ms. Shweta Meena**



**DELHI TECHNOLOGICAL UNIVERSITY**

**(Formerly Delhi College of Engineering)**

**Bawana Road, Delhi-110042**

**INDEX**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **S.No** | **Experiment Name** | **Date** | **Teacher Sign** | **Remarks** |
| 1. | Perform a comparison of the following data analysis tool: WEKA, KEEL, SPSS, MATLAB, R | 6 Feb 2023 |  |  |
| 2. | Collection of Empirical Studies |  |  |  |
| 3. | Write a program to implement B+ Tree, Insertion, Deletion, and Traversal (Character data type) |  |  |  |
| 4. | Write a C++ program to perform the following operations for a Red-Black Tree (RBT) while ensuring that no property of the RedBlack Tree is violated |  |  |  |
| 5. | Write a C++ program to perform following operations for Interval Tree:  Searching, Inserting, and Preorder Traversal. |  |  |  |
| 6. | Write a program to insert element in 2-3 Tree. |  |  |  |
| 7. | Write a program to detect if cycle is present or not using the concept of disjoint set. |  |  |  |
| 8. | Write a program to implement following operation on Binomial Heap:  Make Heap, Insertion, Find Minimum element. |  |  |  |
| 9. | Write a program to implement Fibonacci Heap. |  |  |  |
| 10. | Write a program to count the total number of Spanning Trees for a given graph. |  |  |  |

**EXPERIMENT 1 – Comparison of Data Analysis Tools**

**OBJECTIVE**

Perform a comparison of the following data analysis tool: WEKA, KEEL, SPSS, MATLAB, R

**THEORY**

1. **Weka:**

**Features**: Weka is an open-source machine learning software with a comprehensive collection of algorithms for data mining tasks such as classification, regression, clustering, association rules, and feature selection.

**Ease of Use**: It provides a graphical user interface (GUI) making it accessible to users without extensive programming experience. However, some advanced features may require scripting.

**Flexibility**: Being open-source, Weka allows for customization and integration with other tools or libraries.

**Community Support**: It has an active community with resources like forums, documentation, and tutorials.

1. **KEEL:**

**Features**: KEEL is a Java-based software tool for a wide range of data mining tasks. It offers algorithms for classification, regression, clustering, pattern mining, etc.

**Ease of Use**: It provides a user-friendly interface but may require some learning curve, especially for users unfamiliar with Java.

F**lexibility**: KEEL allows for customization and supports the integration of new algorithms.

**Community Support**: While it has a user community, it might not be as extensive as other more widely used tools.

1. **SPSS:**

**Features**: SPSS (Statistical Package for the Social Sciences) is a statistical software suite offering a broad range of data analysis capabilities including descriptive statistics, hypothesis testing, regression analysis, and more.

**Ease of Use**: It provides a user-friendly interface with point-and-click functionalities, making it suitable for non-programmers.

**Flexibility**: SPSS offers some customization options, but it might be limited compared to open-source alternatives.

**Community** **Support**: It has a large user base, with extensive documentation and support available.

1. **MATLAB:**

**Features**: MATLAB is a programming language and environment primarily focused on numerical computing. It offers various toolboxes for data analysis, including statistics, machine learning, signal processing, etc.

**Ease of Use**: MATLAB provides an interactive development environment (IDE) with easy-to-use functions and visualization tools. However, proficiency in MATLAB programming is required for complex tasks.

**Flexibility**: MATLAB offers high flexibility and customization options, allowing users to create custom algorithms and functions.

**Community** **Support**: MATLAB has a large user base and comprehensive documentation, with active forums and support channels.

1. **R:**

**Features**: R is a programming language and environment specifically designed for statistical computing and graphics. It offers a vast collection of packages for data analysis, visualization, and machine learning.

**Ease** **of** **Use**: While R has a steep learning curve for beginners, it provides powerful functionalities once mastered. Various IDEs and graphical interfaces like RStudio make it more user-friendly.

**Flexibility**: R is highly flexible, allowing users to write custom functions and packages. Its open-source nature encourages community contributions and extensions.

**Community** **Support**: R has a large and active user community with extensive documentation, numerous packages, and online resources.

**LEARNING FROM EXPERIMENT**

1. **Consider your specific needs**: Each tool has its strengths and weaknesses. Consider what tasks you need to accomplish and choose the tool that best aligns with your requirements. For example, if you primarily need statistical analysis, SPSS might be a good choice. If you're focusing on machine learning, Weka, MATLAB, or R might be more suitable.
2. **Evaluate ease of use**: Depending on your familiarity with programming and your team's skill set, consider the ease of use of each tool. If you're a beginner or prefer a point-and-click interface, SPSS or Weka might be better options. If you're comfortable with programming, MATLAB or R might provide more flexibility.
3. **Flexibility and customization**: If you anticipate needing to customize algorithms or integrate with other systems, consider the flexibility of each tool. Open-source tools like R and Weka offer high levels of customization, while commercial tools like SPSS might have limitations in this regard.
4. **Community support**: Look into the availability of community support, documentation, tutorials, and forums for each tool. A strong user community can provide valuable assistance and resources as you learn and use the tool.
5. **Cost considerations**: While some tools like R and Weka are open-source and free to use, others like SPSS and MATLAB may require purchasing licenses. Consider your budget and the cost-effectiveness of each option.

**EXPERIMENT 2 – Collection of Empirical Studies**

**OBJECTIVE**

Collection of Empirical Studies

THEORY

LEARNING FROM EXPERIMENT

**EXPERIMENT 3 – Collection of Empirical Studies**

**OBJECTIVE**

Collection of Empirical Studies

**EXPERIMENT 4 – Collection of Empirical Studies**

**OBJECTIVE**

Collection of Empirical Studies

**EXPERIMENT 5 – Feature Reduction Techniques**

**OBJECTIVE**

Write a program to perform following feature reduction technique for the collected dataset

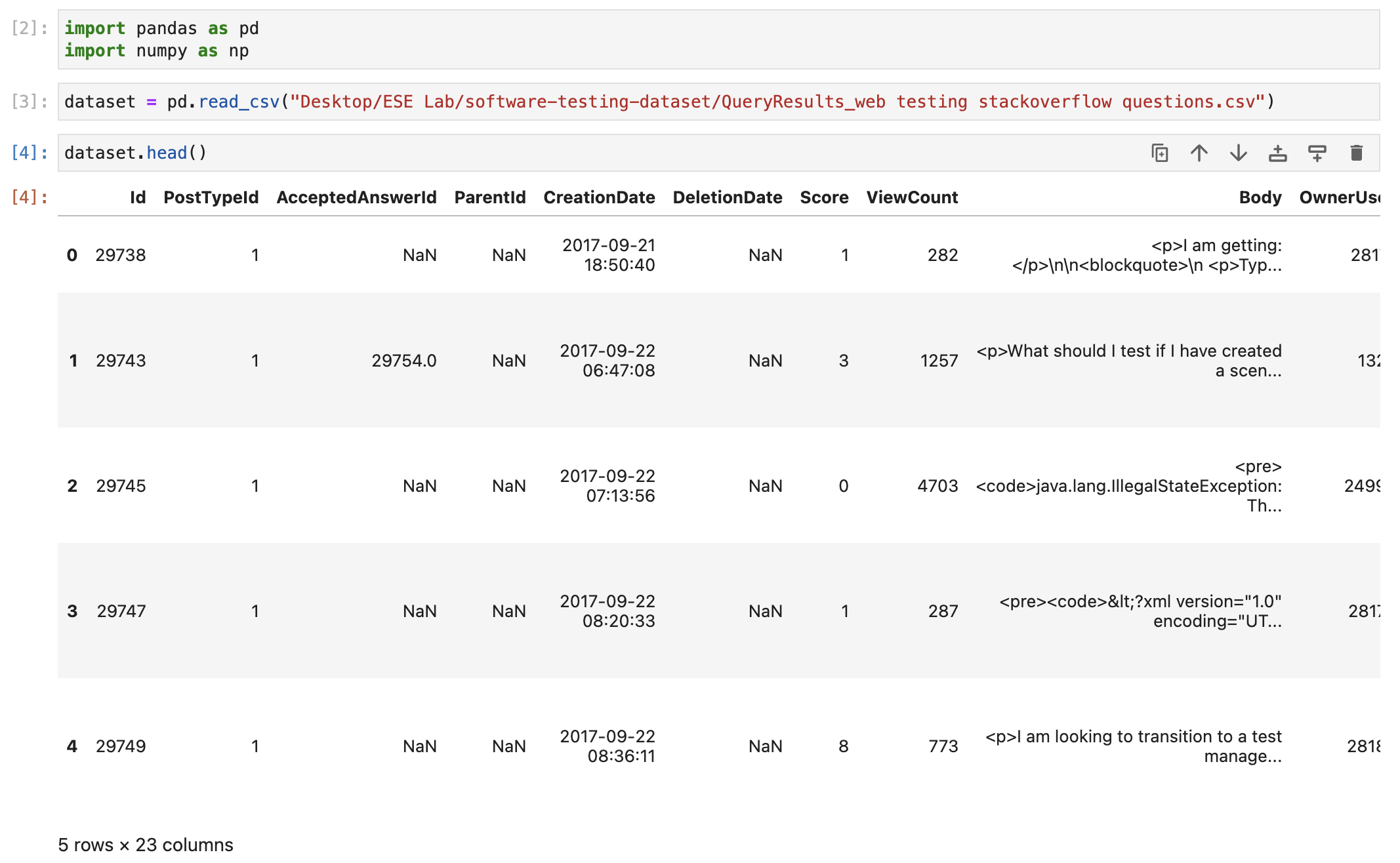
1. Correlation-based feature evaluation
2. Relief attribute feature evaluation
3. Information gain feature evaluation
4. Principle component analysis

**THEORY**

1. **Correlation-based feature evaluation**: This approach evaluates the relationship between each feature and the target variable by calculating their correlation coefficient. Features with high correlation values with the target variable are considered important and are retained, while those with low correlation values may be discarded. However, it's essential to note that correlation doesn't imply causation, so this method might overlook certain important features that are not highly correlated but still influential.
2. **Relief attribute feature evaluation**: The Relief algorithm estimates the importance of features by considering their ability to distinguish between instances of the same and different classes. It works by iteratively sampling instances and adjusting feature weights based on the differences in feature values between the nearest instances of the same and different classes. Features with higher weights are considered more relevant. This method is particularly useful for classification tasks and is robust to noisy data.
3. **Information gain feature evaluation**: Information gain measures the reduction in entropy or uncertainty about the target variable achieved by knowing the value of a particular feature. Features that lead to significant reductions in entropy are considered more informative and are thus selected. This method is commonly used in decision tree algorithms, where features with higher information gain are preferred for splitting nodes. However, it may prioritize features with many distinct values or categories.
4. **Principal component analysis (PCA)**: PCA is a dimensionality reduction technique that identifies the directions (principal components) that capture the most variance in the data. These principal components are linear combinations of the original features. By retaining only the most significant principal components, PCA reduces the dimensionality of the data while preserving most of its variance. This technique is particularly useful for visualizing high-dimensional data and for feature extraction in scenarios where the original features are highly correlated or redundant.

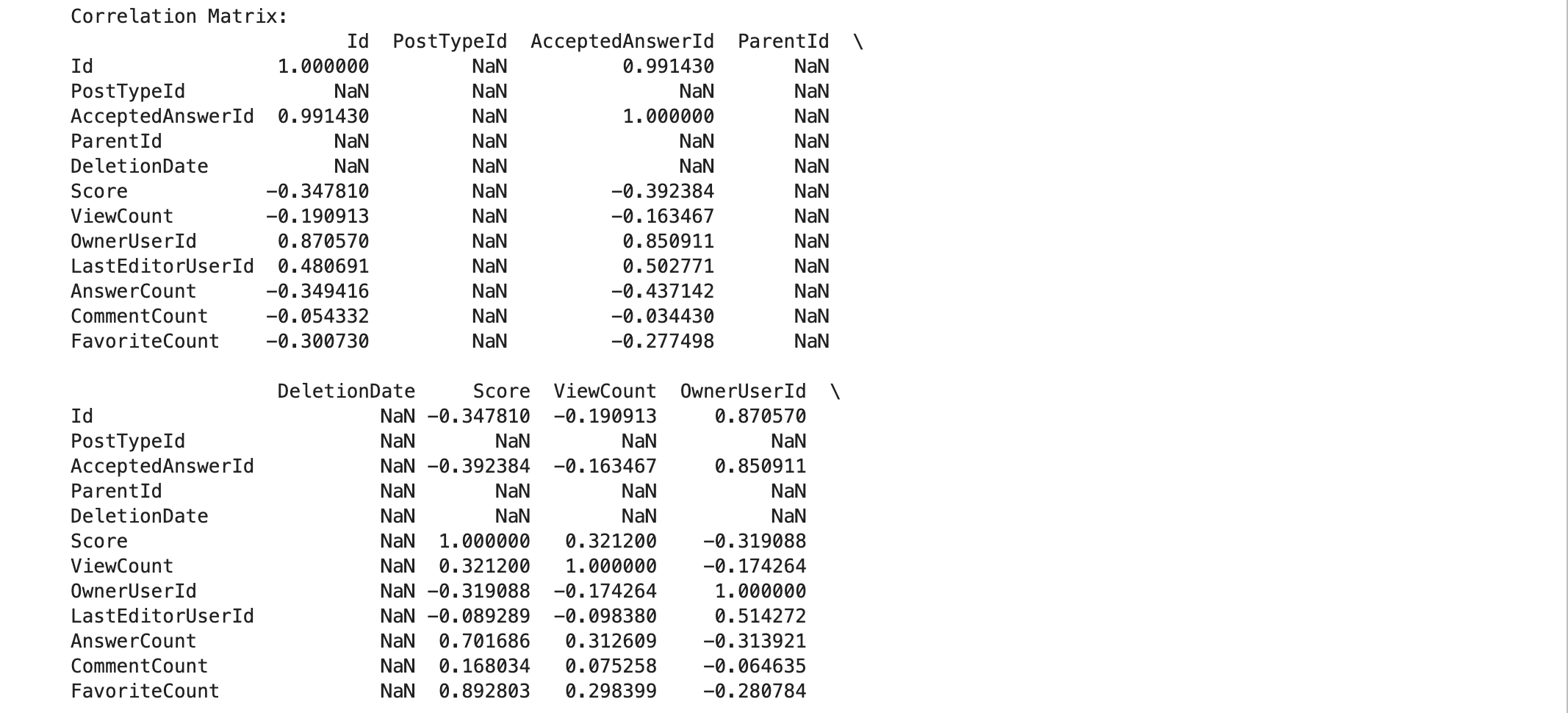
**CODE AND OUTPUT**

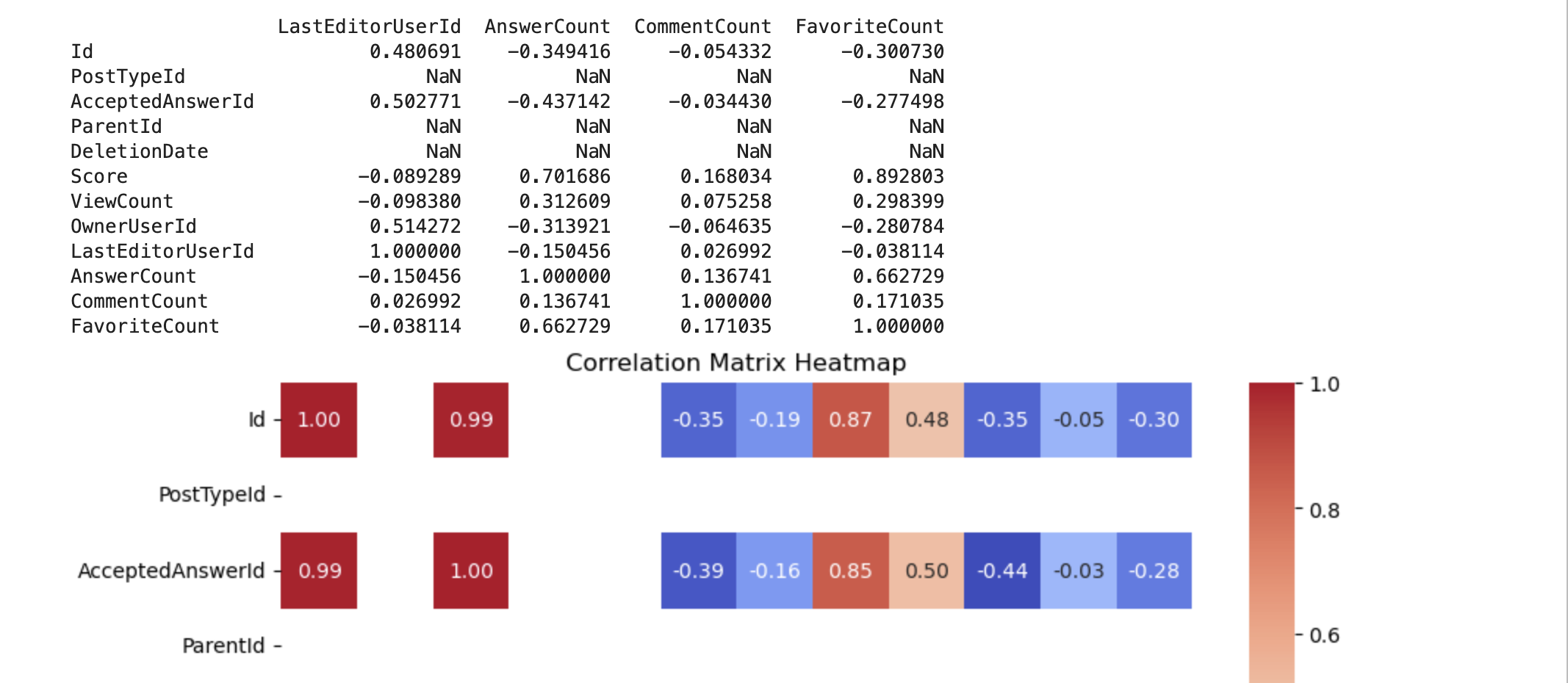
Importing Libraries and Dataset

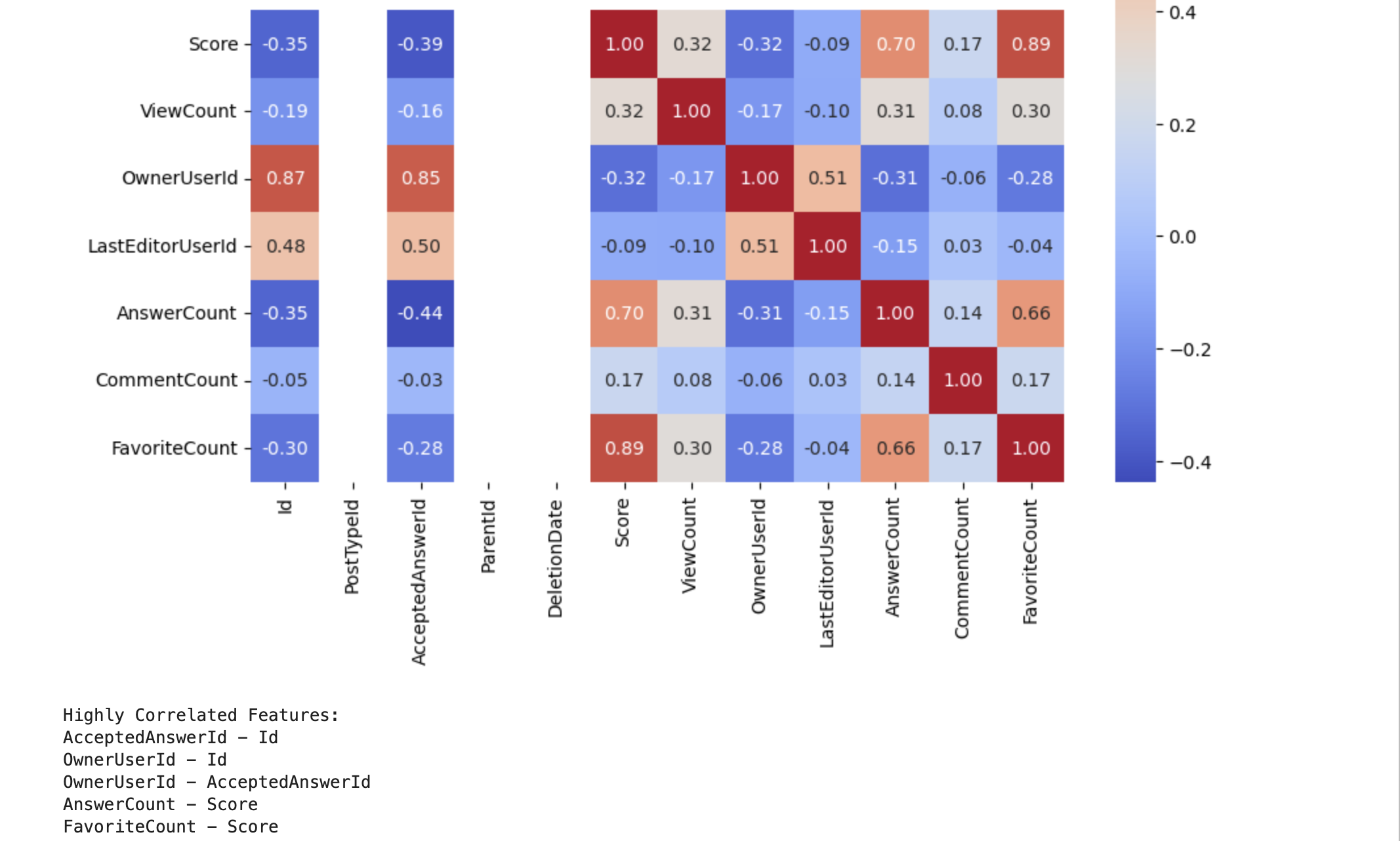


Correlation Based Feature Evaluation









Relief Attribute Feature Evaluation

LEARNING