**README: Feature Selection using VSI**

**Overview**

**This script performs feature selection based on the Vendi Score, utilizing cosine similarity and Bayesian updating to rank features. It evaluates feature importance in a dataset using a cross-validation approach and selects features iteratively based on their posterior probabilities.**

**Key Features**

* **Vendi Score Evaluation: Uses the Vendi library to calculate scores based on cosine similarity matrices.**
* **Bayesian Updating: Implements Bayes' theorem to update feature selection probabilities iteratively.**
* **Cross-Validation: Employs k-fold cross-validation to ensure robust feature evaluation.**
* **Visualization: Outputs a bar plot of feature importance ranked by their Vendi scores.**
* **Scalability: Designed to handle datasets with multiple features efficiently.**

**Requirements**

**Libraries**

* **numpy**
* **pandas**
* **scikit-learn**
* **matplotlib**
* **vendi\_score**

**Data Files**

**The following datasets are studied in the corresponding paper cited at the end of this readme file:**

1. **Boston Dataset (Regression):**
   * **Input: Boston.xlsx**
   * **Sheet Name: FL**
   * **Features: Columns A:M**
   * **Target: Column N**
2. **Concrete Dataset (Regression):**
   * **Input: Concrete.xlsx**
   * **Sheet Name: FL**
   * **Features: Columns A:H**
   * **Target: Column I**
3. **Red Wine Dataset (Classification):**
   * **Input: Wine\_Red.xlsx**
   * **Sheet Name: FL**
   * **Features: Columns A:K**
   * **Target: Column L**
4. **White Wine Dataset (Classification):**
   * **Input: Wine\_White.xlsx**
   * **Sheet Name: FL**
   * **Features: Columns A:K**
   * **Target: Column L**

**Installation**

1. **Clone this repository.**
2. **Install the required Python libraries:**
3. **pip install numpy pandas scikit-learn matplotlib vendi\_score**
4. **Place the input Excel files (Boston.xlsx, Concrete.xlsx, Wine\_Red.xlsx, Wine\_White.xlsx) in the same directory as the script.**

**How to Use**

1. **Prepare Data: Ensure the dataset matches the expected format for features and target columns.**
2. **Run the Script:**
3. **python feature\_selection\_vendi.py**
4. **Select Dataset: Modify the script to load the desired dataset.**

**Example for loading the Boston dataset:**

**# Load Boston data**

**X = pd.read\_excel(' Boston.xlsx', sheet\_name='FL', usecols='A:M', skiprows=0)**

**y = pd.read\_excel(' Boston.xlsx', sheet\_name='FL', usecols='N', skiprows=0)**

1. **Output:**
   * **Selected features are printed in the console.**
   * **A plot of feature importance is saved as Figures/Boston\_VS.pdf (or another file depending on the dataset used).**

**Output Description**

**Console Logs**

* **Selected Features: Displays features in order of importance.**
* **Metric Value: Indicates the metric used for feature ranking (based on Vendi score and uncertainty).**

**Visualization**

* **Feature Importance Plot: A bar plot showing features ranked by importance based on the final Vendi scores.**

**Code Walkthrough**

**Data Preprocessing**

* **Reads feature (X) and target (y) data from the selected Excel file.**
* **Standardizes features and target values using StandardScaler.**

**Feature Selection**

* **Iteratively evaluates features using cosine similarity and cross-validation.**
* **Updates probabilities using Bayesian inference to refine feature rankings.**

**Visualization**

* **Generates a bar plot with features ranked by importance based on the final Vendi scores.**

**Example Usage**

**After running the script, the following results are produced:**

1. **Selected features are printed in the terminal.**
2. **Feature importance plot is saved in the Figures directory.**

**Customization**

* **Adjust Cross-Validation Folds: Modify n\_folds to change the number of cross-validation splits.**
* **Change Data File: Replace the data loading section with the desired dataset.**
* **Add Features: Update usecols and sheet names for different datasets.**

**Notes**

* **Ensure the input files are structured correctly; otherwise, preprocessing will fail.**
* **The script automatically creates the Figures directory if it does not exist.**

**Citation**

**If you use this script in your work, please cite the following paper:**

**Mousavi, M., & Khalili, N. (2025). VSI: An interpretable Bayesian feature selection method based on Vendi score. *Knowledge-Based Systems*, 112973.** [**https://doi.org/10.1016/j.knosys.2025.112973**](https://doi.org/10.1016/j.knosys.2025.112973)

**Contact**

**For further questions or feedback, please reach out via the project's repository or email. For further information, contact Dr Mohsen Mousavi at** [**mohsen.mousavi@unsw.edu.au**](mailto:mohsen.mousavi@unsw.edu.au)**.**