**Feature Selection Model**

When analyzing a child's handwriting data, a total of 32 features are extracted. However, not all of these features are equally relevant for identifying dysgraphia. Including all 32 features in the model may lead to overfitting, where the model learns noise in the data and performs poorly on new, unseen data. To address this issue, it is crucial to select a subset of the most informative features that are best suited for screening dysgraphia. One way to accomplish this is by utilizing techniques like mutual information, which measures the relationship between each feature and the target variable (presence or absence of dysgraphia). Features with high mutual information are more likely to be relevant for dysgraphia detection. By selecting a smaller set of these high-information features, a more focused and effective model that is less prone to overfitting can be created, ultimately enhancing the accuracy of dysgraphia screening. So, this feature selection process is a critical step in improving the overall performance and interpretability of your dysgraphia detection model.

**Information Gain**

It is the amount of information provided by the feature for identifying the target value and measures reduction in the entropy values. Information gain of each attribute is calculated considering the target values for feature selection. The lower the entropy, the more ordered and predictable is the data. Information Gain quantifies how much splitting the data based on a feature reduces this uncertainty.

Formula: The Information Gain (IG) for a feature can be calculated as follows:

IG(Feature) = Entropy before splitting - Weighted average of Entropy after splitting

where Entropy is calculated based on the distribution of the target variable in the subsets created by the split.

By using the above formula, features with higher Information Gain are considered more important and are often selected for use in classification tasks. Information Gain provides a ranking of features based on their relevance to the target variable.

This ranking can be useful for selecting a subset of the most informative features where this method is considered as appropriate and effective in the case of handwriting analysis.



Preprocessing of time-series handwriting data into discrete data

Calculate mutual information of

each feature with target variable

Original Feature Vector of 82 Data samples

32 iterations

Final Optimal Feature Set

Feature Ranking using Information Gain

Fig: Block Diagram of Feature Selection Process

**Algorithm**

1. Import the required libraries and load your dataset into a Pandas DataFrame.

2. Specify your target variable 'y' as 'label’ (whether dysgraphia is present or not) and your features 'X' as all columns except 'label' and 'Id’.

3. The mutual\_info\_classif function from scikit-learn to calculate the mutual information between each feature and the target variable.

4. The mutual\_info\_classif function estimates mutual information by comparing the distributions of feature values against the distribution of target variable values. It calculates the mutual information for each feature individually, treating them as random variables.

5. By specifying discrete\_features='auto', the function will automatically determine if a feature is discrete or continuous based on the data type of the feature and the number of unique values. This is essential because the calculation differs for discrete and continuous features and sort this DataFrame by the 'Mutual\_Information' column in descending order.

**Significant Features**

* StartTime
* RelativeDurationofPrimary
* AverageNormalizedJerkPerTrial
* AverageNormalizedyJerkPerTrial
* AbsoluteSize
* Roadlength
* RelativeInitialSlant
* StartVerticalPosition
* AveragePenPressure
* NumberOfPeakAccelerationPoints
* RelativePenDownDuration

**Results**

* Raw data is processed and features are extracted using MovAlyzer software.
* Data is pre-processed so that it is suitable to build the machine learning model.
* Features are reduced by analysing and understanding the significance of each of them in detecting the dysgraphia using information gain technique.

