**EC 9560 – DATA MINING**

**LAB 01**

**DARMILA.T**

**2020/E/027**

**SEMESTER 07**

**02th OCTOBER 2024**

**Title:** Predict the level of Problematic Internet Use Among Children and Adolescents Using based on demographic, physical health, internet use, and fitness data.

**Objective**: To predict the level of problematic internet usage exhibited by children and adolescents based on their physical activity and fitness data. By developing a predictive model, the aim is to identify early indicators of problematic internet use, allowing for timely interventions that promote healthier digital habits.

**Methodology:**

1. Data Collection

* Load and Explore the Dataset: Start by loading the train.csv file, which contains features such as demographic information, internet usage behavior, fitness measures, and other health indicators. Analyze the data\_dictionary.csv to fully understand the meaning and context of each field.

1. Data Preprocessing

* Handle Missing Data: Address missing values using techniques like mean or median imputation, or consider removing features that have an excessive amount of missing data.
* Feature Engineering: Create new features if they can provide additional insight or improve model performance.
* Data Cleaning: Identify and resolve invalid or inconsistent data points. Convert categorical variables into numerical representations using methods like one-hot encoding or label encoding.

1. Data Splitting

* Train-Validation Split: Divide the preprocessed training data into training and validation sets to enable model tuning and evaluation.
* Stratified Sampling: Since the target variable (Severity Impairment Index, sii) is ordinal with categories (None, Mild, Moderate, Severe), use stratified sampling to preserve the class distribution during the split.

1. Model Selection

* Algorithms for Classification:
  + Logistic Regression
  + Decision Trees
  + k-Nearest Neighbors (KNN)
  + Naive Bayes
  + Support Vector Machines (SVM)
  + Random Forest Classifier
  + Gradient Boosting Classifier

1. Model Training and Hyperparameter Tuning

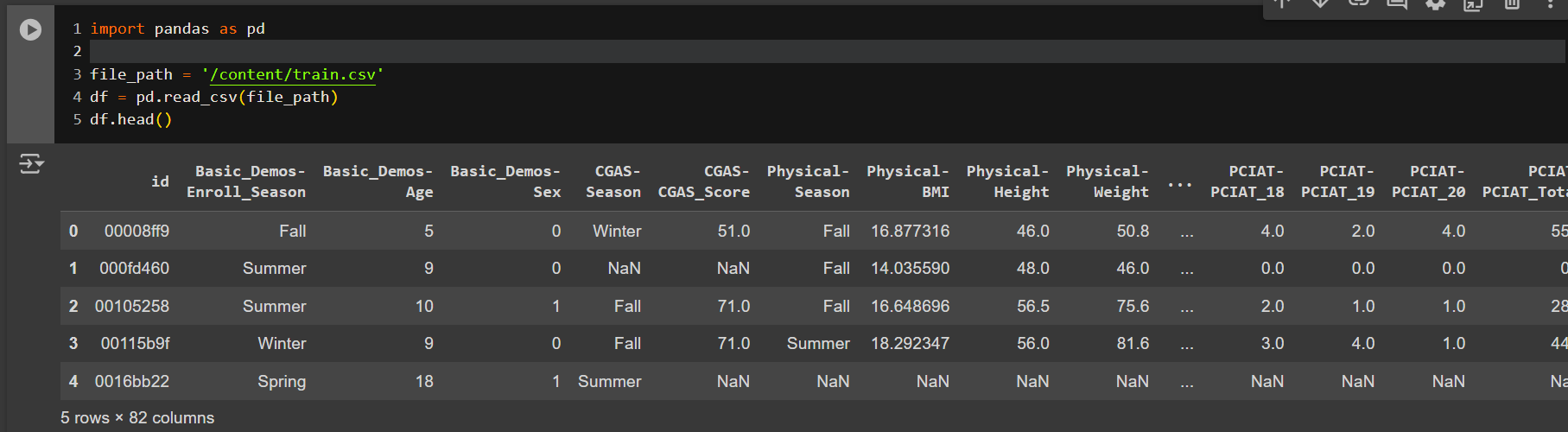
* Model Training: Train several models on the training set and use cross-validation to prevent overfitting.
* Hyperparameter Optimization: Utilize hyperparameter tuning techniques like Grid Search or Random Search to find the optimal configuration for each model.

1. Evaluation
   * Metrics for Evaluation:
     + Accuracy: Measure how often the model predicts the correct category.
     + Precision, Recall, and F1-Score: These metrics are important for assessing performance, especially with imbalanced classes.
     + Confusion Matrix: Use this to visualize misclassifications among the different severity levels.
     + AUC-ROC Curve: Evaluate the model's ability to distinguish between the different severity levels of problematic internet use.
2. Final Model and Predictions

* Once the best-performing model has been finalized, apply it to the test.csv dataset (test set) to predict the Severity Impairment Index (sii).

**Data description with a link to data in data repository:**

<https://www.kaggle.com/competitions/child-mind-institute-problematic-internet-use/data?select=train.csv>



A screenshot of a computer

Description automatically generated

A screenshot of a computer program

Description automatically generated

Reference :

https://www.kaggle.com/competitions/child-mind-institute-problematic-internet-use/overview