# **Child Mind Institute - Problematic Internet Use**

**Challenge description:**  
In today’s digital age, problematic internet use among children and adolescents is a growing concern. Better understanding this issue is crucial for addressing mental health problems such as depression and anxiety.

Current methods for measuring problematic internet use in children and adolescents are often complex and require professional assessments. This creates access, cultural, and linguistic barriers for many families. Due to these limitations, problematic internet use is often not measured directly, but is instead associated with issues such as depression and anxiety in youth.

Conversely, physical & fitness measures are extremely accessible and widely available with minimal intervention or clinical expertise. Changes in physical habits, such as poorer posture, irregular diet, and reduced physical activity, are common in excessive technology users. We propose using these easily obtainable physical fitness indicators as proxies for identifying problematic internet use, especially in contexts lacking clinical expertise or suitable assessment tools.

This competition challenges you to develop a predictive model capable of analyzing children's physical activity data to detect early indicators of problematic internet and technology use. This will enable prompt interventions aimed at promoting healthier digital habits.

Your work will contribute to a healthier, happier future where children are better equipped to navigate the digital landscape responsibly.

**Data description:**

The Healthy Brain Network (HBN) dataset is a clinical sample of about five-thousand 5-22 year-olds who have undergone both clinical and research screenings. The objective of the HBN study is to find biological markers that will improve the diagnosis and treatment of mental health and learning disorders from an objective biological perspective. Two elements of this study are being used for this competition: physical activity data (wrist-worn accelerometer data, fitness assessments and questionnaires) and internet usage behavior data. The goal of this competition is to predict from this data a participant's **Severity Impairment Index** (sii), a standard measure of problematic internet use.

* Demographics - Information about age and sex of participants.
* Internet Use - Number of hours of using computer/internet per day.
* Children's Global Assessment Scale - Numeric scale used by mental health clinicians to rate the general functioning of youths under the age of 18.
* Physical Measures - Collection of blood pressure, heart rate, height, weight and waist, and hip measurements.
* FitnessGram Vitals and Treadmill - Measurements of cardiovascular fitness assessed using the NHANES treadmill protocol.
* FitnessGram Child - Health related physical fitness assessment measuring five different parameters including aerobic capacity, muscular strength, muscular endurance, flexibility, and body composition.
* Bio-electric Impedance Analysis - Measure of key body composition elements, including BMI, fat, muscle, and water content.
* Physical Activity Questionnaire - Information about children's participation in vigorous activities over the last 7 days.
* Sleep Disturbance Scale - Scale to categorize sleep disorders in children.
* Actigraphy - Objective measure of ecological physical activity through a research-grade biotracker.
* Parent-Child Internet Addiction Test - 20-item scale that measures characteristics and behaviors associated with compulsive use of the Internet including compulsivity, escapism, and dependency.

Note in particular the field PCIAT-PCIAT\_Total. The target sii for this competition is derived from this field as described in the data dictionary: 0 for None, 1 for Mild, 2 for Moderate, and 3 for Severe. Additionally, each participant has been assigned a unique identifier id.

## **Challenges:**

The journey to develop an effective predictive model for the SII using the HBN dataset was fraught with challenges that required careful consideration and strategic planning.

1. **Data Quality and Noise**: The dataset is characterized by noise and variability, which can obscure the underlying signals we aim to measure. Many features within the dataset are subjective and influenced by a range of factors, including the time of day, the emotional state of the participants, and even their last meal. This variability introduces a level of randomness that complicates the modeling process.
2. **Bias and Incompleteness**: The data is also subject to systemic biases that can skew results. For instance, parental reports of problematic internet use may be influenced by the parents' own internet habits, leading to biased perceptions of their children's behavior. Additionally, the dataset contains incomplete records, which can further complicate analysis and interpretation. The presence of these biases necessitates careful consideration during the modeling process to avoid drawing misleading conclusions.
3. **Heterogeneity**: The diversity within the dataset is both a strength and a challenge. The sample includes participants with varying degrees of internet use and mental health status, leading to a wide range of responses. Many participants are not severely impacted by internet use, which reflects the prevalence of the issue within the community. This heterogeneity can lead to imbalances in the data, making it easy for models to overfit to the training data rather than generalizing well to unseen cases.
4. **Complex Relationships**: The relationships between variables in the dataset are often intricate and multifaceted. For example, factors such as accessibility to technology, socio-economic status, and individual psychological traits can all influence internet use and its associated impacts. Understanding these complex relationships is crucial for developing an effective model, yet it adds another layer of difficulty to the analysis.
5. **Real-World Context:** The challenges presented by the HBN dataset mirror those encountered in real-world psychiatric research. By working with this raw, unfiltered data, we aim to uncover hidden relationships that can inform our understanding of problematic internet use among children and adolescents. This approach allows us to apply what we learn in practice, rather than relying on a sanitized version of the data that may not reflect the complexities of real life.

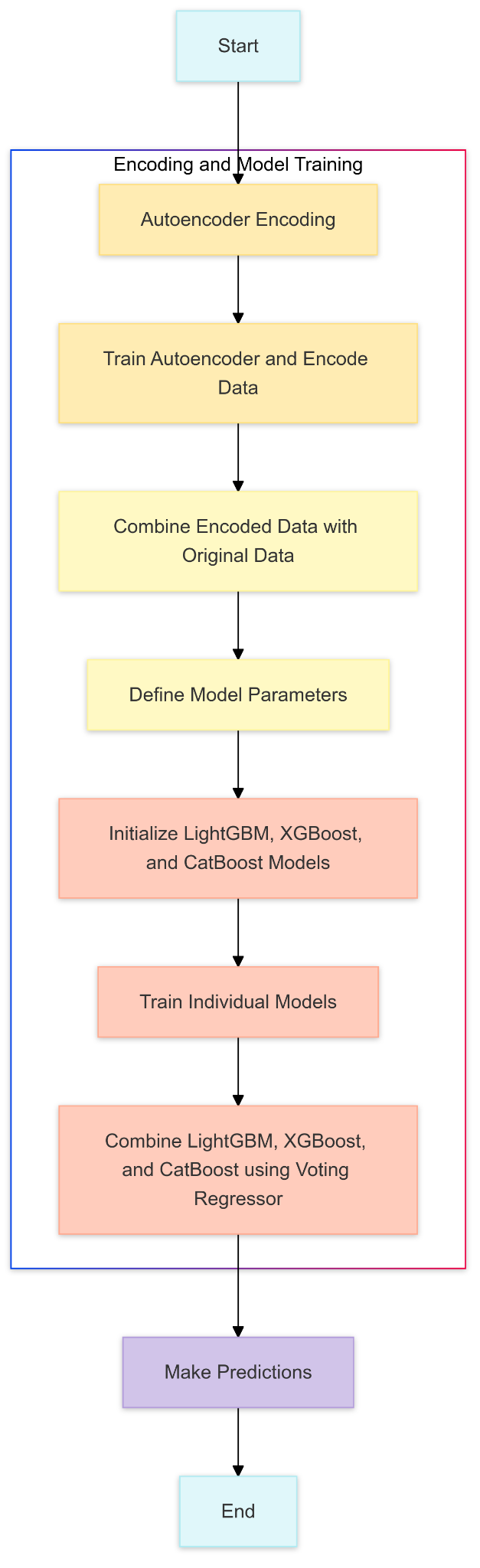
## 

## **Comparison of State of Art Architectures:**

| **Architecture** | **Description** | **Advantages** | **Disadvantages** | **Application in Project** |
| --- | --- | --- | --- | --- |
| **Vanilla Autoencoder + LightGBM + XGBoost + CatBoost** | Dimensionality reduction and boosting ensemble | High accuracy, handles noise | High training time | Versatile ensemble, useful in high-dimensional data |
| **Sparse Autoencoder + LightGBM + XGBoost + CatBoost** | Sparse features, boosting models | Compact features, robust models | Computationally heavy | Ideal for sparse data with boosting-based ensembles |
| **Sparse Autoencoder + LightGBM + XGBoost + CatBoost + RandomForest + GradientBoost** | Sparse Autoencoder with diverse tree-based ensemble | Robust, handles noise, high accuracy | High resource requirement | High-dimensional tasks needing model diversity |
| **Sparse Autoencoder + ADA +GPR + ETR + XGB + ElasticNet** | Sparse features with ensemble of boosting, reg, RF | Robust ensemble, good for noisy data | Complex, computationally heavy | Suitable for complex regression with noise |
| **SAE+LightGBM** | Sparse features with LightGBM | Efficient, interpretable | Limited diversity, may miss complex patterns | Suitable for high-dimensional tabular data |
| **SAE+XGBoost** | Sparse Autoencoder + XGBoost | High accuracy, handles non-linear relationships | Complex tuning, high computation. | Best for complex regression with non-linear data |
| **SAE+CatBoost** | Sparse features with CatBoost | Efficient, handles categorical data | Limited diversity | Structured data with categorical features |

## **Our model architecture:**

**SPARSE AUTOENCODER:**

* **Encoder**: Consists of four fully connected layers:
  + - **Layer 1**: Linear transformation to 256 units, followed by Layer Normalization, Leaky ReLU activation, and Dropout.
    - **Layer 2**: Linear transformation to 128 units, followed by Layer Normalization, Leaky ReLU activation, and Dropout.
    - **Layer 3**: Linear transformation to 64 units, followed by Layer Normalization, Leaky ReLU activation.
    - **Layer 4**: Linear transformation to 32 units, followed by Layer Normalization and Leaky ReLU activation.
* **Decoder**: Mirrors the encoder with four fully connected layers:
  + - **Layer 1**: Linear transformation from 32 to 64 units, followed by Layer Normalization, Leaky ReLU activation, and Dropout.
    - **Layer 2**: Linear transformation from 64 to 128 units, followed by Layer Normalization, Leaky ReLU activation, and Dropout.
    - **Layer 3**: Linear transformation from 128 to 256 units, followed by Layer Normalization and Leaky ReLU activation.
    - **Output Layer**: Linear transformation back to input\_dim, followed by a Sigmoid activation function to ensure output values are in the range [0, 1].
  + 

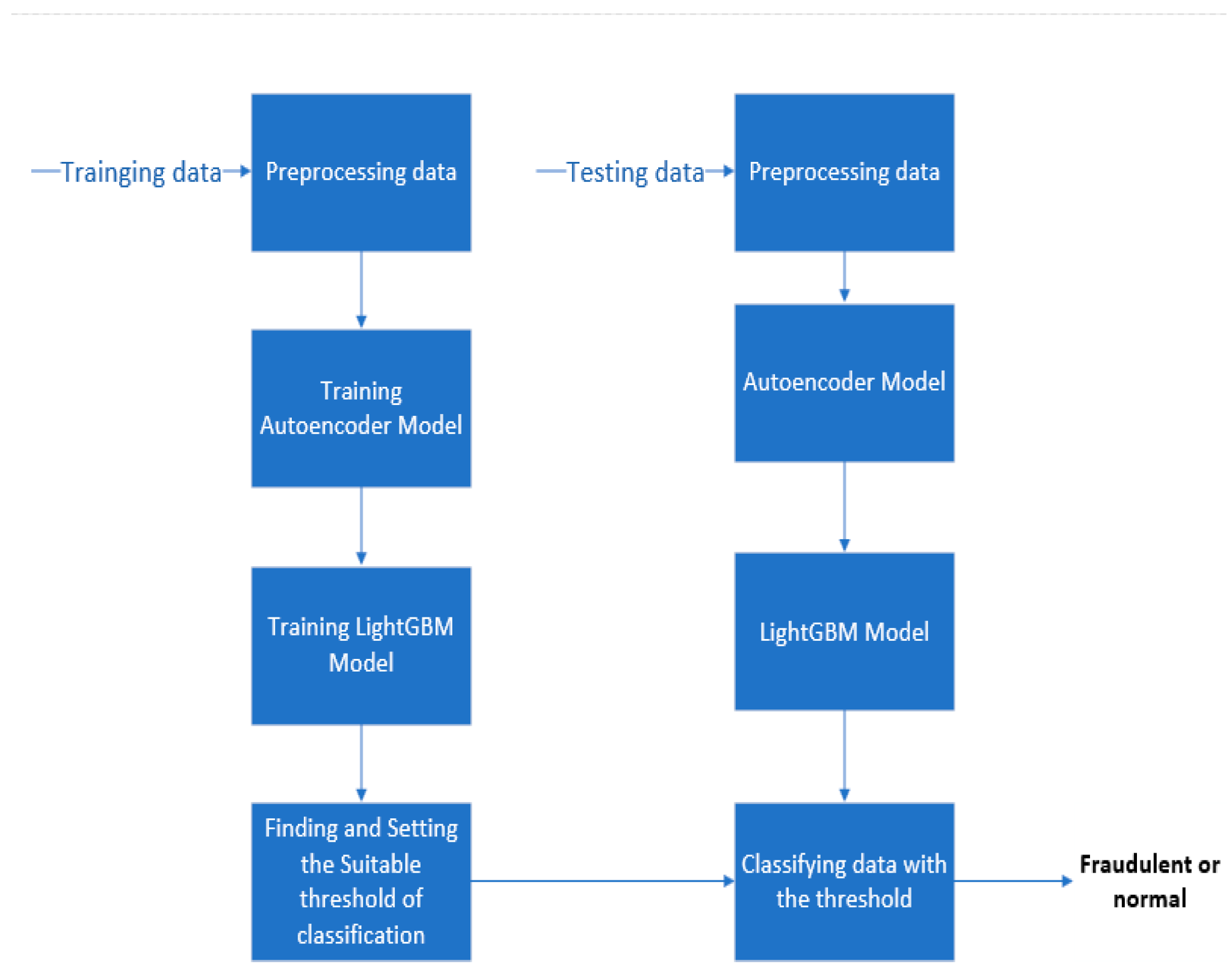
## **Variations attempted:**

Several variations were attempted during the project to enhance the model

1. **Enhanced Normalization**: Layer normalization improves training stability.
2. **Advanced Activation Function**: Leaky ReLU adds non-linearity and improves gradient flow.
3. **Regularization**: Dropout helps prevent overfitting.
4. **Encouragement of Sparse Representations**: Aims for more efficient encodings.
5. **Complexity and Depth**: More layers allow for better feature extraction.

## **Baseline Architecture: (Vanilla Autoencoder)**

* **Input Layer**: Receives the original data, often a feature vector with a specific dimension (e.g., 784 for 28x28 image pixels).
* **Encoder**:
  + - **Hidden Layer(s)**: One or more dense layers with fewer neurons than the input dimension, gradually reducing the dimensionality.
    - **Bottleneck (Latent Representation)**: The final layer of the encoder network, often with the smallest number of neurons. This compressed latent space holds a semantically meaningful representation of the input data.
    - **Activation Functions**: Typically uses simple activation functions like ReLU or Sigmoid.
* **Decoder**: **Layer 1**:
  + - **Hidden Layer(s)**: One or more dense layers that gradually increase dimensionality, mirroring the encoder's structure.
    - **Output Layer**: A dense layer with the same dimension as the input layer to produce the reconstructed output.
    - **Activation Function**: Generally Sigmoid (if the input data is normalized between 0 and 1) or linear for regression tasks.



**Comparison of results:**

| **Model** | **Mean Train QWK** | **Mean Validation QWK** | **Optimized QWK SCORE** |
| --- | --- | --- | --- |
| **Vanilla Autoencoder + LightGBM + XGBoost + CatBoost** | 0.8108 | 0.4885 | 0.541 |
| **Sparse Autoencoder + LightGBM + XGBoost + CatBoost** | 0.8115 | 0.4860 | 0.539 |
| **Sparse Autoencoder + LightGBM + XGBoost + CatBoost + RandomForest + GradientBoost** | 0.9175 | 0.3803 | 0.450 |
| **Sparse Autoencoder + ADA +GPR + ETR +XGB + ElasticNet** | 0.8518 | 0.4457 | 0.457 |
| **SAE + LightGBM** | 0.8603 | 0.5051 | 0.531 |
| **SAE + XGBoost** | 0.9255 | 0.4996 | 0.538 |
| **SAE + CatBoost** | 0.5163 | 0.4271 | 0.512 |

**Quadratic Weighted Kappa (QWK) Scoring**

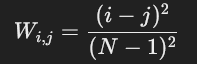
Submissions are scored based on the quadratic weighted kappa, a metric that measures the agreement between two outcomes. The QWK typically ranges from 0 (indicating random agreement) to 1 (indicating complete agreement). In some cases, when there is less agreement than expected by chance, the QWK can be negative.

**Steps to Compute QWK:**

**Matrix Construction:**

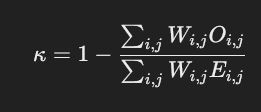
**O (Observed matrix):** An N×N histogram matrix where Oi,jOi,j​ represents the number of instances with an actual value ii and a predicted value jj.

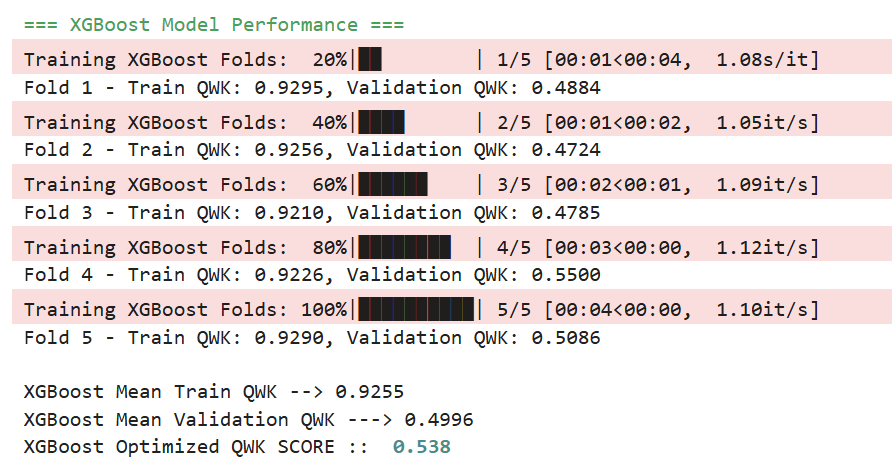
**W (Weight matrix):** An N×N matrix with weights calculated using the formula:



**E (Expected matrix):** An N×N histogram matrix of expected outcomes. It is computed assuming no correlation between actual and predicted values:

**E**=outer product of the actual and predicted histogram vectors, normalized This ensures that EEE and OOO have the same sum.

**Quadratic Weighted Kappa Calculation:** The QWK is computed using the formula:

**Maximum Result:**  


**Future Scopes:**

* Integration with Wearable Technology
* Scope: Develop models to run on wearable devices for real-time monitoring.
* Benefit: Immediate feedback and interventions for problematic internet use.
* Personalized Interventions
  + Scope: Tailor interventions based on individual physical and mental health data.
  + Benefit: More effective strategies for reducing problematic internet use.
* Cross-Disciplinary Approaches
  + Scope: Collaborate with experts in psychology, education, and technology.
  + Benefit: Holistic understanding and addressing of internet use issues.
* Use of Natural Language Processing (NLP)
  + Scope: Analyze text data from social media or online communication for early signs of problematic behavior.
  + Benefit: Provides an additional layer of behavioral insights.

**References:**<https://www.researchgate.net/publication/375902058_A_Hybrid_Ensemble_Learning_Approach_Utilizing_Light_Gradient_Boosting_Machine_and_Category_Boosting_Model_for_Lifestyle-Based_Prediction_of_Type-II_Diabetes_Mellitus>  
  
<https://www.sciencedirect.com/science/article/abs/pii/S0957582024006487>