# HSA 2024 Dataset - Metadata Description

**Dataset Title:** High School Assessment (HSA) 2024 Student Academic Records - English Version

**Version:** 1.0

**Date:** December 2024

**Authors:** Educational Data Research Team

**Institution:** Vietnam National University Repository

## Executive Summary

The HSA 2024 Dataset represents a comprehensive collection of academic performance data from Vietnamese high school students who participated in the High School Assessment examination in 2024. This dataset contains detailed academic records spanning three years of high school education (grades 10-12) for 57,174 students, encompassing 115 variables that capture academic performance, demographic information, and institutional characteristics. The dataset serves as a valuable resource for educational research, particularly in the areas of academic achievement prediction, educational data mining, and learning analytics.

This English version of the dataset has been standardized using international terminology to facilitate publication in Q1 journals and cross-cultural educational research. All categorical variables have been translated and standardized according to international educational assessment frameworks, while maintaining the analytical integrity of the original data.

The dataset is meticulously structured to support reproducible research and has been anonymized to ensure compliance with ethical standards and privacy regulations. All student identities have been removed while preserving the analytical value of the academic and demographic information. The data collection process followed rigorous protocols to ensure data quality, completeness, and consistency across all participating institutions.

## Dataset Overview

### Scope and Coverage

The HSA 2024 Dataset encompasses academic records from high school students across Vietnam who took the High School Assessment examination in 2024. The dataset provides a comprehensive view of student academic performance throughout their three-year high school journey, from grade 10 through grade 12. The geographic coverage spans all 63 provinces and cities of Vietnam, representing diverse educational contexts including urban and rural schools, different socioeconomic backgrounds, and various institutional types.

The temporal scope of the dataset covers academic performance data from the 2021-2022 school year (grade 10) through the 2023-2024 school year (grade 12), culminating in the HSA examination results. This longitudinal perspective enables researchers to track academic progression and identify patterns in student performance over time.

### Data Collection Methodology

The data collection process was conducted through a systematic approach involving multiple stakeholders in the Vietnamese education system. The primary data sources included:

**School Administrative Systems:** Academic records were extracted from official school management systems maintained by participating high schools. These systems contain comprehensive student academic information including semester grades, annual assessments, and behavioral evaluations.

**Ministry of Education and Training (MOET) Database:** Official HSA examination results were obtained from the centralized database maintained by MOET, ensuring accuracy and completeness of the outcome variable.

**Institutional Records:** School-level information including geographic location, institutional characteristics, and regional classifications were collected from official educational directories and administrative databases.

The data collection protocol involved several quality assurance measures to ensure data integrity and consistency. All participating schools were required to follow standardized data extraction procedures, and multiple validation checks were implemented to identify and correct potential data quality issues.

## Data Processing and Quality Assurance

Prior to inclusion in the final dataset, all collected data underwent rigorous processing and quality assurance procedures. The data processing pipeline included several critical steps designed to ensure data quality, consistency, and analytical utility.

**Data Validation:** Each record was subjected to comprehensive validation checks to identify missing values, outliers, and inconsistencies. Range checks were applied to all numerical variables to ensure values fell within expected bounds. For example, grade point averages were verified to be within the 0-10 scale used in Vietnamese education, and categorical variables were checked against predefined value sets.

**Standardization and Translation:** Academic performance measures were standardized across different schools and regions to account for potential variations in grading practices. Additionally, all categorical variables were translated from Vietnamese to English using internationally recognized educational terminology:

* Gender categories: "Nam" → "Male", "Nữ" → "Female"
* Academic performance levels: "Giỏi" → "Excellent", "Khá" → "Good", "Trung bình" → "Average", "Yếu" → "Poor"
* Conduct ratings: "Tốt" → "Good", "Khá" → "Satisfactory", "Trung bình" → "Average", "Yếu" → "Poor"

**Anonymization:** A comprehensive anonymization process was implemented to protect student privacy while preserving analytical value. All direct identifiers such as student names, identification numbers, and specific school names were removed or replaced with anonymous codes. Quasi-identifiers were carefully evaluated and modified as necessary to prevent re-identification while maintaining data utility.

**Completeness Assessment:** The dataset demonstrates exceptional completeness with zero missing values across all 115 variables. This completeness is the result of careful data collection protocols and extensive data cleaning procedures that addressed any initial gaps in the source data.

## Data Structure and Organization

### File Structure

The HSA 2024 Dataset is organized in a systematic directory structure designed to facilitate ease of access and usage:

HSA2024\_Dataset\_English/

├── HSAdataset\_English.xlsx # Main dataset file

├── Variable\_Codebook\_HSA.xlsx # Comprehensive variable documentation

├── Metadata\_Description\_HSA.pdf # This documentation file

└── Data\_Conversion\_Summary.md # Translation and standardization report

### Variable Categories

The 115 variables in the dataset are organized into several logical categories that reflect different aspects of student academic experience and characteristics:

**Demographic and Geographic Variables (3 variables):**

* Gender (Male/Female)
* Province (63 provinces and cities across Vietnam)
* Region (Geographic regions of Vietnam)

**Grade 10 Academic Records (37 variables):**

* Overall GPA for first semester, second semester, and annual average
* Academic performance classification (Excellent, Good, Average, Poor)
* Conduct classification (Good, Satisfactory, Average, Poor)
* Subject-specific grades for nine core subjects: Mathematics, Literature, Physics, Chemistry, Biology, History, Geography, Civic Education, and Foreign Language
* Foreign language type indicator

**Grade 11 Academic Records (37 variables):**

* Identical structure to Grade 10 variables
* Overall GPA for first semester, second semester, and annual average
* Academic performance and conduct classifications
* Subject-specific grades for all nine core subjects
* Foreign language type indicator

**Grade 12 Academic Records (37 variables):**

* Identical structure to Grades 10 and 11
* Overall GPA for first semester, second semester, and annual average
* Academic performance and conduct classifications
* Subject-specific grades for all nine core subjects
* Foreign language type indicator

**Outcome Variable (1 variable):**

* HSA examination score - the primary outcome measure for educational research

### Measurement Scales and Data Types

The dataset employs multiple measurement scales appropriate to the nature of each variable:

**Ratio Scale Variables:** Academic performance measures including GPA scores, subject-specific grades, and HSA examination scores are measured on a ratio scale using the Vietnamese 10-point grading system. These variables allow for meaningful mathematical operations and statistical analyses.

**Ordinal Scale Variables:** Academic performance classifications (Excellent, Good, Average, Poor) and conduct classifications (Good, Satisfactory, Average, Poor) represent ordered categorical variables that reflect hierarchical performance levels.

**Nominal Scale Variables:** Demographic characteristics such as gender, provinces, regions, and foreign language types are measured on a nominal scale, representing distinct categories without inherent ordering.

## Methodological Details

### Sampling Framework

The HSA 2024 Dataset represents a comprehensive census rather than a sample, including all students who completed the HSA examination in 2024 and had complete academic records available from their three-year high school experience. This census approach eliminates sampling bias and provides a complete picture of the target population.

The inclusion criteria for the dataset were:

* Completion of the HSA examination in 2024
* Availability of complete academic records for grades 10, 11, and 12
* Enrollment in Vietnamese high schools following the standard curriculum
* Consent for data use in research (obtained through institutional agreements)

### Data Collection Protocols

The data collection process followed a standardized protocol developed in collaboration with the Ministry of Education and Training and participating educational institutions. The protocol specified:

**Timing:** Data collection occurred in two phases. Academic records for grades 10-12 were collected at the end of the 2023-2024 academic year, while HSA examination results were obtained following the official release of examination scores.

**Data Sources:** Multiple data sources were utilized to ensure comprehensiveness and accuracy. School administrative systems provided academic performance data, while the centralized MOET database supplied HSA examination results and institutional information.

**Quality Control:** Each participating institution designated trained personnel responsible for data extraction and initial quality checks. Standardized data extraction templates were provided to ensure consistency across institutions.

**Verification Procedures:** Cross-validation procedures were implemented to verify data accuracy. Academic performance data was cross-checked against official transcripts, and HSA scores were verified against official examination records.

### Anonymization and Translation Procedures

The anonymization and translation process was designed to protect student privacy while creating an internationally accessible dataset. The procedures implemented include:

**Direct Identifier Removal:** All direct identifiers including student names, national identification numbers, and specific addresses were completely removed from the dataset.

**Categorical Variable Translation:** All categorical variables were systematically translated using internationally recognized educational terminology. The translation process followed established frameworks from educational assessment literature and Q1 journal standards.

**Geographic Generalization:** While province-level information was retained for regional analysis, more specific geographic identifiers were removed to prevent re-identification.

**K-Anonymity Assessment:** The final dataset was evaluated to ensure k-anonymity with k≥5 for all combinations of quasi-identifiers, providing strong privacy protection.

## Ethical Considerations

The development and release of the HSA 2024 Dataset adhered to strict ethical guidelines and regulatory requirements:

**Institutional Review:** The data collection and processing procedures were reviewed and approved by relevant institutional review boards and educational authorities.

**Consent and Authorization:** Data use agreements were established with all participating institutions, ensuring proper authorization for research use of student academic records.

**Privacy Protection:** Comprehensive privacy protection measures were implemented throughout the data lifecycle, from initial collection through final dataset preparation and international standardization.

**Data Security:** Secure data handling procedures were maintained throughout the project, including encrypted data transmission, secure storage systems, and restricted access controls.

**Transparency:** This metadata documentation provides complete transparency regarding data collection, processing, anonymization, and translation procedures to support informed use of the dataset.

## Usage Guidelines and Recommendations

### Intended Use Cases

The HSA 2024 Dataset is designed to support a wide range of educational research applications and analytical purposes:

**Academic Achievement Prediction:** Researchers can utilize the longitudinal academic performance data to develop predictive models for HSA examination outcomes. The three-year progression of grades provides rich temporal information for understanding academic trajectory patterns and identifying early indicators of academic success or risk.

**Educational Data Mining:** The comprehensive nature of the dataset makes it suitable for data mining applications aimed at discovering patterns in student academic behavior, identifying factors associated with academic success, and understanding relationships between different academic subjects and overall performance.

**Cross-Cultural Educational Research:** The English standardization of categorical variables enables comparison with international educational datasets and facilitates publication in international journals and conferences.

**Learning Analytics:** The dataset supports learning analytics research focused on understanding student learning patterns, academic progression trajectories, and the relationship between different aspects of academic performance over time.

**Educational Policy Research:** Policymakers and educational researchers can use the dataset to examine the effectiveness of educational interventions, understand regional variations in academic performance, and inform evidence-based policy decisions.

### Analytical Considerations

Researchers utilizing the HSA 2024 Dataset should consider several important analytical factors:

**Temporal Dependencies:** The longitudinal nature of the data requires careful consideration of temporal dependencies and autocorrelation in academic performance measures. Time series analysis techniques may be appropriate for examining academic progression patterns.

**Hierarchical Structure:** Students are nested within schools and regions, creating a hierarchical data structure that may require multilevel modeling approaches to account for clustering effects and institutional influences.

**Cultural Context:** While the dataset has been standardized for international use, researchers should be aware of the Vietnamese educational context and consider cultural factors when interpreting results or making cross-cultural comparisons.

**Translation Effects:** The standardization process may have introduced subtle changes in meaning. Researchers should refer to the Data Conversion Summary for detailed information about translation decisions.

### Recommended Statistical Approaches

Based on the structure and characteristics of the HSA 2024 Dataset, several statistical approaches are particularly well-suited for analysis:

**Regression Analysis:** Multiple regression models can be used to examine relationships between academic performance variables and HSA outcomes, with careful attention to multicollinearity among predictor variables.

**Machine Learning Methods:** The large sample size and comprehensive variable set make the dataset suitable for machine learning applications including random forests, support vector machines, and neural networks for predictive modeling.

**Longitudinal Analysis:** Growth curve modeling and other longitudinal analysis techniques can be used to examine academic progression patterns and identify factors associated with different trajectory types.

**Cluster Analysis:** Unsupervised learning methods can be applied to identify distinct groups of students based on academic performance patterns and demographic characteristics.

## Technical Specifications

### File Format and Structure

**Primary Dataset File:** HSAdataset\_English.xlsx

* Format: Microsoft Excel (.xlsx)
* Size: Approximately 34.2 MB
* Encoding: UTF-8
* Worksheet Structure: Single worksheet containing all 57,174 records

**Variable Codebook:** Variable\_Codebook\_HSA.xlsx

* Format: Microsoft Excel (.xlsx) with multiple worksheets
* Worksheets: Variable\_Definitions, Descriptive\_Statistics, Dataset\_Overview
* Comprehensive documentation for all 115 variables

**Metadata Documentation:** Metadata\_Description\_HSA.pdf

* Format: Portable Document Format (.pdf)
* Content: Complete methodological documentation and usage guidelines

### Data Quality Metrics

The HSA 2024 Dataset demonstrates exceptional data quality across multiple dimensions:

**Completeness:** 100% complete data across all 115 variables and 57,174 records

**Consistency:** Standardized formatting and encoding across all variables

**Accuracy:** Multiple validation checks and cross-verification procedures

**Timeliness:** Current data reflecting the most recent HSA examination cycle

**Validity:** All variables validated against expected ranges and categorical values

### System Requirements

**Software Compatibility:**

* Microsoft Excel 2016 or later
* R (version 4.0 or later) with readxl package
* Python (version 3.7 or later) with pandas library
* SPSS (version 25 or later)
* Stata (version 15 or later)

**Hardware Requirements:**

* Minimum 8GB RAM for optimal performance
* 1GB available disk space
* 64-bit operating system recommended for large-scale analyses

## Citation and Attribution

When using this dataset in research publications, please cite as:

Educational Data Research Team. (2024). HSA 2024 Dataset - High School Assessment Student Academic Records (English Version). Vietnam National University Repository. Version 2.0.

## Contact Information

For questions about the dataset, technical support, or collaboration opportunities, please contact:

**Dataset Maintainer:** Educational Data Research Team

**Institution:** Vietnam National University, Ha Noi

**Email:** tiennb@vnu.edu.vn

## Version History