CHAPTER 3

RESEARCH METHODOLOGY

3.1 Introduction

This chapter systematically expounds the research methodology and technical implementation path for intelligent prediction of university course satisfaction. Centering on the multi-source integration of structured scoring data and unstructured text comments, the entire text focuses on the design of the "from data to intelligent decision-making" process. This chapter first clarifies the research framework and stage division, then elaborates on the data sources and feature structures in detail, and explains the data preprocessing, feature extraction and fusion strategies, model construction and experimental process in stages. Finally, it discusses the model evaluation system and explainability methods.

3.2 Research Framework

This study proposes a methodology framework for intelligent course satisfaction prediction that integrates structured scoring and text comments, and features both automation and interpretability. The framework strictly adheres to the research paradigms of data science and educational evaluation, systematically covering the entire process from problem definition, data processing, feature construction, model training to interpretation and analysis, aiming to enhance the scientific nature, practical decision-making support capabilities, and management transparency of the model. The

overall research process is shown in Figure 3.1 and consists of the following six stages:

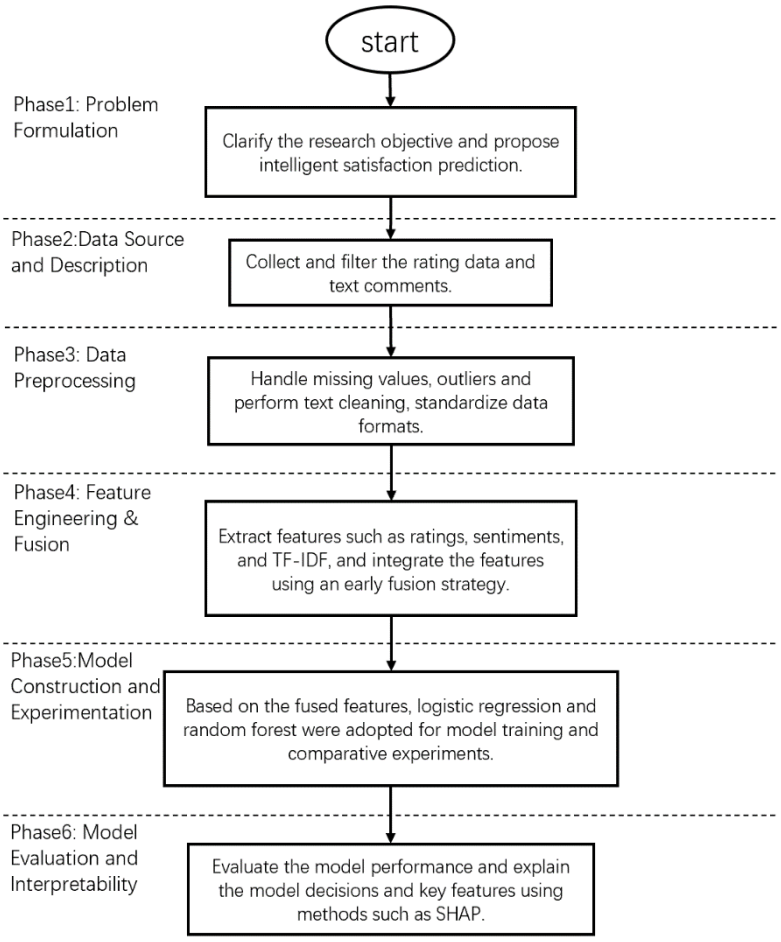

Figure 3.2 Framework diagram of research methodology workflow

Phase 1: Problem Formulation

This study clearly defines its core scientific issue, which is how to achieve intelligent and highly interpretable prediction of university course satisfaction by integrating structured student ratings and unstructured text comments. Based on previous literature and practical needs, this paper systematically reviews the shortcomings of existing satisfaction evaluation methods in terms of fine-grained emotion, subjective bias, and data fusion, and proposes the theoretical goals and practical significance of this research.

Phase 2: Data Source and Description

Select a real student evaluation dataset that covers a rich variety of courses, teachers, and student backgrounds, including rating items (such as Quality, Difficulty, Would Take Again) and open-ended text comments. Based on project requirements, conduct field screening on the original data, focusing on the most representative key variables, and systematically describe the sample structure, variable types, and initial distribution to lay a data foundation for subsequent analysis.

Phase 3: Data Preprocessing

For the distinct attributes of structured data and text data, separate preprocessing plans are formulated for data cleaning, standardization, handling of missing and outlier values, and text normalization, to ensure the high quality, consistency and analyzability of the model input data. At this stage, the verification of data correspondence and sample deduplication are also emphasized to enhance the overall rigor of the data engineering.

Phase 4: Feature Engineering

Combining structured features such as course ratings and learning difficulty with text features like sentiment polarity, TF-IDF, and text length, the system designs a feature extraction and fusion method. Through early fusion, multi-source features are encoded and integrated into a unified vector, enhancing the model's expressive power and its ability to capture complex student feedback.

Phase 5: Modeling and Experimentation

Based on the fused features, multiple machine learning and deep learning models such as random forest, logistic regression, and LSTM were adopted to conduct experiments on course satisfaction prediction. Through cross-validation and parameter optimization, the stability and generalization ability of the models were ensured, and comparative experiments were designed to evaluate the performance differences of different feature combinations and algorithms.

Phase 6: Model Evaluation and Interpretability

The performance of the model is evaluated by comprehensively applying indicators such as accuracy rate and F1 score, and introducing explainability analysis tools like SHAP to systematically dissect the key influencing features of the model's prediction results. By combining feature importance ranking with local explanation visualization, the transparency of the model is further enhanced, providing evidence-based intelligent decision-making references for educational management practices.

In summary, the research method framework of this study takes "multi-source data fusion, deep feature modeling and result interpretability" as the main thread, and advances the theoretical innovation and practical application of intelligent course satisfaction analysis in stages. This process not only conforms to the current development trend of the intersection of data science and educational technology, but also provides a solid methodological foundation for subsequent experimental links and application analysis.

3.3 Problem Formulation

This study aims to leverage modern data analysis techniques, specifically natural language processing and machine learning, to integrate structured rating data and text comments for the intelligent prediction of university course satisfaction. By doing so, it endeavors to offer more valuable insights to educational administrators and decision - makers.However, attaining accurate and interpretable outcomes is beset with several challenges. These include the intricate nature of emotional expression in text, the high degree of data heterogeneity, and the difficulty in integrating structured and unstructured information.

I. Structured scoring and text comments have significant differences in data representation and information structure. It is necessary to ensure data quality and consistency through means such as missing value handling, data cleaning, and standardization. Secondly, the emotional expression in text comments is complex and

highly subjective. Traditional manual or simple rule-based methods are difficult to efficiently and accurately extract effective information. Therefore, this study adopts natural language processing and sentiment analysis techniques to achieve automated feature extraction. II. The effective integration of structured and unstructured features is the key to improving model performance. It is necessary to design reasonable feature engineering methods to fully leverage the advantages of various types of information. Finally, to balance the accuracy and interpretability of the model, this study not only uses high-performance algorithms such as random forests but also introduces interpretability tools such as SHAP to facilitate transparency and decision support of the results.

Focusing on these core issues, this paper pursues intelligence, integration, and interpretability. It presents a methodological framework grounded in "multi - source data fusion, deep semantic modeling, and result interpretation". This framework aims to improve the accuracy and generalization ability of satisfaction prediction while simultaneously taking into account its practical utility in educational management and decision - making support.

3.4 Data Source and Description

The data used in this study is sourced from the public education evaluation platform RateMyProfessor, covering student feedback on various courses and instructors from multiple universities. The original data consists of two major categories: the first is structured rating data, mainly including core variables such as course quality (Quality), course difficulty (Difficulty), and "Would Take Again"; the second is unstructured text comments, where students can freely express their subjective feelings about course content, teaching methods, and personal experiences.

Table 3.4: Variable Description and Examples

Field Name	Data Type	Example Value	Description
professor_name	Text	Leslie Looney	Name of the professor being

			evaluated
star_rating	Numeric	4.7	Overall course rating given by students (1–5 scale, supports decimals)
diff_index	Numeric	2	Course difficulty index as assessed by the professor
student_difficult	Numeric	3	Course difficulty as perceived by students (subjective rating, 1–5 scale)
would_take_agains	Categorical	Yes	Whether the student would take the course again (Yes/No)
comments	Text	This class is hard...	Free-text review written by students, containing detailed subjective feedback and sentiments

3.5 Data Pre-Processing

In the data preprocessing phase of this research, a comprehensive and systematic processing protocol was implemented. This involved integrating both structured scoring and unstructured text information derived from the actual dataset. As depicted in Figure 3.X, the entire data preprocessing process consists of several crucial steps. These include dealing with missing and outlier values, removing duplicate samples, standardizing and encoding the data, cleaning and normalizing the text, and validating sample consistency.

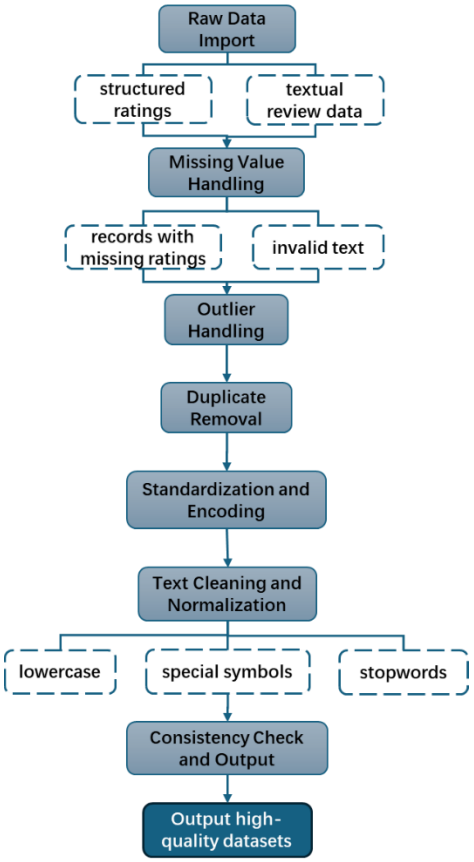

Figure 3.1 Data Preprocessing Flowchart

Step 1: Handling Missing and Outlier Values

Firstly, a comprehensive check for missing values was conducted on key fields in the dataset, including professor_name, star_rating, diff_index, student_difficult, would_take_again, and comments. Through statistical analysis, it was found that a very small number of samples had missing values in key variables such as ratings, difficulty, or text comments. To ensure the integrity and validity of the data analysis, all samples with missing values in the above fields were removed. Additionally, outlier screening was performed on numerical variables such as star_rating, diff_index, and student_difficult. Considering that the theoretical rating range should be between 1 and 5, all values outside this range were removed using descriptive statistics and visualization methods. For the would_take_again categorical field, the value format was standardized, and Yes/No was mapped to binary encoding to ensure standardized and consistent data input.

Step 2: Duplicate Value Removal and Variable Standardization

After the initial cleaning, duplicate values in the dataset were detected and all redundant samples with identical fields were removed to prevent the influence of duplicate data on subsequent statistical inference and model training. For numerical rating fields, the z-score standardization method was used to normalize star_rating, diff_index, and student_difficult to a standard normal distribution with a mean of 0 and a variance of 1, eliminating the dimensional differences between different features. For categorical variables, binary encoding or one-hot encoding was applied to ensure that all variables were in a format suitable for input into subsequent machine learning models.

Step 3: Text Cleaning and Data Consistency Verification

For the unstructured comments field, systematic text cleaning and normalization were carried out. This included converting all text to lowercase, removing HTML tags, punctuation, and special characters, eliminating extra spaces, and deleting high-frequency words without actual meaning based on an English stop word list. At the same time, a lower limit was set for the text length to filter out overly short or invalid comments, ensuring that each comment had analytical value. After preprocessing, a consistency check was conducted on all samples to ensure that each rating data corresponded to its corresponding text comment without information mismatch or omission. Ultimately, a high-quality dataset was formed, free of missing values, outliers, duplicates, with standardized structure and complete information.

3.6 Feature Fusion Strategy

To effectively integrate structured rating information with unstructured text features and achieve precise prediction of course satisfaction, this study designed a systematic feature extraction and fusion strategy. In terms of structured features, the main variables extracted include students' overall course ratings (star_rating), teachers' self-assessed difficulty (diff_index), students' self-assessed difficulty (student_difficult), and whether they would choose the course again (would_take_again). All these variables are numerical data ranging from 1 to 5.

Categorical variables were uniformly processed through binary encoding (e.g., Yes=1, No=0). To eliminate the influence of different feature scales, all structured numerical features were standardized using the z-score method to ensure consistent distribution.

For text features, this study first employed the VADER (Valence Aware Dictionary and sEntiment Reasoner) sentiment analysis tool to model the subjective sentiment orientation of English student comments (comments field). Specifically, by calling the SentimentIntensityAnalyzer in the nltk library, the compound score of each text was obtained as the sentiment polarity score, with a numerical range from -1 to 1, representing strongly negative to strongly positive. This feature quantifies the emotional information in students' subjective evaluations, supplementing the rational ratings with emotional expressions.

Furthermore, to deeply explore the text content, the TF-IDF (Term Frequency-Inverse Document Frequency) method was used to model the key words of all comments. Through sklearn.feature_extraction.text.TfidfVectorizer, with max_features set to 1000, ngram_range set to (1,2), and English stop words filtered, the most representative high-frequency keywords and phrases in the comments were extracted and transformed into sparse vector features to reflect the core concerns of students' text feedback. Additionally, the text length of each comment (such as the number of tokens after word segmentation) was counted as a supplementary feature reflecting the richness of expression and participation.

All text features were standardized simultaneously with the structured rating features. Finally, in the feature fusion stage, an Early Fusion strategy was adopted to concatenate the standardized structured features, text sentiment polarity scores, TF-IDF sparse vectors, and text length features into a unified high-dimensional feature vector, serving as the input for subsequent machine learning and deep learning models. Considering the high-dimensional nature of TF-IDF features, dimensionality reduction methods such as Principal Component Analysis (PCA) could be applied before model training to balance information integrity and computational efficiency.

3.7 Model Construction and Innovation

3.7.1 Comparison of Methods

3.7.1.1 Comparative Analysis

In recent years, for the intelligent prediction of course satisfaction in universities, existing studies have mainly used structured scores as the primary feature input and widely adopted traditional machine learning methods such as logistic regression, decision trees, and support vector machines to classify or regress students' course evaluations (Lopez-Cueva et al., 2024). Some literature has attempted to introduce sentiment dictionaries or basic sentiment scoring to a certain extent to make up for the limitations of structured data, but generally only used a single feature source as input, lacking systematic mining and deep modeling of open-ended text evaluations (Wilbrod & Joshua, 2024). In terms of feature fusion, most works still remain at simple concatenation or the sole use of score data, making it difficult to effectively capture the complex interaction between students' subjective text emotions and rational scores (Koufakou, 2023). Moreover, in the model training and evaluation process, existing research often fails to fully consider issues such as sample class imbalance, feature diversity, model interpretability, and parameter tuning, resulting in certain limitations in the generalization ability and decision transparency of the models (Oubraime et al., 2025).

Based on the above deficiencies, this study proposes a new strategy of multi-source feature input in the model construction phase, systematically integrating structured score variables (such as star_rating, diff_index, student_difficult, would_take_again) with high-dimensional features extracted based on natural language processing and sentiment analysis techniques, including text sentiment polarity, TF-IDF keyword weights, and text length. On this basis, two main classification algorithms, logistic regression (LR) and random forest (RF), are selected for the model: logistic regression has good interpretability and can quantify the influence direction and weight of each input variable on satisfaction prediction; while random forest can fully utilize high-dimensional, multi-type, and non-linear features to enhance the overall performance of the model and output feature importance

rankings. To ensure the scientific nature of model evaluation, the experimental process sets up a main model (multi-source feature input) and a comparison model (only structured features or only text features input), and uses cross-validation, stratified sampling, accuracy rate, and F1-score and other multi-dimensional indicators to systematically compare the applicability, efficiency, and interpretability of the models. In terms of parameter optimization, grid search is used to systematically tune the hyperparameters of logistic regression (such as regularization coefficient) and random forest (such as the number of trees and maximum depth), further enhancing the stability and generalization ability of the models.

3.7.1.2 Summarisation and Research Gap

In the model construction stage of intelligent prediction for course satisfaction in this study, not only have the limitations of previous research on single features, weak model interpretability and low generalization been broken through, but also systematic innovations have been made in multi-source feature fusion, model evaluation and optimization. However, the current model still faces challenges such as the deep mechanism explanation of feature interaction, the handling of extremely imbalanced samples and the mining of higher-dimensional semantic features. Further exploration and improvement can be made in the directions of deep feature learning, sample augmentation and enhanced model interpretability in subsequent research.

Table 3.6.1.2 Summarisation of The Comparison Between Previous and Current Methods and The Research Gap

Aspect	**Previous Studies**	**Current Study**	**Research Gap**
Feature Sources	Structured ratings only; basic text sentiment	Fusion of ratings, NLP-based sentiment, TF-IDF, length	Deeper semantic/text feature extraction
Main Algorithms	LR, SVM, DT, sometimes basic ensemble	LR, RF; systematic parameter tuning, model comparison	Deep models, attention, or advanced fusion
Feature Fusion	Simple concatenation or rating only	Early fusion of multi-type features	Explore advanced fusion (e.g., attention)
Evaluation	Accuracy, sometimes recall/F1; little	Cross-validation, F1, accuracy, feature importance	Address class imbalance and more robust metrics

	cross-validation		
Interpretability	Limited (mostly LR coefficients)	Both LR weights & RF feature importance, SHAP planned	Enhanced interpretability (e.g., SHAP/LIME)
Optimization	Basic/manual parameter tuning	Grid search for key hyperparameters	Automated/advanced optimization techniques

3.6.1 Model Construction

During the model construction phase, this study focused on selecting two classic machine learning algorithms, Logistic Regression (LR) and Random Forest (RF), as the main models for the course satisfaction classification task based on multi-source fused features. Logistic Regression, with its simple structure and strong parameter interpretability, can intuitively display the influence direction and intensity of each input feature on satisfaction prediction, and is widely used in binary or multi-classification problems in fields such as education and social sciences. Random Forest, on the other hand, excels in modeling high-dimensional features and nonlinear relationships, has good robustness, and is resistant to overfitting. It is particularly suitable for complex data scenarios that integrate structured scores and sparse text features. Based on the idea of ensemble learning, it builds a large number of decision trees and integrates votes to significantly improve the generalization performance and prediction accuracy of the model, while also providing feature importance rankings for subsequent explainability analysis.

In the process of model selection and construction, this study fully referred to the literature review in Chapter Two, combined with the actual data scale and feature dimensions of this project, and systematically compared the applicability, computational efficiency, and compatibility with complex feature structures of various models. In the experimental design, the fused feature vectors were used as model inputs to construct the main models (LR and RF) and corresponding control experiments (such as only structured features, only text features, etc.). In terms of parameter setting, L2 regularization was adopted for Logistic Regression to prevent overfitting, and the penalty coefficient was optimized through cross-validation; for Random Forest, several different tree numbers (such as 100, 200, etc.) and maximum depths were set, and the optimal parameter combination was found through grid search.

To comprehensively evaluate the model performance, a cross-validation strategy was adopted in the training process to ensure the stability and generalization ability of the results. At the same time, stratified sampling was conducted based on the category distribution of the training set and test set to ensure the fairness of model evaluation. Finally, the outputs of all models were compared horizontally using metrics such as accuracy and F1 score, and in-depth analyses were conducted on the feature importance rankings and decision boundaries of each model.

3.8 Model Evaluation and Interpretability

In the stage of model evaluation and interpretability analysis, this research integrates state-of-the-art approaches from the domains of educational data mining and machine learning to establish a multi-level and systematic evaluation framework.

1. Performance Evaluation

In the aspect of performance evaluation, this study employs mainstream classification metrics, namely accuracy, precision, recall, and F1-score, to quantitatively assess the generalization ability of the logistic regression and random forest models in the task of predicting course satisfaction. Given the possible imbalance in category distribution within educational scenarios, particular emphasis is placed on balance metrics, such as the F1-score and confusion matrix. This is to comprehensively evaluate the model's discriminatory power and practical applicability across diverse categories.

Furthermore, by plotting ROC curves and calculating the Area Under the Curve (AUC), the overall performance of the models at various decision thresholds is further quantified. This significantly enhances the rigor and scientific nature of the evaluation system.

2. Interpretability Analysis

Regarding the interpretability analysis of the models, this study introduces the SHAP (Shapley Additive Explanations) framework to conduct in-depth global and local analyses of ensemble learning models (e.g., random forest) and linear models (e.g., logistic regression). Rooted in game theory, the SHAP method can assign accurate feature contribution values to each prediction result. This effectively reveals the key variables influencing the model's discriminatory outcomes and their underlying mechanisms.

Through visualizations such as SHAP summary plots and dependence plots, the importance and interactions of structured scores, text sentiment polarity, and other features during the prediction process are clearly depicted. Additionally, in combination with the regression coefficient analysis of logistic regression, the influence direction and significance of different features on course satisfaction are further verified from the perspective of statistical modeling. This interpretability work not only enhances the transparency and credibility of the models but also offers theoretical support for educational managers to comprehend the basis of algorithmic decisions, optimize the allocation of teaching resources, and implement personalized interventions.

3. Presentation of Analysis Results

All analysis results at this stage are presented using multiple visualization methods, including performance metric comparison charts, confusion matrix heatmaps, and SHAP feature importance rankings. This facilitates the intuitive understanding of the model's strengths and weaknesses by both the academic community and educational management practitioners.

Overall, through a rigorous evaluation system and interpretability analysis, this study not only ensures the scientific validity and practical value of the integrated models but also provides a solid methodological foundation for data-driven educational management and course improvement.

CONFERENCE

López-Cueva, J., Ares, S., García, M. C., & Martínez, F. (2024). A Comparative Study of Decision Tree and Logistic Regression for Predicting University Student Satisfaction. *Pakistan Journal of Life and Social Sciences, 22*(2), 7844–7856.https://doi.org/10.57239/PJLSS-2024-22.2.00591

Wilbrod, R., & Joshua, A. (2024). Sentiment Analysis of Student Feedback: An Implementation of a Natural Language Processing (NLP) Algorithm. *International Journal of Computer Applications, 47*(4), 58–65.https://www.researchgate.net/publication/385299533

Koufakou, A. (2023). Deep Learning for Opinion Mining and Topic Classification in Course Reviews. *arXiv preprint arXiv:2304.03394*. https://arxiv.org/abs/2304.03394

Oubraime, A., Oulad Haj Thami, R., & Chahhou, M. (2025). Predicting Student Satisfaction in Career Choices Using Machine Learning: A Case Study. *International Journal of Educational Technology in Higher Education, 22*, Article 56.https://www.researchgate.net/publication/388083302