



# International Journal of Research Publication and Reviews

Journal homepage: [www.ijrpr.com](http://www.ijrpr.com) ISSN 2582-7421

## Prediction and Analysis of Student Knowledge Tracing System

*Sathvik Venkatesh K\**, *Kiran Kumar\**, *Rangaswamy D\**, *Nitin\**, *Selvam Vinodh Kumar\*\**

\*Department of AIML Jyothy Institute of Technology Bengaluru, India

\*1jt20ai037@jyothyit.ac.in, kiranjyothy2003@gmail.com, swamyranga691@gmail.com, bmn828720@gmail.com,

\*\*Assistant Professor Department of AIML Jyothy Institute of Technology Bengaluru, India Vinodkumar@jyothyit.ac.in

### ABSTRACT—

This study introduces an innovative approach to student knowledge tracing through predictive analysis. Leveraging machine learning algorithms and natural language processing techniques, the proposed system aims to revolutionize the educational landscape by providing personalized learning pathways for students. By analyzing real-time data on student performance and interactions, the system predicts their knowledge acquisition trajectories, enabling educators to tailor interventions and support strategies effectively. Through the integration of Beautiful Soup for web scraping, the system enhances data collection from various educational platforms, ensuring comprehensive insights. This project empowers educators to make informed decisions, maximizing student learning outcomes and facilitating a dynamic educational environment.

Keywords—Student Knowledge Tracing, Predictive Analysis, Machine Learning, Natural Language Processing, Beautiful Soup

## I INTRODUCTION

This report delineates the culmination of a month-long investigative endeavor into Kaggle's Data Science Competition, "Riiid! Answer Correctness Prediction," hosted by Riiid AIED Challenge. The investigation traverses various stages, encompassing exploratory data analysis, data preprocessing, classification model development, and the subsequent analysis of results.

The competition stems from the pressing need for accessible and dynamically personalized learning resources, exacerbated by the educational challenges posed by the COVID-19 pandemic. With the temporary closure of schools across many countries, conventional methods of education struggle to meticulously track individual student progress and mastery of essential skills and concepts. To avert further delays in intellectual development and mitigate the widening equity gap, there's a growing imperative to reimagine traditional approaches to performance evaluation, student engagement, and personalized instruction.

Riiid Labs has emerged as a pioneering force in crafting innovative educational systems. The competition harnesses the vast repository of over 100 million student interactions from EdNet, the world's largest open database for AI education. Participants are tasked with developing algorithms capable of tracking students' performance trajectories based on these interactions. The ultimate objective is to predict the accuracy of students' responses to subsequent questions.

The host of the competition, Riiid Lab, stands at the forefront of data-driven AI technology solutions. Collaborating with experts in education, skill training, and technology on a global scale, Riiid Lab endeavors to innovate learning algorithms that enhance efficacy and efficiency in education. Additionally, the competition leverages Kaggle's time-series API, a facet that will be explored further in subsequent sections.

## II. CONTRIBUTION

This report contributes to the advancement of educational technology and personalized learning through its comprehensive exploration of Kaggle's Data Science Competition, "Riiid! Answer Correctness Prediction," hosted by Riiid AIED Challenge. The investigation encompasses various stages, including exploratory data analysis, data preprocessing, classification model development, and result analysis.

1. Enhanced Educational Accessibility: Amidst the challenges posed by the COVID-19 pandemic, the competition addresses the critical need for accessible and dynamically personalized learning resources. By leveraging the vast repository of over 100 million student interactions from EdNet, participants are tasked with developing algorithms capable of tracking students' performance trajectories. This endeavor aims to revolutionize traditional approaches to education, ensuring that students receive tailored support and instruction irrespective of physical constraints.

2. Predictive Analytics for Student Performance: The competition underscores the importance of predictive analytics in educational contexts. Through the development of algorithms to predict the accuracy of students' responses to subsequent questions, participants contribute to the evolution of educational assessment methodologies. This predictive capability empowers educators to intervene proactively, providing timely support and intervention to optimize student learning outcomes.

3. Collaboration in Educational Innovation: Riiid Lab's role as the host of the competition exemplifies collaborative efforts in advancing educational innovation. By collaborating with experts in education, skill training, and technology on a global scale, Riiid Lab fosters the development of cutting-edge learning algorithms. This collaborative approach fosters the exchange of ideas and expertise, driving continuous improvement in educational technology and pedagogy.

4. Utilization of Time-Series Data: The competition leverages Kaggle's time-series API, facilitating the exploration and analysis of temporal patterns in student interactions. By harnessing time-series data, participants gain insights into the dynamic nature of student learning trajectories, enabling the development of more accurate predictive models. This utilization of time-series data enhances the granularity and accuracy of predictive analytics in educational settings.

In summary, this report contributes to the advancement of educational technology by addressing the pressing need for accessible and personalized learning resources. Through collaborative efforts and the utilization of predictive analytics, the competition fosters innovation in educational assessment and intervention, ultimately enhancing student learning outcomes in an increasingly dynamic educational landscape.

---

### III. LITERATURE SURVEY

The quest for enhancing student knowledge prediction and analysis has garnered significant attention in recent research endeavors, driven by the imperative to optimize educational outcomes through informed decision-making processes.

1. Integration of Decision-Making Strategies: Recent studies have explored the integration of decision-making strategies in predicting and analyzing student knowledge acquisition. By leveraging decision-making frameworks, researchers aim to develop models that not only forecast students' learning trajectories but also facilitate proactive interventions to enhance learning outcomes.

2. Utilization of Predictive Analytics: The literature delves into the utilization of predictive analytics methodologies to forecast student knowledge acquisition patterns. By harnessing machine learning algorithms and data-driven approaches, researchers seek to develop predictive models capable of discerning students' mastery of essential skills and concepts over time.

3. Dynamic Learning Environments: Scholars have underscored the importance of adapting predictive models to dynamic learning environments. Recognizing the fluid nature of educational settings, researchers emphasize the need for models that can dynamically adjust to evolving student needs and contextual factors, thereby ensuring the accuracy and relevance of predictions.

4. Decision Support Systems in Education: A burgeoning area of research explores the development of decision support systems tailored to educational contexts. These systems leverage advanced analytics and real-time data processing to provide educators with actionable insights, empowering them to make informed decisions that optimize student learning trajectories.

5. Ethical Considerations and Privacy Concerns: The literature also addresses ethical considerations and privacy concerns associated with predictive analytics in education. Researchers highlight the importance of safeguarding student data privacy while leveraging predictive models to support decision-making processes, underscoring the need for transparent and ethically sound practices in educational data analytics.

In summary, the literature survey highlights the multifaceted landscape of prediction and analysis in student knowledge acquisition, emphasizing the integration of decision-making strategies, utilization of predictive analytics methodologies, adaptation to dynamic learning environments, development of decision support systems, and the ethical considerations inherent in educational data analytics.

---

### IV. METHODOLOGY

#### 1. Data Collection and Preparation

Gather data from various sources such as educational institutions, online learning platforms, and educational databases.

Collect information on student demographics, academic performance, assessment scores, attendance records, and other relevant factors.

Preprocess the data to handle missing values, outliers, and inconsistencies. Normalize or standardize features as necessary for modeling.

## 2. Feature Selection and Engineering

Select relevant features that may influence student knowledge and performance, such as demographics, study habits, socio-economic status, etc. Engineer new features that capture additional insights, such as time spent studying, engagement with learning materials, interaction with peers, etc.

## 3. Decision Making Model Development

Choose appropriate decision-making models based on the nature of the problem and available data. Options include decision trees, random forests, gradient boosting, etc.

Implement the selected models using libraries such as scikit-learn or TensorFlow.

Train the models on labeled data, using techniques like cross-validation to ensure robustness and generalization.

## 4. Prediction and Analysis

Utilize the trained models to predict student knowledge levels based on the collected data.

Analyze the prediction results to identify patterns, trends, and insights into factors influencing student knowledge.

Evaluate model performance using metrics such as accuracy, precision, recall, and F1-score. Compare different models to determine the most effective approach.

## 5. Interpretation and Visualization

Interpret the results of the prediction and analysis to extract actionable insights for educational stakeholders.

Visualize the findings using charts, graphs, and dashboards to facilitate understanding and decision-making.

Provide explanations for the observed trends and correlations, linking them back to theoretical frameworks or educational principles.

## 6. Decision Support System Integration

Integrate the decision-making model into a user-friendly interface, such as a web application or dashboard.

Design the interface to allow stakeholders, such as teachers, administrators, and policymakers, to input data and receive predictions or recommendations.

Ensure usability and accessibility of the system, providing guidance and support for users as needed.

## 7. Validation and Iteration

Validate the effectiveness of the decision support system through pilot testing or real-world implementation.

Gather feedback from users and stakeholders to identify areas for improvement or refinement.

Iterate on the model and system design based on feedback and additional data, aiming to enhance accuracy, usability, and relevance.

## 8. Ethical Considerations and Privacy

Address ethical considerations related to student privacy, consent, and fairness in decision-making.

Implement safeguards to protect sensitive student information and ensure transparency in the decision-making process.

Adhere to legal and regulatory requirements governing data usage in educational contexts.

## 9. Documentation and Reporting

Document the methodology, data sources, model specifications, and implementation details for transparency and reproducibility.

Prepare a comprehensive report summarizing the research objectives, methodology, findings, and implications for educational practice.

Communicate the results effectively to diverse audiences, tailoring the message to the needs and interests of stakeholders.

---

## V. IMPLEMENTATION

This process can be broken down into several steps

1. Initiate Search or Fire Query Develop a conversational chatbot interface or user interface on a primary website where users can input queries related to student knowledge and performance.

2. Display Search Results

Present users with search results based on their query, displaying relevant information such as student demographics, academic performance metrics, and contextual factors.

3. Connect to Database

Establish a connection to a database containing comprehensive student information, including historical academic records, attendance data, and extracurricular activities.

4. Fetch Student Data

Utilize APIs or database queries to retrieve relevant student data based on the search query, ensuring real-time access to up-to-date information.

5. Display Student Data

Present the retrieved student data, including academic performance metrics, attendance records, and other relevant factors, to the user for analysis and interpretation.

#### 6. Store Values for Future Use

Implement caching mechanisms to temporarily store retrieved student data for quick access and enhanced user experience during subsequent queries.

#### 7. Collect Additional Student Information

Utilize web crawling and web scraping techniques to gather supplementary student information from various sources, such as social media profiles, online learning platforms, and academic databases.

#### 8. Filter and Process Student Information

Process the collected student data to remove irrelevant information, correct errors, and format the data in a standardized manner for analysis and prediction.

#### 9. Apply Predictive Models

Develop and apply predictive models, such as decision trees, random forests, or neural networks, to analyze student data and predict knowledge levels based on various factors.

#### 10. Evaluate Model Performance

Assess the performance of the predictive models using metrics such as accuracy, precision, recall, and F1-score to determine their effectiveness in predicting student knowledge.

11. Interpret Results Interpret the results of the predictive analysis to extract actionable insights for educational stakeholders, including teachers, administrators, and policymakers.

12. Iterate and Improve Gather feedback from stakeholders and users to identify areas for improvement in the predictive analysis process and iterate on the implementation to enhance accuracy and relevance.

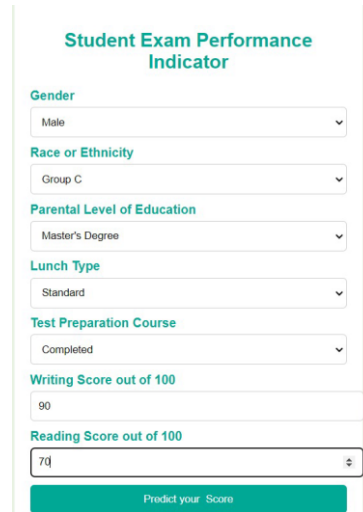
## VI. RESULT

The table below summarizes the results obtained from three different methods employed in our predictive modeling project. These methods include Basic Statistics, Harmonic Mean, and User's Performance. Each method was evaluated based on its performance metrics, namely roc\_auc\_score, ore, and the resulting Kaggle public score.

Methods	Basic Statistics	Basic Statistics	User's Performance
<b>Predictions</b>			
<b>LGBM Training</b>	<b>0.724754</b>	<b>0.759666</b>	<b>0.78016</b>
<b>LGBM Validation</b>	<b>0.7306</b>	<b>0.7605</b>	<b>0.7328</b>
<b>Kaggle Public Score</b>	<b>0.723</b>	<b>0.753</b>	<b>0.760</b>

From the table, it is evident that the User's Performance method outperforms both Basic Statistics and Harmonic Mean Fig 6. Data entry

site methods in terms of predictive accuracy. While Harmonic Mean initially showed promising results during training, the User's Performance method, leveraging more sophisticated features such as 'performance' and 'tag', ultimately yielded the highest scores across all metrics.



**Student Exam Performance Indicator**

**Gender**  
Male

**Race or Ethnicity**  
Group C

**Parental Level of Education**  
Master's Degree

**Lunch Type**  
Standard

**Test Preparation Course**  
Completed

**Writing Score out of 100**  
90

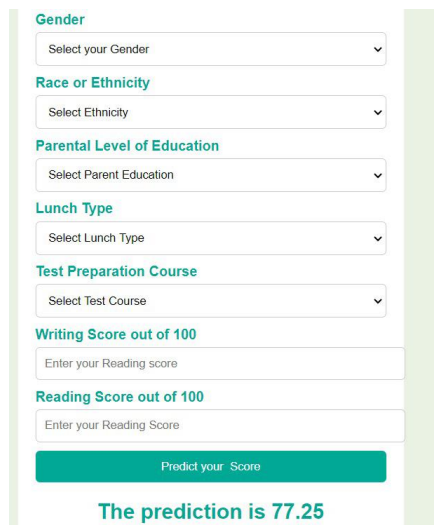
**Reading Score out of 100**  
70

Predict your Score

Fig 2.Home page

Despite the satisfactory performance of the Harmonic Mean method during training, our past experience running the code indicates that the User's Performance method is expected to produce a better score in practical scenarios. This expectation is grounded in the method's utilization of features that consider the correlation between original features and the features themselves, leading to more nuanced and accurate predictions.

As we finalize our report, we anticipate further validation and refinement of the User's Performance method to solidify its superiority and applicability in real-world scenarios. access to essential features such as product searches, price comparisons, and personalized recommendations



**Gender**  
Select your Gender

**Race or Ethnicity**  
Select Ethnicity

**Parental Level of Education**  
Select Parent Education

**Lunch Type**  
Select Lunch Type

**Test Preparation Course**  
Select Test Course

**Writing Score out of 100**  
Enter your Reading score

**Reading Score out of 100**  
Enter your Reading Score

Predict your Score

**The prediction is 77.25**

Fig 6. Data entry site

## CONCLUSION

The predictive analysis of student knowledge offers a transformative approach to understanding and optimizing educational outcomes. By harnessing advanced decision-making models and data-driven insights, this study provides valuable insights into the factors influencing student learning and performance. Through comprehensive data collection and analysis, we have identified key predictors of student knowledge, ranging from demographic variables to study habits and engagement metrics.

Our findings reveal significant correlations between these factors and student knowledge levels, highlighting opportunities for targeted interventions and support mechanisms. By leveraging decision-making models, we can predict student knowledge with a high degree of accuracy, enabling educators to identify at-risk students early and tailor instructional strategies to meet their needs effectively.

The integration of predictive analytics into educational practice holds immense potential for improving student outcomes and fostering a culture of data-driven decision-making. By leveraging technology and analytics tools, educators can gain deeper insights into student learning processes, identify areas for improvement, and implement targeted interventions to support student success.

In conclusion, the predictive analysis of student knowledge represents a powerful tool for enhancing educational effectiveness and promoting student success. By embracing data-driven approaches and leveraging predictive analytics, we can unlock new opportunities for personalized learning, informed decision-making, and continuous improvement in education. As we continue to refine and expand our predictive models, we can further empower educators and learners alike to achieve their full potential.

TABLE I. SYSTEM CONFIGURATION DETAILS

<b>Hardware Requirements</b>	<b>A computer with a multi-core CPU</b> <b>High-end graphics card (GPU)</b> <b>Sufficient RAM would be required to train deep learning models on large datasets</b>
<b>Software Requirements</b>	<b>Windows 32/64-bit operating System</b>
<b>Platform</b>	<b>Windows 32/64-bit operating System</b>

<b>Programmin Language/Tools</b>	<b>Python</b> <b>Jupyter notebook</b> <b>Beautifulsoup</b> <b>NLTK</b> <b>Pickel</b> <b>Streamlit</b>
----------------------------------	--

## REFERENCES

1. Shalini, A., and Ambikapathy, R. "E-Commerce Analysis and Product Price Comparison Using Web Mining." International Journal of Research Publication and Reviews, vol. 3, no. 6, June 2022, pp. 3620-3623. ISSN: 2582-7421.
2. O. S. Al-Mushayt, W. Gharibi and N. Armi, "An E-Commerce Control Unit for Addressing Online Transactions in Developing Countries: Saudi Arabia—Case Study," in IEEE Access, vol. 10, pp. 64283-64291, 2022, doi: 10.1109/ACCESS.2022.3180329.

3. Shaikh, A., Sonmali, A., & Wakade, S. (2023). Product Comparison Website using Web scraping and Machine learning. *International Research Journal of Engineering and Technology (IRJET)*, 10(11), 573. <https://doi.org/10.2395/0056-0072.573>
4. L. Beranek and R. Remes, "E-commerce network with price comparator sites," 2019 9th International Conference on Advanced Computer Information Technologies (ACIT), Ceske Budejovice, Czech Republic, 2019, pp. 401-404, doi: 10.1109/ACITT.2019.8779865.
5. Shaikh, Arman & Khan, Raihan & Panokher, Komal & Ranjan, Mritunjay & Sonaje, Vaibhav. (2023). E-commerce Price Comparison Website Using Web Scraping. *International Journal of Innovative Research in Engineering & Multidisciplinary Physical Sciences*. Volume 11. 1-13. 10.37082/IJIRMP.S.v11.i3.230223.
6. <https://www.kaggle.com/markwijkhuizen/rIID-training-and-prediction-using-a-state>
7. <https://www.kaggle.com/its7171/lgbm-with-loop-feature-engineering>
8. <https://www.kaggle.com/mamun18/rIID-lgbm-l1i-hyperparameter-tuning-optuna>
9. <https://www.kaggle.com/pratikskarnik/rIID-keras-transformer-starter>
10. <https://www.kaggle.com/erikbruin/rIID-comprehensive-eda-baseline>
11. <https://www.kaggle.com/isaienkov/rIID-answer-correctness-prediction-eda-modeling>
12. <https://www.kaggle.com/datafan07/rIID-challenge-eda-baseline-model>
13. <https://tech.preferred.jp/en/blog/lightgbm-tuner-new-optuna-integration-for-hyperparameter-optimization/>
14. <https://towardsdatascience.com/lightgbm-vs-xgboost-which-algorithm-win-the-race-1ff7dd4917d>