Student Performance Prediction

Varikuti Madhurima (210953314), Meda Bharath Reddy (210911308),
Likhith Balaji Reddy Thodimi (210911144), Nallamilli Naga Venkata Reddy (210953334)

SDG 4 – Quality Education

***Abstract*—This project addresses the challenge of predicting student performance by using advanced machine learning models to reveal patterns in academic data. At the core of this research is the support vector machine (SVM), tested with linear, polynomial, and Gaussian (RBF) kernels, each of which divides the optimized data into those of training, validation, and testing processes greater accuracy, we ensure that the models are reliable and fit well. In addition to SVMs, we use K-Nearest Neighbors (KNN) and Logistic Regression to find models that better capture the effect of different factors on student achievement. This work does not focus solely on numbers; It also highlights the key characteristics that most affect success, helping to make sense of what drives academic performance. The goal is to find the best predictive models and provide insights that can support targeted, data-driven decisions to help students succeed.**

***Index Terms*—Student Performance Prediction, K-Nearest Neighbors (KNN), Machine Learning in Education, Logistic Regression, Linear Kernel, Support Vector Machines (SVM), Polynomial Kernel, Gaussian (RBF) Kernel, Hyperparameter Tuning, Model Optimization, Accuracy, F1-Score, ROC-AUC, Evaluation Metrics, Feature Importance Analysis, Comparative Model Evaluation, Academic Success Factors**

I. Introduction

As schools and universities seek to better support students, predictive modeling has emerged as a powerful way to make sense of factors affecting academic success. By applying machine learning to student data, we can identify patterns in performance and truly understand what affects outcomes, enabling teachers to provide targeted intervention. This project takes a deep dive into the use of support vector machines (SVM) and other classification models to predict student success, aimed at helping educators make informed, data-driven decisions.

We examine three different SVM models—Linear, Polynomial, and Gaussian (RBF)—to see which performs best for this task. Each model is optimized to improve accuracy, and is analyzed based on metrics such as accuracy, F1-score, and ROC-AUC. This tuning helps ensure that each model captures meaningful patterns in the data. We go beyond the numbers by examining the importance of factors to identify key factors affecting student performance, bringing valuable context to predictions.

To get a broader perspective, we compare the performance of SVM with two other popular models: K-Nearest Neighbors (KNN) and Logistic Regression. We use visual tools such as ROC curves and comparative tables to show how each model stacks up and highlight the strengths of each method. By the end, this project aims to explore the most effective models for measuring student performance, providing insights that can inform practical, data-driven approaches to helping students succeed. Specifically, this research brings a human-centered approach to data science, blending research with real-world applications in education.

II. Literature Review

A. Students Performance Prediction Using KNN and Naïve Bayesian

The paper investigates the predictability of student performance in Gaza secondary schools using two machine learning methods The goal is to identify students who may need additional support by examining factors such as gender, parental employment status, and past grades. After processing the data and selecting key characteristics, the researchers applied both algorithms to a dataset of 500 student records. Naïve Bayes was found to have the highest accuracy 93.6 percentage, making it a reliable method for predicting early performance. This approach can help teachers and the Ministry of Education provide timely support, ultimately enhancing student educational outcomes[1].

B. An Ordinal Logistic Regression Model to Identify Factors Influencing Students Academic Performance at Njala University

The study looks at what affects student academic success at the University of Nzala. It showed that four main factors played a role: time spent studying. Students who read more and whose fathers earn more money do better. Interestingly, although higher educational attainment for mothers positively influences student performance, higher maternal income is associated with lower academic achievement, which may be due to social factors[2].

C. Classification and prediction of student performance data using various machine learning algorithms

The paper explores how machine learning can help shape student performance, enabling schools to better support students who may need additional support. After analyzing the students' data, the researchers tested different algorithms to see which ones were the best at predicting academic success. The findings showed that the SVM model was the most accurate in terms of performance prediction. Such a data-driven approach can help teachers focus on students who need it most, improve instructional methods, and potentially reduce dropout rates[3].

D. A Comparative Analysis of Student Performance in an Online vs. Face-to-Face Environmental Science Course From 2009 to 2016

This study examines how students performed in environmental science courses taught online compared to traditional, individual case study data from 548 students examined whether there were any meaningful and also looked at gender and grade differences systematic Interestingly, the study found no significant performance between online and in-person students, despite these internal factors. They don't know the difference. This suggests that, in non-STEM majors, online learning can effectively teach scientific concepts, providing flexibility that many students value without compromising instructional quality. Although the study faced some limitations, such as the convenience-based sample and the lack of control for prior online experience, the results support the concept of online education can be a strong and flexible alternative to the traditional classroom when done right[4].

III. Methodology

THIS section describes the methodology used to develop a model for analyzing student performance using data processing and machine learning techniques. The approach involves feature engineering, model analysis, model development, and Data preprocessing.

A. Dataset Collection

*1) **Dataset:*** This section describes the methodology used to develop a model for analyzing student performance using data processing and machine learning techniques. The approach involves feature engineering, model analysis, model development, and Data preprocessing.

B. Data Preprocessing and Cleaning

*1) **Data Loading and Initial Inspection**:* The dataset was loaded and checked for null values and duplicate records to ensure data quality. This step ensured that only valid and unique entries were used for model training.

*2) **Data Preprocessing:*** In order to prepare the student performance dataset for analysis, several preprocessing steps were taken to ensure data quality and consistency. These steps include handling missing values, encoding categorical features, and normalizing numerical data.

1) **Handling Missing Values**
   Missing data can impact model performance, so we addressed missing values through the following methods:
   - **Imputation**: We used techniques like mean or median imputation for numerical variables, depending on the data distribution. For categorical variables, we filled in missing values with the most frequent category.
   - **Removal of Records**: If records had too many missing fields that couldn't be reliably imputed, they were removed from the dataset to maintain data integrity.
2) **Encoding Categorical Variables**
   To make the dataset compatible with machine learning models, categorical variables were converted to numerical format:
   - **One-Hot Encoding**: This method was used for nominal features, such as gender and study program, generating binary columns for each category.
   - **Ordinal Encoding**: For features with an inherent order (e.g., education levels or grade categories), we assigned an ordinal scale, preserving the natural ordering.
3) **Normalizing/Standardizing Data**
   Features with varying scales, such as age, scores, or participation metrics, were normalized or standardized to ensure consistent model input:
   - **Normalization**: Applied to features that had a non-Gaussian distribution, rescaling them to a 0-1 range.
   - **Standardization**: For Gaussian-distributed data, we standardized features to have a mean of 0 and standard deviation of 1.
4) **Feature Selection**
   To reduce complexity, we analyzed feature correlations and selected the most relevant features based on their relationships with the target variable. Techniques such as correlation matrices and feature importance from initial models helped to identify key predictors of student performance.

C. Data Visualization

we explore various aspects of the student performance dataset through visualizations to gain insights and understand underlying patterns. The visualizations focus on analyzing distributions, relationships, and trends in the data, aiding in identifying key factors that may impact student performance

*1) **Feature Visualization**:* Firstly, we are going to look deeper into each feature by using multiple visualization methods, such as distribution plot and density. After the visualization, we will understand which features are most important for students' performances.

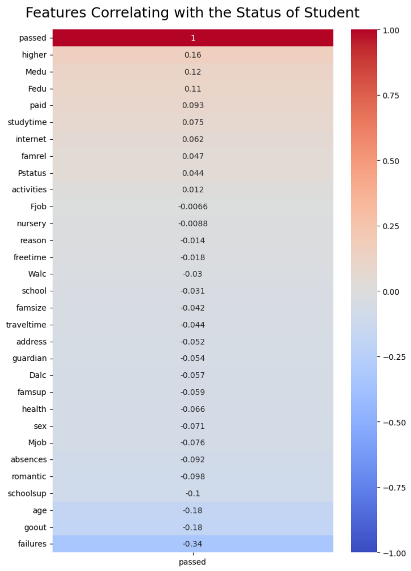


Fig. 1: heatmap

- ***correlation heat-map:*** This heat map visualization illustrates the correlation between the target variable, ”passed” (indicating student status), and other features in the dataset. Showing the strength and direction of each feature’s relationship with student success helps identify which factors may be most predictive or influential in determining whether a student passes. This insight guides feature selection and can enhance model performance by focusing on relevant attributes.

*2) **Distribution plot**:* After dealing with the most relevant features, the valedictorian of excellent conditions for high academic potential is likely to have this profile:

- Does not go out with friends frequently
- Is not in a romantic relation
- Parents receive higher education, especially woman
- Have a strong desire to receive higher education
- Mother is a health care professional
- The father is a teacher
- No absences to classes
- have access to the internet
- study more than 10 hours a week
- Is healthy

after the distribution plot in my project, I got this about student performance analysis

D. Model Development

*1) **Model Selection**:*

- **logistic Regression:** A reliable binary classifier used here to assess its effectiveness in predicting student success based on available resources.
- **Support Vector Machine (SVM):** Selected for its ability to create a clear margin between classes, potentially offering high accuracy in distinguishing performance levels.
- **K-Nearest Neighbors (KNN):** Chosen for its simplicity and effectiveness in identifying patterns by comparing a student’s performance to similar cases, providing an intuitive basis for predictions.

These models together allow for a comprehensive comparison of predictive performance in student success analysis.

E. Model Training

1) Data Splitting and Preprocessing:: TUse an 80/20 split of the data, with 80% for training and 20% for testing. for evaluation. **StratifiedKFold cross-validation** was employed to maintain balanced class distributions across folds. Each model was trained on the preprocessed data to predict student performance. This involved:

- **Logistic Regression:** Applied to predict the probability of student success/failure based on available features.
- **Support Vector Machine (SVM):** Used for its capability to handle complex boundaries between success and failure classes.
- **K-Nearest Neighbors (KNN):** Compared each student’s profile with similar cases to determine likely outcomes.

F. Model Evaluation and Hyperparameter Tuning

*1) **Evaluation Metrics**:* Model performance was evaluated using the following metrics.

- **Accuracy:** The proportion of accurate predictions made by a model.
- **Specificity:** The proportion of positives that are certain in all positive predictions.
- **Remember:** The real-positive fraction of all real-positive cases.
- **F1 Score:** A balanced metric that combines precision and recall into a single value, especially useful on imbalanced data.
- A **confusion matrix** was constructed for each model to visually evaluate performance for true positives, false positives, true negatives, and false negatives

2) Cross-Validation:: StratifiedKFold cross-validation was used to ensure the robustness of the model, dividing the data into several clusters to ensure that a balanced class was assigned to each fold is generalizes to subsets of data.

3) Hyperparameter Tuning with RandomizedSearchCV:: For further optimization, RandomizedSearchCV was used to fine-tune the hyperparameters of the model:

Logistic regression: We adjusted for parameters such as regularizing power (C) and punishment type.

IV. Results and Discussion

In this study, we examined key factors affecting student success and identified the most effective model for predicting academic achievement. We examined positive and negative effects on student achievement and compared the performance of three machine learning models: **Logistic Regression**, **K-Nearest Neighbors (KNN)**, and **Support Vector Machine (SVM)**.

A. Key Factors Affecting Student Success

Positive Impact Factors:

- **Father's Education**: Students with more educated fathers tended to perform better academically, possibly because educated parents provide guidance and support in schoolwork.
- **Parental Support**: Interestingly, there was no significant effect from parents alone; this may indicate additional support from extended family or counselors, but further research is needed to confirm this.
- **Desire for Higher Education**: Students who aspired to further study performed well. Clear learning objectives seem to encourage student focus.
- **Study Time**: Regular study time is essential for success. Generally, students who followed a structured curriculum performed better on tests.
- **Father's Occupation**: Stability in the father's job appeared to support academic success, possibly through access to educational resources like tutoring or internet access.

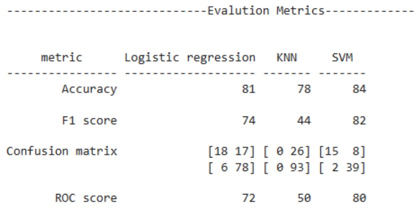
Negative Impact Factors:

- **Age**: Older students showed slight difficulties, potentially due to unique challenges or later high school entry.
- **Health**: Health challenges, though minor in impact, can affect school attendance or concentration.
- **Socializing**: Students who spent more time with friends outside school showed slightly lower academic performance. While socializing is important, excessive time spent can reduce study hours.
- **Absenteeism**: Absenteeism had a strong negative impact on performance, with regular attendance being essential to maintain academic performance.
- **Past Failures**: Students with histories of failure often continued to struggle, suggesting the need for early intervention to support these students.

B. Model Comparison and Selection

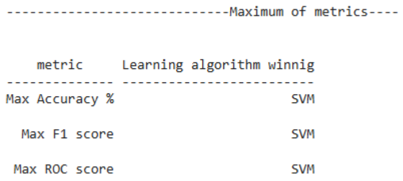
To determine the most accurate prediction model, we evaluated each model using metrics such as confusion matrix, accuracy, F1 score and ROC AUC score:

- **Logistic Regression**: This model achieved an **accuracy of 81%** and an **F1 score of 0.74**. Although slightly less accurate than SVM, its interpretability makes it useful for understanding factors that most affect success.
- **K-Nearest Neighbors (KNN)**: KNN attained an **accuracy of 78%** and an **F1 score of 0.44**. Its lower performance may stem from sensitivity to data imbalance or calibration issues.
- **Support Vector Machine (SVM)**: SVM was the best-performing model, with an **accuracy of 84%** and a maximum **F1 score of 0.82**. The confusion matrix indicated a strong ability to differentiate between passing and failing students, with minimal misclassification. Its **ROC AUC score of 0.80** further confirmed its predictive strength.
- **Linear Kernel SVM**: This kernel came out very well, achieving an **accuracy of 84%** and an **F1 score of 0.82**, with its confusion matrix indicating a strong ability to discriminate between classes accurately. A **ROC AUC score of 0.80** reconstructed its prediction, emphasizing the complexity of its performance.
- **Gaussian Kernel SVM**: Achieved an **accuracy of 82.8%** and an **F1 score of 0.77**, with a slightly lower performance measure compared to the linear kernel. Nevertheless, it remains a viable method with moderate classification accuracy.
- **Polynomial Kernel SVM**: Recorded an **accuracy of 78.1%** and an **F1 score of 0.74**. Combined with the lower ROC AUC results, these metrics indicate limitations in predictive power for this dataset.



```
----------------------------Evalution Metrics------------

     metric       Logistic regression   KNN     SVM
---------------- ------------------- ------- -------
        Accuracy                  81      78      84

        F1 score                  74      44      82

Confusion matrix             [18 17] [ 0 26] [15  8]
                             [ 6 78] [ 0 93] [ 2 39]

       ROC score                  72      50      80
```

Fig. 2: Evaluation Metrics



```
----------------------------Maximum of metrics----

    metric      Learning algorithm winnig
-------------- -------------------------
Max Accuracy %                       SVM

  Max F1 score                       SVM

 Max ROC score                       SVM
```

Fig. 3: Maximum of Metrics

C. Insights and Suggestions

Educational Implications: Findings such as parental education, study time, and attendance offer valuable insights for **teachers**, **school administrators**, and **parents**. Supporting effective study habits, encouraging further education, and ensuring consistent attendance can positively impact academic success.

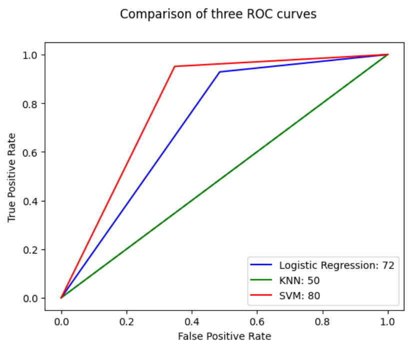


Fig. 4: Graph of ROC Curve

Recommended Model: SVM emerged as the most suitable model for this prediction task, owing to its high accuracy and balanced performance across all metrics. **Logistic Regression** remains valuable for understanding individual characteristic impacts.

Support and Policy Recommendations: Schools and communities should focus on **reducing absenteeism**, identifying students with repeated failures for early intervention, and engaging parents to provide academic guidance. Addressing the unique challenges faced by students from underserved areas can also help improve academic achievement.

V. Conclusion

In conclusion, this study highlights the effectiveness of the Support Vector Machine (SVM) model in predicting student success, outperforming logistic regression and K-nearest neighbor (KNN) in terms of accuracy and balance in key dimensions in, especially F1-score and ROC- AUC . Through comparative analysis, SVM models with Gaussian (RBF) kernels exhibited good prediction capabilities, making them a strong choice for educational data processing

The research also provides important insights into the factors affecting academic achievement, providing educators and policy makers with valuable frameworks for targeted interventions. These findings highlight the potential of machine learning to drive data-driven decisions that improve educational outcomes. Future research with expanded and diverse datasets could further strengthen these findings and explore additional predictors of student success, thus strengthening the foundation of data use in education and it contributes to the continuous improvement of educational data science.

References

[1] P. Kaur, M. Singh, and G. Singh, “Classification and prediction based data mining algorithms to predict slow learners in education sector,” *Procedia Comput. Sci.*, vol. 57, pp. 500–508, 2015.

[2] F. Haghanikhameneh, P. H. Shariat Panahy, N. Khanahmadliravi, and S. A. Mousavi, “A comparison study between data mining algorithms over classification techniques in Squid dataset,” *Int. J. Artif. Intell.*, vol. 9, no. 12 A, pp. 59–66, 2012.

[3] K. M. Osei Bryson, “Towards Supporting Expert Evaluation of Clustering Results Using a Data Mining Process Model,” *Inf. Sci.*, vol. 180, no. 3, pp. 414-431, 2010.

[4] S. Kovalev, A. Kolodenkova, and E. Muntyan, “Educational Data Mining: Current Problems and Solutions,” in *V International Conference on Information Technologies in Engineering Education (Inforino)*, 2020, pp. 1-5, doi: 10.1109/Inforino48376.2020.9111699.

[5] S. R. Hamidi, Z. A. Shaffiei, S. M. Sarif, and N. Ashar, “Exploratory study of assessment in teaching and learning,” in *International Conference on Research and Innovation in Information Systems (ICRIIS)*, 2013, pp. 398–403.

[6] D. Xu and S. S. Jaggars, “Performance gaps between online and face-to-face courses: differences across types of students and academic subject areas,” *J. High. Educ.*, vol. 85, no. 5, pp. 633–659, 2016. doi: 10.1353/jhe.2014.0028.

[7] L.-C. Zhang and A. C. Worthington, “Scale and scope economies of distance education in Australian universities,” *Stud. High. Educ.*, vol. 42, pp. 1785–1799, 2017. doi: 10.1080/03075079.2015.1126817.

[8] G. Considine and G. Zappala, “Influence of social and economic disadvantage on the academic performance of school students in Australia,” *J. Sociology*, vol. 38, pp. 129–148, 2002.

[9] D. E. Ukpong and I. N. George, “Length of study-time behavior and academic achievement of social studies education students in the University of Uyo,” *Int. Educ. Stud.*, vol. 6, no. 3, pp. 113–120, 2013.

Team Member	Contribution Details
Varikuti Madhurima	**Introduction**: Explained the purpose of predicting student success for the benefit of teachers and policymakers. **Literature Review**: Researched ”Classification and Prediction of Student Performance Using Various Machine Learning Algorithms.” **EDA**: Created pair plots to visualize relationships between features. **Visualization**: Used pair plots to illustrate key factor relationships. **Modeling (Polynomial and Gaussian SVM)**: Developed and optimized SVM models with Polynomial and Gaussian kernels. **Results and Discussion**: Compared polynomial and Gaussian kernels, discussing improvements for student success prediction. **Conclusion**: Summarized conclusions on the comparative performance of different SVM models.
Meda Bharat Reddy	**Introduction**: Explained classification problems and scaling needs in machine learning. **Literature Review**: Researched ”Predicting Student Performance Using KNN and Naive Bayesian.” **Data Processing**: Enhanced non-binary features to improve KNN model performance. **Visualization**: Utilized boxplots to make predictions based on factors like study time. **Modeling (K-Nearest Neighbors)**: Developed and evaluated a KNN algorithm for student performance prediction. **Results and Discussion**: Observed impact of characteristics, like study time, on student achievement. **Conclusion**: Highlighted strengths and limitations of the KNN algorithm.
Likhith Balaji Reddy Thodimi	**Data Operations**: Conducted basic data analysis using functions like `head()` and `describe()`. **Literature Review**: Researched ”Ordinary Logistic Regression Model to Identify Factors Affecting Student Academic Achievement.” **EDA**: Analyzed interactions using a correlation matrix and a heatmap. **Modeling (Linear SVM)**: Developed and optimized the Linear Support Vector Machine (SVM) algorithm, with tuning of the C parameters **Results and Discussion**: Evaluated the performance of Linear SVM using accuracy and F1-score metrics. **Conclusion**: Summarized the main findings on the efficacy of Linear SVM.
Nallamilli Naga Venkata Reddy	**Data Creation**: Configured data sources and structured the dataset. **Literature Review**: Summarized research from ”Comparative assessment of student performance in online and face-to-face courses.” **Data Manipulation**: Converted categorical data (e.g., parental background) to statistical values for modeling. **Visualization**: Created histograms for factors like internet usage and study time. **Modeling (Logistic Regression)**: Used Logistic Regression to predict student performance. **Results and Discussion**: Identified patterns, such as the impact of internet usage on academic success. **Conclusion**: Provided a summary, emphasizing the effectiveness of Logistic Regression.

TABLE I: Team Member Contributions