1. **What are the key academic and behavioural factors that most strongly influence student performance according to your models?**
   * **Answer**: Factors such as academic history (grades and entrance scores) and behavioural data (attendance, participation in assignments) showed the strongest correlation with student performance. These features helped differentiate successful students from those at risk, supporting early interventions.
2. **How do your findings contribute to the existing body of literature on predictive models in education?**
   * **Answer**: This study advances existing literature by comparing a range of machine learning models in the context of student performance prediction. While previous work has focused on specific models or datasets, this study presents a comprehensive comparison of both traditional and ensemble-based methods, offering deeper insights into model performance with real-world educational data.
3. **What are the limitations of the models used, and how can they be improved in future research?**
   * **Answer**: The limitations mainly stem from the inability of some models, particularly **ANN** and **Bernoulli Naive Bayes**, to handle feature dependencies effectively. In future work, incorporating more sophisticated data preprocessing techniques or using models that can better capture feature interactions, such as deep learning models or hybrid ensemble methods, could yield improvements.
4. **Why was feature selection crucial in your model, and how did it affect performance?**
   * **Answer**: Feature selection was vital to reduce overfitting, improve model interpretability, and enhance prediction accuracy. By focusing on the most significant features—such as student grades, attendance, and participation—unnecessary noise was minimized, leading to better performance and clearer decision boundaries.
5. **What practical impact could your findings have on educational institutions and policymakers?**
   * **Answer**: This study offers actionable insights that educational institutions can use to identify at-risk students early and provide tailored interventions. Policymakers can use the findings to allocate resources more effectively, ensuring that students in need receive the support they require to succeed.
6. **How do ensemble models like Random Forest and XGBoost outperform individual models in this context?**
   * **Answer**: Ensemble models combine the strengths of multiple base models, which helps reduce the model’s variance and bias. **Random Forest** and **XGBoost** effectively handle feature interactions and imbalances, providing more robust predictions, particularly for complex datasets with subtle patterns.
7. **How do you envision your model being implemented in real-world educational settings?**
   * **Answer**: The model could be integrated into learning management systems (LMS) to track student performance in real-time. It could flag students at risk of underperforming, enabling instructors and counsellors to provide timely support, such as tutoring or mentoring programs, thus improving student retention rates and overall success.
8. **Could your model be adapted to predict performance in other domains beyond education?**
   * **Answer**: Yes, the underlying machine learning models are highly adaptable. With minor modifications, the same approach could be applied to areas like employee performance, patient health outcomes, or customer retention, where predictive analytics could play a crucial role in decision-making and resource allocation.
9. **How did the model’s performance differ between training and testing datasets, and what does this indicate about generalization?**
   * **Answer**: The models exhibited good generalization, with similar performance on both training and testing datasets, indicating that overfitting was minimized. However, some models showed slight discrepancies between training and testing, suggesting that additional tuning or feature engineering could improve generalization further.
10. **What role did hyperparameter tuning play in improving model performance?**
    * Answer: Hyperparameter tuning played a critical role in enhancing model performance, especially for complex models like ANN and SVM. Fine-tuning parameters such as learning rates, tree depth (for Random Forest), and kernel choices (for SVM) significantly improved accuracy and minimized error, leading to better model robustness and predictive
11. What does it reflect, the negative or week correlation of my all-other features (x) with respect to target feature (y)?
    * Answer: The week or negative correlation reflect that the target variable is not depended on other variables. In this journal, I used feature engineering technique to create a new feature i.e. class which is a target variable, this variable is obtained by classifying the average into 2 classes.
12. Why there is 280 values in confusion matrix and sometime 279 in confusion matrix?
    * Answer: So, the value 280 reflects the number of datasets used by K-Fold Cross-Validation process. The total dataset is of 1397 rows, and I used 5 K-Fold Cross-Validation, which means 1397/5 = 280(approx.). Each K-fold used 280 dataset value to process the cross-validation. Sometimes, due to model internal process this value becomes 279, still it is approximately value. This does not cause any issue at all, because the change in value is 1 only but if this value is higher like 10 then it will affect the model performance and result as well.
13. Why Naive Bayes have lowest accuracy as compared to other models? Answered in NOTE.
    * Answer: So, Naïve bayes