**Project Documentation: Student Graduation Prediction System**

**Created By**: Yakub Junaid  
**Project Title**: Predicting Student Graduation Status Using Machine Learning

**1. Project Objectives**

The primary objective of this project was to develop a machine learning model capable of predicting whether a student will graduate based on various academic metrics. The project aimed to:

* **Objective 1**: Build a data-driven solution to help educational institutions identify at-risk students early on and take proactive measures to improve their chances of success.
* **Objective 2**: Develop an easy-to-use, web-based interface for making predictions based on student data.
* **Objective 3**: Select the most important academic features that impact graduation rates to improve model performance and interpretability.

**2. Achievements**

The project successfully met its objectives with the following outcomes:

* A Random Forest Classifier model was developed and fine-tuned to predict student graduation with high accuracy.
* Feature selection was conducted to identify the 10 most important features influencing student graduation, resulting in a more efficient and interpretable model.
* The model was integrated into a user-friendly web application using Streamlit, allowing real-time predictions based on user input.
* The app was successfully tested and deployed, providing educational institutions with a tool to predict graduation outcomes based on student data.

**3. Challenges Faced and Solutions**

**Data Imbalance**

* **Description**: The initial dataset was imbalanced, with a disproportionate number of students who graduated versus those who did not. This imbalance could have led to biased model performance.
* **Solution**: Applied techniques such as under sampling the majority class and over-sampling the minority class to create a balanced dataset. The Random Forest algorithm also helped mitigate bias by building multiple trees, some focused on under-represented samples.

**Feature Selection**

* **Description**: The model initially included many features, which made it harder to interpret and could result in overfitting.
* **Solution**: Implemented feature selection based on feature importance scores from the Random Forest model. The 10 most important features were selected, resulting in a simpler and more interpretable model without significant loss of accuracy.

**Data Scaling**

* **Description**: Without scaling, the model performed poorly as some features had different ranges, which negatively impacted the training process.
* **Solution**: Used a StandardScaler to normalize the data, ensuring that all features were on the same scale. This improved model performance significantly.

**4. Future Works**

To enhance the current system, the following future work is planned:

1. **Model Improvement via Hyperparameter Tuning**
   * Optimize the Random Forest model by tuning hyperparameters such as the number of trees, maximum depth, and minimum samples split to further improve model accuracy and generalization.
2. **Incorporation of Additional Features**
   * Future iterations of the model could include more features related to student behavior, such as attendance records, extracurricular activities, and financial aid data, to improve predictive accuracy.
3. **Expansion to Other Educational Institutions**
   * Currently, the model is based on data from a specific institution. Future work could involve collecting data from multiple institutions to develop a more generalizable model.
4. **Improved Visualization and Reporting**
   * Add visualization dashboards to provide more detailed insights into predictions, such as heatmaps showing feature importance or trends in student dropout rates over time.
5. **Deployment on Cloud Platforms**
   * Consider migrating the app to a more robust cloud platform like AWS or Google Cloud for better performance, scalability, and integration with other educational tools.
6. **Early Intervention System**
   * Develop a feature to alert administrators and counselors when a student is predicted to drop out, enabling them to intervene and provide targeted support.

**5. Conclusion**

The project successfully created a machine learning model to predict student graduation with high accuracy. The model's deployment via a user-friendly web app allows institutions to use data for early interventions and improve graduation rates. Despite challenges like data imbalance and scaling issues, the solutions implemented resulted in a robust and interpretable model. Future improvements will focus on expanding the model's applicability and enhancing user experience.