cat 2 data science
student prediction model

1. **Which type of machine learning algorithm would be most suitable for this task? Explain your reasoning.**

Answer:

A **:Random Forest Classifier**---is most suitable for this task. Here's why:

- Handles Mixed Data Types: Random Forest can handle both numerical and categorical features without requiring extensive preprocessing.

- **Captures Non-linear Relationships:** It can capture complex interactions between features, which is important for predicting enrollment and graduation success.

- Feature Importance: Provides feature importance scores, helping identify key predictors.

- Robust to Overfitting-- Less prone to overfitting compared to single decision trees, making it reliable for smaller datasets.

---

2. What features from the student data would be most relevant for predicting enrollment and graduation success?

Answer:

The most relevant features include:

- Historical Student Enrollment Data

  - Previous enrollment status (Previous qualification, Curricular units 1st sem (approved)`, etc.).

  - Academic performance metrics (`Admission grade, Curricular units 1st sem (grade)`).

- Student Academic Records:

cat 2 data science
student prediction model

- GPA, test scores, attendance rates, study hours, course completion rates.
- Performance indicators like grades and evaluations.

- **Student Demographic Data**
- Marital status, gender, age at enrollment, international status.
- Financial aid status (`Tuition fees up to date`, `Scholarship holder`).

3. How can you protect the privacy of student data while still using it to develop predictive models?

Answer:

To protect student privacy:

1. **Anonymize Identifiers:** Remove or hash personally identifiable information (PII) like names, IDs, and addresses.

2. **Add Noise:** Introduce small random noise to sensitive numerical features (e.g., GPA, test scores) to obscure exact values.

3**. Aggregate Data**: Use aggregated statistics instead of individual-level data where possible.

4. **Access Controls**: Implement strict access controls to ensure only authorized personnel can view raw data.

5. **De-identification**: Replace sensitive attributes with generic categories (e.g., age ranges instead of exact ages).

---

4. **How can you communicate the results of your model to educational institutions in a way that is actionable and informative?**

cat 2 data science
student prediction model

Answer:

To communicate results effectively:

1**. Detailed Reports**: Provide comprehensive reports summarizing model predictions and insights.

2. **Visualizations:** Use charts and graphs (e.g., bar plots, heatmaps) to highlight key findings.

3. **Actionable Insights**: Identify high-risk students and recommend targeted interventions.

4. **Interactive Dashboards**: Develop dashboards allowing institutions to explore predictions by program, demographics, or other factors.

5**. Recommendations**: Offer specific recommendations based on feature importance (e.g., focus on improving attendance rates).