Detailed Report for Grade 10th : Predicting College Major Choices

1. Overview

This report describes the Python code used for predicting college major choices based on educational survey data from 10th-grade students. The code employs machine learning techniques to analyze and model student preferences and academic performance.

2. Data Overview

2.1. Variables and Features

Fall Semester Variables:

- GAMTH1C, GAMTH1D, GAMTH1E, GAMTH1H
- GAMTH2C
- GASCI1D, GASCI1E, GASCI1H, GASCI2C, GASCI2D
- GAENG1C, GAENG1D, GAENG1E, GAENG2C, GAENG2D
- GASSTC, GASSTD
- GACOMC, GACOMD, GAFORC
- GAARTC, GAMUSC, GAVOCC
- GA32A, GA32D, GA33A, GA33D

Spring Semester Variables:

- HAMTH1C, HAMTH1D, HAMTH2C
- HASCI1C, HASCI2C
- HAENG1C, HAENG2C
- HASSTC, HACOMC
- HAFORC, HAARTC, HAMUSC, HAVOCC

Target Variable:

- LAMAJOR8I: Encoded as LAMAJOR8I_encoded

3. Data Variables

3.1 Subject Preference and Teacher Effectiveness

1. **Math Classes:**

- **GAMTH1C:** Liking for the first math class (fall)
- **GAMTH1D:** Teacher clarity in the first math class (fall)
- **GAMTH1E:** Challenge level in the first math class (fall)
- **GAMTH1H:** Difficulty level in the first math class (fall)
- **GAMTH2C:** Liking for the second math class (fall)
- **HAMTH1C:** Liking for the first math class (spring)
- **HAMTH2C:** Liking for the second math class (spring)

2. **Science Classes:**
   - **GASCI1D:** Teacher clarity in the first science class (fall)
   - **GASCI1E:** Challenge level in the first science class (fall)
   - **GASCI1H:** Difficulty level in the first science class (fall)
   - **GASCI2C:** Liking for the second science class (fall)
   - **GASCI2D:** Teacher clarity in the second science class (fall)
   - **HASCI1C:** Liking for the first science class (spring)
   - **HASCI2C:** Liking for the second science class (spring)
3. **English Classes:**
   - **GAENG1C:** Liking for the first English class (fall)
   - **GAENG1D:** Teacher clarity in the first English class (fall)
   - **GAENG1E:** Challenge level in the first English class (fall)
   - **GAENG2C:** Liking for the second English class (fall)
   - **GAENG2D:** Teacher clarity in the second English class (fall)
   - **HAENG1C:** Liking for the first English class (spring)
   - **HAENG2C:** Liking for the second English class (spring)
4. **Social Studies:**
   - **GASSTC:** Liking for social studies class (fall)
   - **GASSTD:** Teacher clarity in social studies class (fall)
   - **HASSTC:** Liking for social studies class (spring)
5. **Other Subjects:**
   - **GACOMC:** Liking for computer class (fall)
   - **GACOMD:** Teacher clarity in computer class (fall)
   - **GAFORC:** Liking for foreign language class (fall)
   - **GAARTC:** Liking for art class (fall)
   - **GAMUSC:** Liking for music/dance class (fall)
   - **GAVOCC:** Liking for business/vocational class (fall)
   - **HACOMC:** Liking for computer class (spring)
   - **HAFORC:** Liking for foreign language class (spring)
   - **HAARTC:** Liking for art class (spring)
   - **HAMUSC:** Liking for music/dance class (spring)
   - **HAVOCC:** Liking for business/vocational class (spring)

3.2 Interests and Career Aspirations

1. **Interest in Issues:**
   - **GA8A:** Interest in foreign policy issues
   - **GA8B:** Interest in space exploration

- **GA8C:** Interest in agricultural issues
- **GA8D:** Interest in science issues
- **GA8E:** Interest in economic issues
- **GA8F:** Interest in minority rights
- **GA8G:** Interest in new technologies
- **GA8I:** Interest in women's rights issues

2. **Occupational Aspirations:**
   - **GA20ATXT:** First choice occupation (fall)
   - **GA20C:** Certainty of first choice occupation (fall)
   - **HA9ATXT:** First choice occupation (spring)
   - **GA32A:** Enjoyment of math
   - **GA33A:** Enjoyment of science
   - **GA32D:** Perception that math is more useful for boys
   - **GA33D:** Perception that science is more useful for boys

3.3 Importance and Values

1. **Values and Aspirations:**
   - **GA6A:** Importance of success in work
   - **GA6C:** Importance of having lots of money
   - **GA2B:** Intention to enroll in a four-year college or university
   - **GA1V:** Discussion about future with friends during the summer
2. **Teacher Knowledge:**
   - **HK15:** Teacher knowledge in subjects

---

4. Analysis

1. **Subject Liking and Teacher Effectiveness:**
   - Students' preferences and perceptions of subject difficulty and teacher clarity are recorded for fall and spring semesters. Analysis of these factors helps determine which subjects and teachers are most effective and engaging for students.
2. **Interest in Issues:**
   - Students' interest in various issues like space exploration, economic issues, and scientific discoveries is tracked. This data helps understand the alignment of student interests with current global and societal topics.
3. **Career Aspirations:**
   - Students' first-choice occupations and their certainty about these choices are analyzed to gauge their career planning and aspirations. Trends in occupational preferences provide insights into future career paths students are considering.
4. **Values and Aspirations:**
   - Importance placed on success, money, and higher education is assessed. This analysis helps in understanding students' long-term goals and motivations.
5. **Teacher Knowledge:**

- Teacher knowledge ratings provide feedback on how well students perceive their teachers' expertise in their subjects, which can influence overall student satisfaction and learning outcomes.

5. Process Overview

5.1. Data Loading and Preprocessing

1. **Data Loading:**
   - The dataset is loaded from a CSV file (`30263-0001-Data.csv`), which includes features and the target variable.
2. **Feature Extraction:**
   - Variables are divided into those representing the fall and spring semesters, based on their relevance and timing.
3. **Data Encoding:**
   - The target variable `LAMAJOR8I` is encoded into numerical values for model compatibility.
4. **Handling Data Warnings:**
   - Warnings related to setting with copy are corrected by creating explicit copies of the DataFrame for fall and spring data.

5.2. Data Splitting and Standardization

1. **Data Splitting:**
   - The dataset is split into training and test sets for both fall and spring semesters using an 80-20 split.
2. **Feature Scaling:**
   - Features are standardized using `StandardScaler` to ensure uniform scaling and improve model performance.

5.3. Model Training and Evaluation

1. **Model Training:**
   - Random Forest Classifier models are trained separately for fall and spring semester data.
2. **Model Evaluation:**
   - Models are evaluated using confusion matrices, classification reports, and ROC curves.
   - Feature importance is visualized to understand the significance of each feature in predictions.

5.4. Visualization and Analysis

1. **Feature Importance:**

    - Bar plots are used to show the importance of different features in the Random Forest model for both fall and spring semesters.
2. **Confusion Matrix:**
    - Confusion matrices are plotted to visualize the performance of the models in classifying college major choices.
3. **ROC Curves:**
    - ROC curves are plotted for each class to assess the performance of the models across different classes.
4. **Distribution and Likability Analysis:**
    - Count plots and box plots are used to analyze the distribution of college major choices and subject likability.

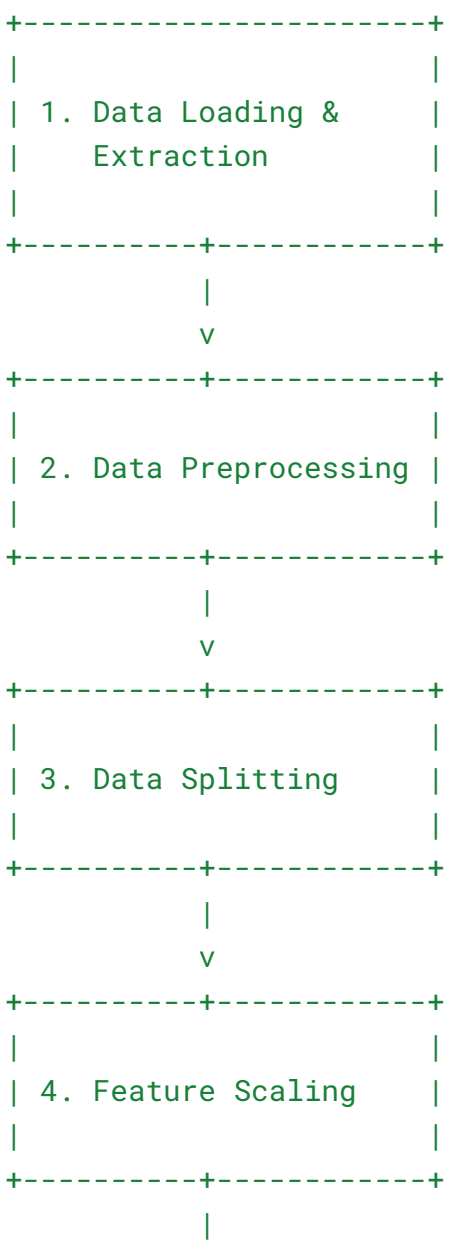
Overall Architecture

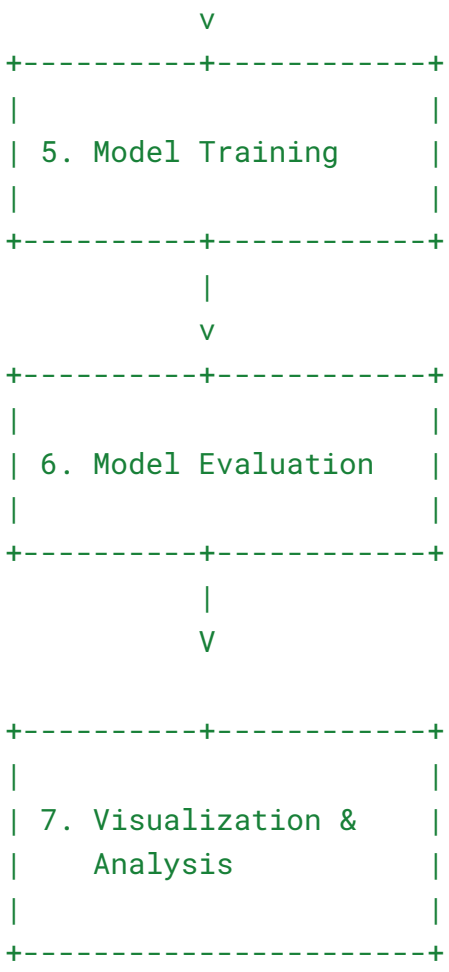
Data Processing and Model Workflow

1. **Data Loading & Extraction:**
    - **Description:** Load the dataset from a CSV file and extract relevant features for the fall and spring semesters.
    - **Output:** DataFrames for fall and spring semester features.
2. **Data Preprocessing:**
    - **Description:**
        - Encode categorical variables (e.g., `LAMAJOR8I`) into numerical values.
        - Handle any data warnings and prepare datasets for modeling.
    - **Output:** Preprocessed DataFrames.
3. **Data Splitting:**
    - **Description:** Split the preprocessed data into training and test sets.
    - **Output:** Training and test DataFrames for both fall and spring semesters.
4. **Feature Scaling:**
    - **Description:** Standardize features to ensure uniform scaling.
    - **Output:** Scaled features for training and test datasets.
5. **Model Training:**
    - **Description:** Train Random Forest classifiers on the training datasets.
    - **Output:** Trained Random Forest models for fall and spring semesters.
6. **Model Evaluation:**
    - **Description:**
        - Evaluate models using metrics such as confusion matrices, classification reports, and ROC curves.
        - Visualize feature importances.
    - **Output:** Evaluation reports and visualizations.
7. **Visualization & Analysis:**

- **Description:**
  - Generate plots for feature importance.
  - Plot confusion matrices and ROC curves.
  - Analyze distribution and likability of subjects.
- **Output:** Visual reports and insights.

Architecture Diagram



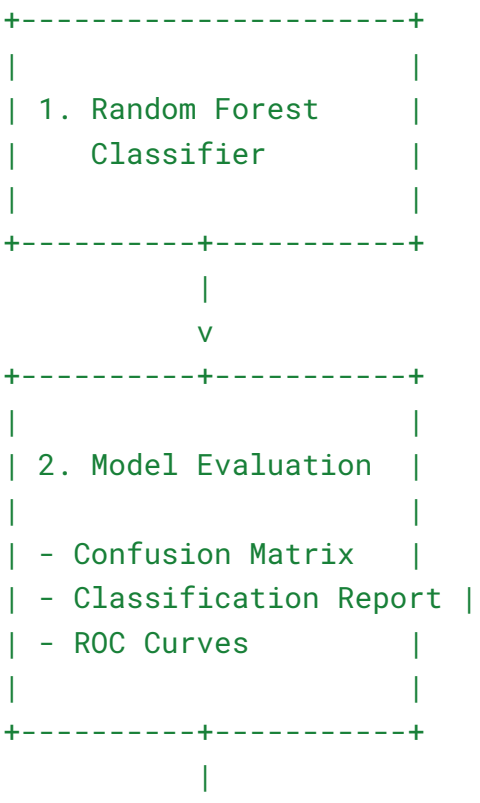

5. Model Training
6. Model Evaluation
7. Visualization & Analysis

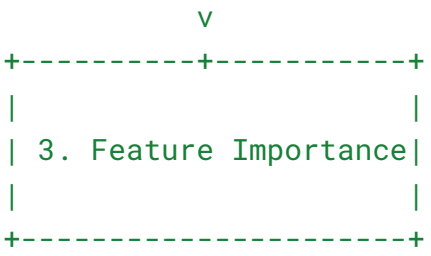
Model Architecture

Random Forest Classifier

1. **Model Training:**
    - **Description:** Train a Random Forest classifier using the training dataset.
    - **Features:** Use both fall and spring semester features.
2. **Model Evaluation:**
    - **Description:** Evaluate the trained model on the test dataset to assess performance.
    - **Metrics:**
        - **Confusion Matrix:** Shows classification performance across different classes.
        - **Classification Report:** Provides precision, recall, and F1-score for each class.
        - **ROC Curve:** Displays the trade-off between true positive rate and false positive rate.
3. **Feature Importance:**
    - **Description:** Analyze and visualize the importance of each feature in the model's predictions.

Model Diagram





These diagrams provide a clearer representation of the overall workflow and model architecture used in the analysis. The steps are sequential and interrelated, ensuring a comprehensive approach to predicting college major choices.

Recommendations

1. **Enhance Teacher Training:**
    - Based on feedback about teacher clarity and effectiveness, additional training and resources could be provided to teachers to improve clarity and engagement.
2. **Support Subject Preferences:**
    - Offer additional resources or support in subjects where students show lower enjoyment or find more challenging.
3. **Career Counseling:**
    - Increase career counseling services to help students better understand and pursue their career interests and aspirations.
4. **Align Interests with Curriculum:**
    - Incorporate topics that align with students' interests into the curriculum to increase engagement and relevance.
5. **Address Gender Perceptions:**
    - Implement programs to address and challenge gender perceptions related to subjects like math and science to promote equity.