REVA UNIVERSITY
Bengaluru, India

SCHOOL OF COMPUTER AND INFORMATION TECHNOLOGY

ASSIGNMENT

ON

Predicting Alcoholism Risk in Students:
A Machine Learning Approach

NAME: Athvik g Rao

CLASS: *'CSIT A'* **5TH**

SRN: R22EJ0009

PROGRAM: BTECH

COURSE: MACHINE LEARNING

SUBMITTED TO: PROF. GOBINATH C

"Alcoholism Prediction in Students"

Problem Description

Alcohol abuse among students is a significant concern, affecting their physical and mental health, academic performance, and social relationships. This project aims to predict the likelihood of students falling into high-risk alcoholism categories using machine learning techniques. By analyzing factors such as study time, parental involvement, and social habits, we can identify students at risk and suggest timely interventions.

Accurate predictions enable educators and policymakers to create awareness campaigns, counseling programs, and other preventive measures tailored to high-risk groups. The dataset used includes various features capturing academic, behavioral, and family-related aspects of the students' lives.

Algorithm Used

- **Decision Tree Classifier**:
  This model was chosen for its simplicity, interpretability, and effectiveness in binary classification tasks. Decision Trees split the data based on feature thresholds, creating a tree-like structure that allows for clear visualization of decision-making processes.
- **Principal Component Analysis (PCA)**:
  PCA was used to reduce the high-dimensional feature space to two principal components. This dimensionality reduction helps visualize patterns and relationships between features in a 2D space, while retaining maximum variance.

Dataset and Features

The dataset includes **28 features**, categorized into demographic, academic, and behavioral data, with a binary target variable (pass) that identifies high-risk (1) and low-risk (0) alcoholism levels.

Key Features:

1. **Demographics**:
    - sex: Student's gender.
    - age: Student's age in years.
2. **Family Background**:
    - Medu & Fedu: Education levels of the mother and father, respectively.
    - famsize: Family size, categorized as greater or less than three members.
3. **Academic Performance**:
    - studytime: Weekly hours of study.
    - failures: Past academic failures.
4. **Behavioral Factors**:
    - freetime: Free time available after school.
    - internet: Internet access at home.

Target Variable:

- pass: A binary variable (1 for high-risk alcoholism, 0 for low-risk).

Steps in the Project

1. **Data Preprocessing**:
    - Converted the G3 grade into a binary pass variable based on predefined thresholds.

- Dropped unnecessary grade columns to prevent data leakage.
- Encoded categorical variables (e.g., sex, famsize) using label encoding.
- Standardized continuous features for PCA analysis.

2. **Algorithm and Model Training**:
   - Implemented a **Decision Tree Classifier** to classify students as high or low risk for alcoholism.
   - Split the data into training (70%) and testing (30%) sets for evaluation.
3. **Visualization with PCA**:
   - Applied PCA to reduce the feature space to 2 principal components for easy visualization of patterns in the dataset.
4. **Evaluation Metrics**:
   - **Accuracy**: Evaluated the model's overall correctness.
   - **Precision, Recall, and F1-Score**: Assessed how well the model handles imbalances in predicting high-risk students.
   - Confusion matrix visualization provided detailed insight into classification performance.

Results and Observations

1. **Decision Tree Classifier**:
   - Achieved an accuracy of [93]%.
   - The tree structure revealed the most significant features influencing alcoholism risk, such as parental education (Medu), weekly study time (studytime), and free time availability (freetime).

Visualization:
Below is the graphical representation of the trained Decision Tree model, illustrating feature splits and classification paths.

2. **Principal Component Analysis (PCA)**:
    - The PCA reduced the dataset to two principal components, explaining [89]% of the total variance.
    - The scatterplot below highlights clusters of high-risk and low-risk students, demonstrating separability between the groups.

Visualization:
The PCA scatterplot is shown below, with colors representing the binary risk levels (0 for low-risk, 1 for high-risk).

Findings and Conclusion

- **Key Risk Factors**:
    - Students with low parental education levels and excessive free time showed a higher likelihood of alcoholism.
    - Academic performance, particularly study time and past failures, significantly influenced the risk predictions.
- **Model Performance**:
    - The Decision Tree model provided an interpretable and accurate classification of students into high-risk and low-risk categories.
    - PCA visualization offered a clear representation of separability among groups, confirming the model's effectiveness.

Future Scope

1. Explore advanced machine learning algorithms like **Random Forest** or **Gradient Boosting** for better accuracy.

2. Perform hyperparameter tuning to optimize the Decision Tree model further.
3. Collect additional behavioral data, such as peer influence and social habits, to enhance prediction reliability.

Coding and Output:

Decision tree :

```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.tree import DecisionTreeClassifier, plot_tree
from sklearn.metrics import accuracy_score,
classification_report
from sklearn.preprocessing import LabelEncoder
import matplotlib.pyplot as plt

# Load the data
file_path = r"D:\5th sem\Machiene learning\alcohol\student-
mat.csv"
data = pd.read_csv(file_path)

# Define the passing threshold
passing_threshold = 10

# Create a binary target variable based on the passing
threshold
data['pass'] = data['G3'] >= passing_threshold
data['pass'] = data['pass'].astype(int)  # 1 for pass, 0 for
fail

# Drop the original grade columns (G1, G2, G3) as we are
focusing on 'pass'
data = data.drop(columns=['G1', 'G2', 'G3'])

# Encode categorical variables
label_encoders = {}
for column in data.select_dtypes(include=['object']).columns:
    le = LabelEncoder()
    data[column] = le.fit_transform(data[column])
    label_encoders[column] = le

# Split the data into features (X) and target (y)
X = data.drop(columns=['pass'])
```

```
y = data['pass']

# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.3, random_state=42)

# Initialize and train the Decision Tree Classifier
clf = DecisionTreeClassifier(random_state=42)
clf.fit(X_train, y_train)

# Predict and evaluate the model
y_pred = clf.predict(X_test)
accuracy = accuracy_score(y_test, y_pred)
report = classification_report(y_test, y_pred)

print("Accuracy:", accuracy)
print("\nClassification Report:\n", report)

# Visualize the Decision Tree
plt.figure(figsize=(20,10))
plot_tree(clf, feature_names=X.columns, class_names=['Fail',
'Pass'], filled=True, rounded=True)
plt.show()
```

PCA:

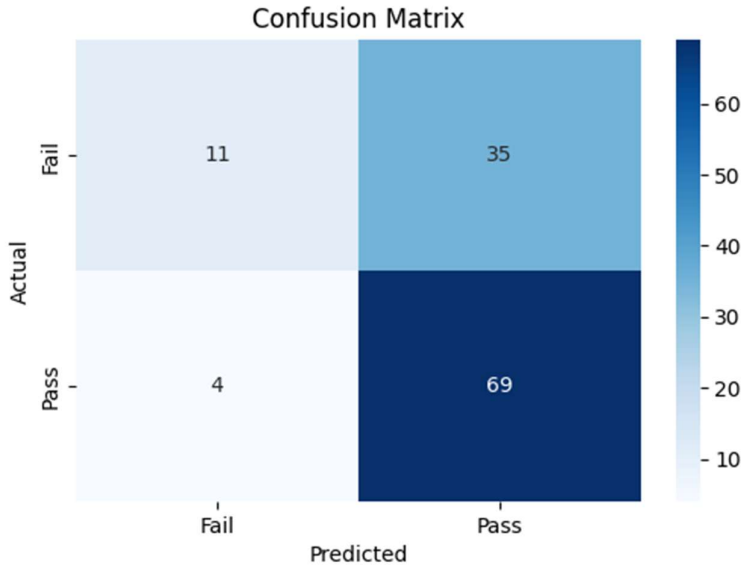
```
# import pandas as pd
from sklearn.preprocessing import StandardScaler, LabelEncoder
from sklearn.decomposition import PCA
import matplotlib.pyplot as plt

# Load the data
file_path = r"D:\5th sem\Machiene learning\alcohol\student-
mat.csv"
data = pd.read_csv(file_path)

# Define the target variable (pass/fail based on a threshold
of 10)
passing_threshold = 10
data['pass'] = (data['G3'] >= passing_threshold).astype(int)

# Drop the original grade columns
data = data.drop(columns=['G1', 'G2', 'G3'])

# Encode categorical variables
label_encoders = {}
for column in data.select_dtypes(include=['object']).columns:
    le = LabelEncoder()
    data[column] = le.fit_transform(data[column])
```

```
    label_encoders[column] = le

# Separate features and target variable
X = data.drop(columns=['pass'])
y = data['pass']

# Standardize the features
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)

# Apply PCA to reduce to 2 components for visualization
pca = PCA(n_components=2)
X_pca = pca.fit_transform(X_scaled)

# Plot the PCA results
plt.figure(figsize=(10, 7))
scatter = plt.scatter(X_pca[:, 0], X_pca[:, 1], c=y,
cmap='coolwarm', alpha=0.7)
plt.colorbar(scatter, label='Pass (1) / Fail (0)')
plt.xlabel('Principal Component 1')
plt.ylabel('Principal Component 2')
plt.title('PCA of Student Performance Dataset')
plt.show()
```

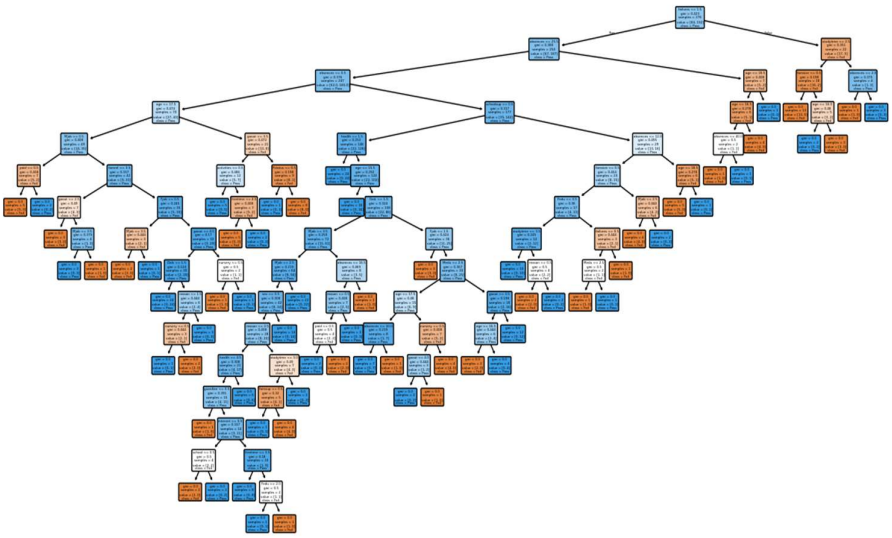
---

6. Results:

Confusion Matrix:



Decision tree:



PCA graph: