Predicting Youth Drug Usage

Abstract:

This research is based on analyzing the National Survey on Drug Usage and Health data using decision tree models. The given survey data has various factors which play important role in predicting youth drug usage. In this study we explored three types of problems one being binary classification for Marijuana usage, multi-class classification for number days marijuana used in past year and regression for number of days school missed with multiple models such as random forest, bagging, boosting. Based on the importance variables and plots we have understood the usage of marijuana by youth, next we have understood how the usage of marijuana in the last year impacted by the demographic and substance factors. Additionally , analysed how usage marijuana effected the youth from missing the school.

Introduction:

In this study, we aimed to analyze and predict marijuana consumption by youth using the National Survey on Drug Use and Health. This survey data consists of responses from all the people above age 12 in the United States. This data have more than 300 columns and vast number of respondents. This study is mostly focused on impact of the usage of marijuana on the youth. We can find answers for how much the marijuana is used, how many days in the last year marijuana used by the youth , how many days of school was missed due to marijuana usage based on the factors such as situations in school as fight, friends impact on usage of marijuana, parents strictness , grades and religious beliefs. In order to analyze all these factors , decision tree and ensembles where used to conclude to the potential factors in order to prevent and reduce the youth from usage of marijuana.

Theoretical background:

Decision Trees: In machine learning, decision trees are an essential tool for both regression and classification problems. They use basic decision rules deduced from the data features to build a model that forecasts the value of a target variable. The goal of the decision tree approach is to maximize the homogeneity of the target variable inside each partition by recursively splitting the data based on feature values. Advantage of decision tree is it is easy to interpret and visualize but it prones to overfitting with complex datasets.

Random Forest: Several decision trees are combined in Random Forest, an ensemble learning technique, to increase prediction accuracy and decrease overfitting. Every tree in the forest receives training using a random subset of the features, a random portion of the training data, and a replacement sample. Because of the diversity that this randomness creates among the trees, the model becomes more reliable and accurate. The number of trees in the forest, the maximum depth of each tree, and the quantity of characteristics to take into account for each split are among the hyperparameters to take into account.It is advantageous by providing robust predictions by reducing overfitting through ensemble averaging but it requires more computational resources and slow to train.

Bagging: Also known as Bootstrap Aggregating, bagging is a broad ensemble technique that builds several models separately before combining their predictions by voting (for

classification) or averaging (for regression). Using bootstrap samples of the training data, multiple decision trees are trained in the context of bagging. The unpredictability that the bootstrapping method introduces is what gives the models their diversity. The number of models to train and the maximum depth of each tree are two examples of hyperparameters to take into account.Bagging reduces variance and improves stability averaging predictions from multiple models but improvement cannot be found if it has strong base model.

Boosting: By gradually adding weaker learners, this ensemble strategy creates a strong learner. Each new model improves overall performance by concentrating on the samples that the prior models struggled to classify. The number of boosting iterations, the maximum tree depth, and the learning rate(shrinkage values) are hyperparameters to take into account.Follow sequential learning models which is advantageous on difficult instances but sensitive to noisy data and outliers which leads to overfitting.

Classifiers and Regressors: Decision trees have the ability to function as both classifiers and regressors. Decision trees divide the feature space in classification tasks so that examples can be assigned class labels. They make continuous value predictions in regression tasks. Both kinds of jobs can be handled by the decision tree method by modifying the splitting criterion appropriately. Applicable for both categorical and continuous predictions but gives trouble with the complex relationships and tuning must be done inorder to avoid overfitting.

Pruning: Pruning is a strategy that removes elements of the tree that do not significantly contribute to forecast accuracy in order to minimize overfitting in decision trees. This entails developing the entire tree and then eliminating any nodes that don't enhance the validation set's performance. Often used pruning methods include reduced-error and cost-complexity pruning, sometimes referred to as weakest link pruning. The pruning method and the pruning criterion are examples of hyperparameters to take into account.It simplifies decision tree and improves interpretability but selection of pruning parameters must be cautioned.

Methodology:

The aim of this study is to predict the use of marijuana by the youth based on the parameters which might influenced the youth.This methodology takes data in from National Survey on Drug Use and Health (NSDUH) dataset . NSDUH consists of uncleaned data inorder to analyse and give as input to our models. So our first step will be cleaning and as our main focus is on youth data and store it as Youth data csv file.

Data Preparation: The dataset youth data csv file provided is loaded using pandas. Inorder to analyse the data present we have printed the columns and went through each column which is represented in code words. Each column is understood using the codebook. While analysing the data set we have found out that the dataset contains missing and null values these values are treated by introducing a simple imputation method in which each columns having missing values are replaced with most frequent values. Now, all the column names are in the form of codes , with the help of codebook the names of required columns are renamed for easy access and to be more efficient.

Exploratory Data Analysis and Feature Selection: To understand the data more deeply and to select the features for the classification regression , the unique values for each column are printed for better analysis.

Modelling for Binary Classification and Multi-Class Classification : In the classification steps we addressed questions such as whether the youth consumed the marijuana or not and how other factors are responsible for the outcome of the usage. These factors are demographic and behavioral. In multi classification we addressed how many days in the past year the youth used marijuana and the influence of other factors is also analyzed by plotting the importance variable plot. In order to analyze these questions we have to split the data into training and testing set of data and evaluate the performance using accuracy score. To get the accuracy score we have used Decision Tree classifier, Random Forest Classifier, Bagging Classifier and Boosting Classifier is used with multiple shrinkage values. For all the classifiers cross validation is done and the best model is chosen with highest accuracy and importance plot is plotted for the analysis.

Modeling for Regression: In regression our problem is to know how many days the school was missed due to drug usage. We define the problem as a target variable. Followed by splitting training and testing data sets. The Decision Tree Regressor ,Random Forest Regressor, Bagging Regressor and Gradient Boosting Regressor are used on the training data. Each model performance is evaluated using cross validation and the best model is selected on the basis of mean squared error(MSE) value. The best model importance of features is plotted. For better understanding decision tree and pruned decision tree is plotted.

Computational Results:

Binary Classication: Our problem in binary classification is to find whether the youth use marijuana or not. For that we have choosen the variable that represents whether youth uses marijuana or not and the other features as the predictors such as the overall health, mother presence in the life of the youth, father presence in the life of the youth, youth friends religious beliefs, gender of the youth and the average grade in the school. Now since the data is prepared for the analysis the data is sent through various classification methods. Before that the data is split into training and testing sets with a test size 20% and random state 42 for reproducibility.

Used Decision Tree Classifier on the training data and evaluated models performance on the testing data using accuracy score. Performed cross validation with 5 folds to asses the model stability and generalization performance . We got accuracy of 0.8548 for Decision Tree Classifier.

Next used Random Forest Classifier with 100 estimators on the training data and evaluated the model's performance using cross validation to obtain the mean accuracy score. We got accuracy of 0.860798560756699 for the Random Forest Classfier.

For Bagging Classifier with 100 estimators on the training data and evaluated the model's performance using cross validation to obtain the mean accuracy score. We got accuracy of 0.8607975640508121 for the Bagging Classfier.

We can observe there is very negligible difference in the accuracy acquired in between bagging and random Forest Classification method.

Boosting Classifier: We have performed boosting with multiple shrinkage values by predefining it with 100 estimators on the training data set. Followed by cross validation and

identify best model in the boosting that is identifying which shrinkage value gives best model. We got best model with shrinkage value 0.1 and accuracy 0.8633177.

From all the classifications we can say that the boosting has the highest accuracy compared to others. So we have plotted feature importance plot .

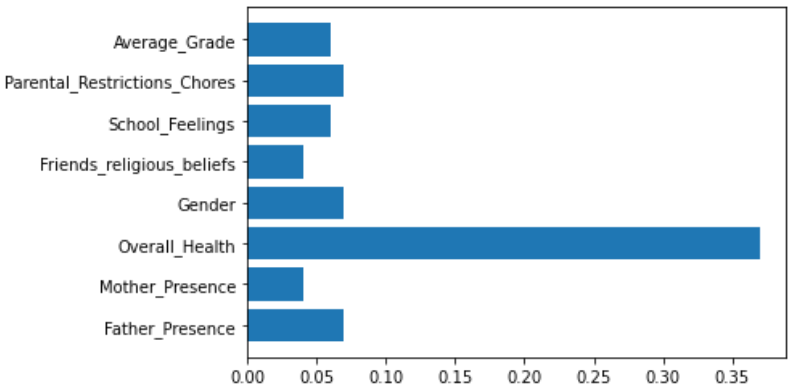

Fig 1. Feature Importance plot -Binary Classification

The above feature importance plot shows that the Overall Health , Parental restriction on youth and school feelings having larger impact and influence on the youth using marijuana. They play key role whether youth will be subjectable to marijuana usage or not.

Multi-Class Classification: Our problem in multi-class classification is to analyse what features influenced the usage of marijuana on days in past year,our target variable will be that and the features will be Drugs sold by youth, Fights between youth at the school, aruguments with parents, average grade, restrictions imposed by the parents, religious beliefs of friends, overall health, thoughts of friends on marijuana and presence of the parent. Now since the data is prepared for the analysis the data is sent through various classification methods. Before that the data is split into training and testing sets with a test size 20% and random state 42 for reproducibility.

Used Decision Tree Classifier on the training data and evaluated models performance on the testing data using accuracy score. Performed cross validation with 5 folds to asses the model stability and generalization performance . We got accuracy of 0.84027 for Decision Tree Classifier.

Next used Random Forest Classifier with 100 estimators on the training data and evaluated the model's performance using cross validation to obtain the mean accuracy score. We got accuracy of 0.86079 for the Random Forest Classfier.

For Bagging Classifier with 100 estimators on the training data and evaluated the model's performance using cross validation to obtain the mean accuracy score. We got accuracy of 0.86557 for the Bagging Classfier.

We can observe there is very negligible difference in the accuracy acquired in between bagging and random Forest Classification method.

Boosting Classifier: We have performed boosting with multiple shrinkage values by predefining it with 100 estimators on the training data set. Followed by cross validation and identify best model in the boosting that is identifying which shrinkage value gives best model. We got best model with shrinkage value 0.1 and accuracy 0.876578408.

From all the classifications we can say that the boosting has the highest accuracy compared to others. So we have plotted feature importance plot .

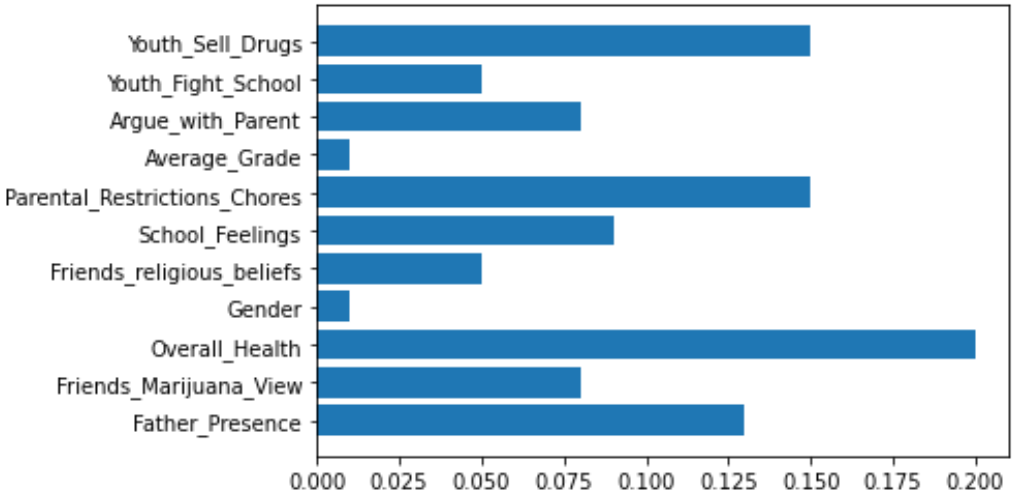

Fig 2. Feature Importance plot - Multi-Class Classification

From the plot we can say that the usage of marijuana is influenced highly because of the overall health ,youth selling drugs, restrictions imposed by the parents , presence of the parent seems to be highly influential lead the youth to take marijuana past year.Where as fights at school, arguments with parents ,friends religious beliefs have moderate influence on the youth using marijuana past year.

Regression: In regression our problem is to find how many days of school is missed due to marijuana usage by the youth, our target variable will be that and the features will be Drugs sold by youth, Fights between youth at the school, students using marijuana, aruguments with parents, average grade, restrictions imposed by the parents, religious beliefs of friends,current grade at school,average grade at school of the youth, overall health, thoughts of friends on marijuana and presence of the parent. Now since the data is prepared for the analysis the data is sent through various classification methods. Before that the data is split into training and testing sets with a test size 20% and random state 42 for reproducibility.

Used Decision Tree Regressor on the training data and evaluated models performance on the testing data using MSE value. Performed cross validation with 5 folds to asses the model stability and generalization performance . We got MSE of 8.511 for Decision Tree Regressor.

Next used Random Forest Regressor with 100 estimators on the training data and evaluated the model's performance using cross validation to obtain the MSE . We got cross validation MSE of 5.84301 for the Random Forest Regressor.

For Bagging Regressor with 100 estimators on the training data and evaluated the model's performance using cross validation to obtain the MSE. We got cross validation MSE of 5.8337 for the Bagging Regressor.

We can observe there is very negligible difference in the MSE acquired in between bagging and random Forest regressor method.

Boosting Regressor: We have performed boosting with multiple shrinkage values by predefining it with 100 estimators on the training data set. Followed by cross validation and identify best model in the boosting that is identifying which shrinkage value gives best model. We got best model with shrinkage value 0.05 and cross validation MSE 4.55355.

From all the classifications we can say that the boosting has the lowest cross validation MSE compared to others. So we have plotted feature importance plot .

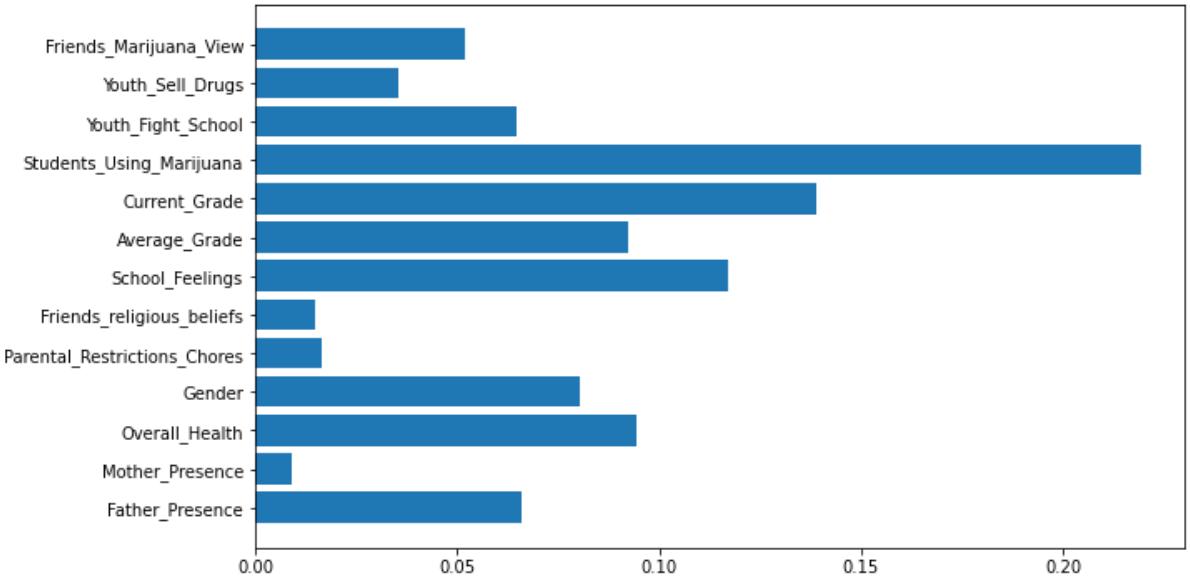

Fig 3. Feature Importance plot – Regression

In this feature importance plot we can observe that the reason or features responsible for missing the school is high for students using marijuana, followed by the grade of the student also determine not going to school. Friends of the youth having thoughts on the marijuana is also influential on missing school and also the youth gets influenced by the parent presence mostly of the father and the drugs sold by the youth also shows moderate influence.

For the regression we have also plotted the pruned tree is shown in Fig 4.

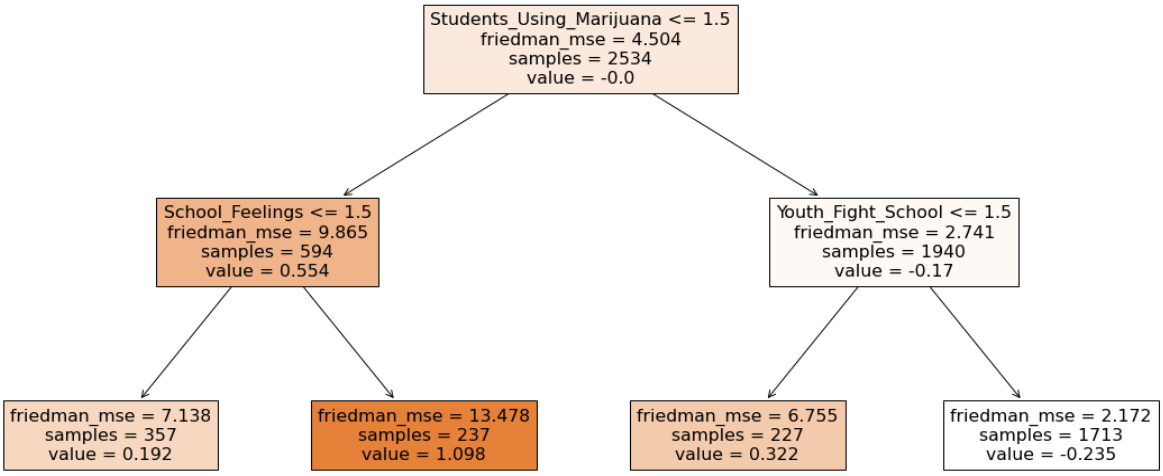

Fig 4. Pruned Tree-Regression

For the above pruned tree is about the factors associated with youth missing school,based on the features given to the regressor. Here we can see that the root node is Student using marijuana with a threshold value . If the total number of students using marijuana is less than or equal to the threshold values we move to left branch , if not we move to right. Coming to the left branch the tree splits into the feature of students feeling on school. If feels about the school is less than the threshold value then that means missing the school is likely lower chance. If the school feelings are more than the threshold value the chances of missing the school is higher. Lastly on the right branch the tree splits on the feature of fights of the youth at school. If the frequency of youth fighting at school is higher than the threshold the chances are low for missing the school. If the frequency of fights at school of youth is lower than equal to the threshold the chances of missing the school is onto certain level.

Discussion:

For Binary Classification feature importance plot shows that the Overall Health , Parental restriction on youth and school feelings having larger impact and influence on the youth using marijuana. They play key role whether youth will be subjectable to marijuana usage or not.

For Multi-Class Classification feature importance plot shows that the usage of marijuana is influenced highly because of the overall health ,youth selling drugs, restrictions imposed by the parents , presence of the parent seems to be highly influential lead the youth to take marijuana past year.Where as fights at school, arguments with parents ,friends religious beliefs have moderate influence on the youth using marijuana past year.

For Regression this feature importance plot we can observe that the reason or features responsible for missing the school is high for students using marijuana, followed by the grade of the student also determine not going to school. Friends of the youth having thoughts on the marijuana is also influential on missing school and also the youth gets influenced by the parent presence mostly of the father and the drugs sold by the youth also shows moderate influence

Overall feature importance plot shows a critical factors influencing youth to use marijuana. Mainly family dynamics , restrictions imposed by the parents and overall health seems to be more of an issue with the marijuana usage. Friends beliefs , adolescent behaviors imposed by the youth which where not controllable by the parents and school as well. Friends thoughts on the marijuana usage , feeling about the school, fights in the school among the youth , low and high grades in school seems to be influential on youth for missing school and using marijuana uncontrollably.

The pruned tree shows the patterns of missing school by youth. It shows factors such as the marijuana usage, school feelings and frequency of fights at school be related with absent in school among the youth.

Conclusion:

In this study, we looked at why young people use marijuana. We used NSDUH survey of people in the US who are 12 years old and up. There were lots of things to consider, like how often they used marijuana, how many days they missed school because of it, and what situations at school might be linked to it, like fights or how they felt about school. We used fancy math to figure out which things were most important in predicting if a young person might use marijuana.

We found that family, friends, and school are really important. For example, if parents are around a lot and set rules, it's less likely a young person will use marijuana. And how a young person feels about school and how well they're doing in school also matter.

Overall, we learned that stopping young people from using marijuana isn't just about telling them not to. It's about understanding what's going on in their lives and helping them make good choices. This study gives us a good start in figuring out how to help young people stay away from marijuana.

References:

[1]National Survey on Drug Use and Health. (2020). https://www.samhsa.gov/data/release/2020-national-survey-drug-use-and-health-nsduh-releases

[2] Scikit-learn Documentation. https://scikit-learn.org/0.21/documentation.html

Appendix:

```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import accuracy_score
from sklearn.tree import export_text
from sklearn.tree import plot_tree
from  sklearn.impute import SimpleImputer
from sklearn.ensemble import GradientBoostingClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.ensemble import RandomForestRegressor
from sklearn.metrics import mean_squared_error,r2_score
from sklearn.metrics import accuracy_score, classification_report
import matplotlib.pyplot as plt
import seaborn as sns

path = "youth_data.csv"
d = pd.read_csv(path)
print(d.columns)
def imp(d):
    return d.mode().iloc[0]
for col in d.columns:
    imp_v=imp(d[col])
    d[col].fillna(imp_v,inplace=True)
n_after=d.isnull().sum()
columns_after_replace=sum(n_after>0)
print(columns_after_replace)

column_mapping = {
    'iralcfy':'Alcohol_frequency',
```

```
    'irmjfy':'Marijuana_freuency',
    'mrjflag': 'Marijuana_Use_Flag',
    'alcflag': 'Alcohol_Use_Flag',
    'tobflag': 'Tobacco_Use_Flag',
    'mrjydays': 'Marijuana_Days_Last_Year',
    'schfelt': 'Feelings_About_School',
    'tchgjob': 'Teacher_Job_Performance',
    'stndscig': 'Cigarette_Standard',
    'stndalc': 'Alcohol_Standard',
    'stnddnk': 'Drinking_Standard',
    'parchkhw': 'Parents_Check_Homework',
    'PRCHORE2': 'Parental_Restrictions_Chores',
    'PRLMTTV2': 'Limiting_TV_Time',
    'parlmtsn': 'Limitations_Snacks',
    'FRDPCIG2': 'Friends_Cigarette_Use',
    'FRDMEVR2': 'Friends_Ever_Marijuana_Use',
    'PREVIOL2': 'Previous_Violence',
    'PRVDRGO2': 'Parental_Encouragement_Do_Good',
    'irsex': 'Respondent_Sex',
    'NEWRACE2': 'Race',
    'imother': 'Presence_Mother',
    'ifather': 'Presence_Father',
    'income': 'Income_Level',
    'POVERTY3': 'Poverty_Status',
    'COUTYP4': 'County_Type'
}


d.rename(columns=column_mapping, inplace=True)
d.columns
```

```
for column in d.columns:
    unique_values = d[column].unique()
    print(f"'{column}': {unique_values}")
features = ['Father_Presence','Overall_Health', 'Gender', 'Friends_religious_beliefs', 'School_Feelings', 'Parental_Restrictions_Chores','Average_Grade']
X = d[features]
y = d['Marijuana_Flag']


X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)


tree = DecisionTreeClassifier()
tree.fit(X_train, y_train)
y_pred = tree.predict(X_test)
accuracy = accuracy_score(y_test, y_pred)
dt_model = DecisionTreeClassifier(random_state=42)
dt_scores = cross_val_score(dt_model, X, y, cv=5)
print("\nDecision Tree Cross-validation scores:", dt_scores)
print("Mean Cross-validation accuracy - Decision Tree:", dt_scores.mean())


rf_model = RandomForestClassifier(n_estimators=100, random_state=42)
rf_model.fit(X_train,y_train)
rf_scores = cross_val_score(rf_model, X, y, cv=5)
print("Random Forest Cross-validation scores:", rf_scores)
print("Mean Cross-validation accuracy - Random Forest:", rf_scores.mean())


bagging_model = BaggingClassifier( n_estimators=100, random_state=42)
bagging_model.fit(X_train,y_train)
bagging_cv_scores = cross_val_score(bagging_model, X, y, cv=5)
print("Bagging Cross-validation scores:", bagging_cv_scores)
print("Mean Cross-validation accuracy - Bagging:", bagging_cv_scores.mean())
```

```
shrinkage_values = [0.1, 0.05,0.025, 0.01]
best_model = None
best_accuracy = 0
for shrinkage in shrinkage_values:
    boosting_model = AdaBoostClassifier(n_estimators=100, learning_rate=shrinkage, random_state=42)
    boosting_model.fit(X_train,y_train)
    cv_scores = cross_val_score(boosting_model, X, y, cv=5)
    mean_accuracy = cv_scores.mean()
    print("Shrinkage :", shrinkage)
    print("Cross-Validation Scores:", cv_scores)
    print("Mean Accuracy:", mean_accuracy)
    print()
    if mean_accuracy > best_accuracy:
        best_accuracy = mean_accuracy
        best_model = boosting_model
print("Best Model:")
print(best_model)
print("Best Accuracy:", best_accuracy)

best_model.fit(X_train,y_train)
importances=model.feature_importances_
plt.figure(figsize=(10,6))
plt.barh(range(len(features)),importances,align="center")
plt.yticks(range(len(features)),features)

features = ['Father_Presence','Friends_Marijuana_View', 'Overall_Health', 'Gender', 'Friends_religious_beliefs', 'School_Feelings', 'Parental_Restrictions_Chores','Average_Grade','Argue_with_Parent'
,'Youth_Fight_School','Youth_Sell_Drugs']
```

```
X = d[features]
y = d['Marijuana_Days_Last_Year']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)



dt_model = DecisionTreeClassifier(random_state=42)
dt_scores = cross_val_score(dt_model, X, y, cv=5)
print("\nDecision Tree Cross-validation scores:", dt_scores)
print("Mean Cross-validation accuracy -Decision Tree:", dt_scores.mean())
rf_model = RandomForestClassifier(n_estimators=100, random_state=42)
rf_scores = cross_val_score(rf_model, X, y, cv=5)
print("Random Forest Cross-validation scores:", rf_scores)
print("Mean Cross-validation accuracy -Random Forest:", rf_scores.mean())



base_estimator=DecisionTreeClassifier()
bagging_model = BaggingClassifier( n_estimators=100, random_state=42)
bagging_model.fit(X_train,y_train)
bagging_cv_scores = cross_val_score(bagging_model, X, y, cv=5)

print("Bagging Cross-validation scores:", bagging_cv_scores)
print("Mean Cross-validation accuracy -Bagging:", bagging_cv_scores.mean())


shrinkage_values = [0.1, 0.05,0.025, 0.01]
best_model = None
best_accuracy = 0
for shrinkage in shrinkage_values:
    boosting_model = AdaBoostClassifier(n_estimators=100, learning_rate=shrinkage,
random_state=42)
    cv_scores = cross_val_score(boosting_model, X, y, cv=5)
```

```
    mean_accuracy = cv_scores.mean()
    print("Shrinkage (Learning Rate):", shrinkage)
    print("Cross-Validation Scores:", cv_scores)
    print("Mean Accuracy:", mean_accuracy)
    print()
    if mean_accuracy > best_accuracy:
        best_accuracy = mean_accuracy
        best_model = boosting_model
print("Best Model:")
print(best_model)
print("Best Accuracy:", best_accuracy)


best_model.fit(X,y)
importances=best_model.feature_importances_
plt.barh(features,importances)


features = ['Father_Presence', 'Mother_Presence', 'Overall_Health', 'Gender', 'Parental_Restrictions_Chores','Friends_religious_beliefs', 'School_Feelings', 'Average_Grade','Current_Grade','Students_Using_Marijuana','Youth_Fight_School','Youth_Sell_Drugs','Friends_Marijuana_View']
X = d[features]
y = d['Days_Missed_School']



X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)


dt_regressor = DecisionTreeRegressor(random_state=42)
dt_cv_scores = cross_val_score(dt_regressor, X, y, cv=5, scoring='neg_mean_squared_error')
print("Decision Tree Cross-Validation MSE:", -dt_cv_scores.mean())
```

```
rf_regressor = RandomForestRegressor(n_estimators=100, random_state=42)
rf_cv_scores = cross_val_score(rf_regressor, X, y, cv=5, scoring='neg_mean_squared_error')
print("Random Forest Cross-Validation MSE:", -rf_cv_scores.mean())

bagging_regressor = BaggingRegressor(n_estimators=100,random_state=42)
bagging_cv_scores = cross_val_score(bagging_regressor, X, y, cv=5, scoring='neg_mean_squared_error')
print("Bagging Cross-Validation MSE:", -bagging_cv_scores.mean())

learning_rates=[0.1,0.05,0.25,0.01]
for i in learning_rates:
    gb_regressor = GradientBoostingRegressor(n_estimators=100, learning_rate=i, random_state=42)
    gb_cv_scores = cross_val_score(gb_regressor, X, y, cv=5, scoring='neg_mean_squared_error')
    print("Gradient Boosting Cross-Validation MSE:", -gb_cv_scores.mean(),'for shrinkage value',i)

mse_values = {
    "Random Forest": -rf_cv_scores.mean(),
    "Bagging": -bagging_cv_scores.mean(),
    "Gradient Boosting": min(-gb_cv_scores),
    "Decision Tree": min(-dt_cv_scores)
}

best_model = min(mse_values, key=mse_values.get)
best_mse = mse_values[best_model]

print("The Best Model is:", best_model)
print("MSE of the Best Model is:", best_mse)
model=GradientBoostingRegressor(n_estimators=100, learning_rate=0.05, random_state=42)
model.fit(X,y)
```

```
importances=model.feature_importances_
plt.figure(figsize=(10,6))
plt.barh(range(len(features)),importances,align="center")
plt.yticks(range(len(features)),features)
from sklearn.tree import plot_tree

plt.figure(figsize=(20, 10))
plot_tree(model.estimators_[0][0], feature_names=features ,filled=True)
plt.show()



gb_regressor_pruned = GradientBoostingRegressor(n_estimators=100,
                                                max_depth=2,
                                                min_samples_split=5,
                                                min_samples_leaf=2,
                                                random_state=42)
gb_regressor_pruned.fit(X_train, y_train)
plt.figure(figsize=(20, 10))
plot_tree(gb_regressor_pruned.estimators_[0][0], feature_names=features, filled=True)
plt.show()
```