**Final Project**

**AP Test**

Thanate Banchonhattakij

Department of Computer Science, Fitchburg State University

CSC 7015: Introduction to Data Science

Dr. Ricky Sethi

**Abstract**

Advance placement (AP) test requires much time and efforts of students to pass the test. This research has shown knowledge of data science supports students decision making based on the statistical database. This study aims to determine the probability of exams in Mathematics and Science categories for students. If a student is only going to take one AP math or science course while in high school, which one should he or she take? In this context, there are many exams in the both categories for high school students who have options to take. Data mining techniques can figure out the answer by using the machine learning model to analyze the students’ data. Students were grouped into two groups (pass and fail), and the model uses the data to figure out which exam is the most critical in the pass group. Analysis of the AP test illustrated that the Science exam has more probability to pass for high school students more than Mathematics. The performance of the model is high accuracy and less bias result. On this basis, it is recommended that high school students should take the science exams. Further research is needed to gather more information about students that could strengthen the analysis more effective and deeper.

**Table of Contents**

[Introduction 4](#_Toc44164222)

[Data Appraisal 5](#_Toc44164223)

[Techniques 6](#_Toc44164224)

[Evaluation 8](#_Toc44164225)

[Model 11](#_Toc44164226)

[Results 14](#_Toc44164227)

[Limitation 16](#_Toc44164228)

[Conclusion 17](#_Toc44164229)

[References 18](#_Toc44164230)

# **Introduction**

Advance placement (AP) provides students an opportunity to tackle college-level work while they’re still in high school and earn college credit and placement. AP tests are standardized exams designed to measure how students have mastered the content and skills of a specific AP course.

This report uses a dataset from the College Board and knowledge of data mining to produce the result of the project. The report contains the processes of data preparation, model, evaluation, and conclusion about data analysis of AP test. Data preparation is a significant task to prepare the proper the proper input dataset, compatible with the machine learning algorithm requirements, and it can improve the performance of the machine learning model. The model in this report is a classification decision tree model, and the design of the tree is Bottom up model to find the best exams having more probability to pass between math and science exams. Evaluation of the model is to use cross validation technique called k-fold cross validation with repetition.

**Research Question**

The preparation for taking AP tests requires time and efforts, and some students may not have enough time to read and take many AP tests. The research question for AP test is that if a student is only going to take one AP math or science course while in high school, which one should he or she take? So, the purpose of this report is to figure out which kind of exam in Mathematics and Science categories is better probability to pass the exams.

# **Data Appraisal**

The data utilized in this study case was the national summary report in 2013 from The College Board as an excel file. According to AP central in the College Board website, the website releases summary reports outlining student and school participation and performance data on a national and state-by-state basis for that year’s AP Exams for students, teachers, and data scientist who are interested in AP test. The report contained the information of students, distributed by gender, 11th and 12th grade student, and college’s location, for each examination and ethnic group in 2013.

This project focuses on the first sheet of the dataset that illustrates the number of all students taking each examination divided by ethnic group in 2013, and the exams in math and science categories. According to The College Board (n.d.), AP math exams contain Calculus AB, Calculus BC, and Statistics; on the other hand, AP Science exams consist of Biology, Chemistry, Environmental Science, Physics B, Physics C: Electricity and Magnetism, and Physics C: Mechanics.

According to the dataset, the dataset consisted of nine ethnic groups and one all nations, and each group contains five rows included each AP score for each exam. The values in the table were the number of students taking each exam and how many score they got in scale one to five. This means the size of the dataset is quite small (50 rows).

# **Techniques**

Data preparation is the process of cleaning and transforming raw data prior to processing and analysis. It is an important step prior to processing and often involves reformatting data, making corrections to data and the combining of data sets to enrich data (Pearlman, 2020). For this report, there are several steps for preparation.

1. Gathering the data and important package in R

This step is to import the dataset excel file as an R object, and install the significant package for data analysis.

1. Filtering and Cleaning the data

Some values in the data were missing and unnecessary because the focused exams in this study case is only math and science exams, so filtering the dataset was the step to clean the data. Using the data having missing or unnecessary values can make the wrong result and performance of the model. The missing values in this dataset was that the number of students who got the AP score is zero or no student took the exam.

1. Transforming the data

This step is to transform the number of students into a percentage format for each exam and ethnic group because the number of students taking each exams is different. The proportion of students was created on the number of students in each AP score divided by the total number of student taking this exam and multiple 100. Transforming the values in the dataset can improve the result and accuracy of the model.

According to the dataset after preparation, the size of the dataset is quite small that has 5 rows per ethnic group and one total nation (totally 50 rows). Due to the small size of training dataset, the decision tree model might miss out on some interesting information about the data which will lead to a higher bias (Sunil, 2018). To solve this problem, the k-fold cross validation with repetition technique was applied to the model to perform the training and testing process multiple times by splitting a dataset into k-folds per ethnic group, and each fold was used to create a model with the training and test data (Brownlee, 2018). Additionally, using the k-fold cross validation with repetition technique can optimize the decision tree model in this situation.

To determine a successful analysis, the diagram of the all trees will show the first splitting of the models at the root that is the most critical exam between math and science exams. Furthermore, to describe the performance of a classification model on a test data, a confusion matrix is a popular table that consists of four different combinations of predicted and actual values that can measure the accuracy, recall, and precision of the model (Brownlee, 2014).

This research has no potential ethical and legal compliance issues because the dataset is available for everyone who is interested in AP test. According to the College Board (n.d), the summary AP reports provide a broad overview of AP Program participation, and the score distribution reports include detailed information about which AP scores were earned. The both report are public and available for everyone to download. Therefore, this report can be a study case for data analysis students who want to learn and practice the real world experience with the real world data.

# **Evaluation**

The model of this analysis is a classification decision tree model because the answer of the research question is a categorical answer (Math or Science). According to Chakure (2019), tree models where the target variable can take a discrete set of values are called classification trees. Many organizations prefer to use decision tree models for decision making because the model is easy to interpret for people who don’t have analytical background (Magee, 1964).

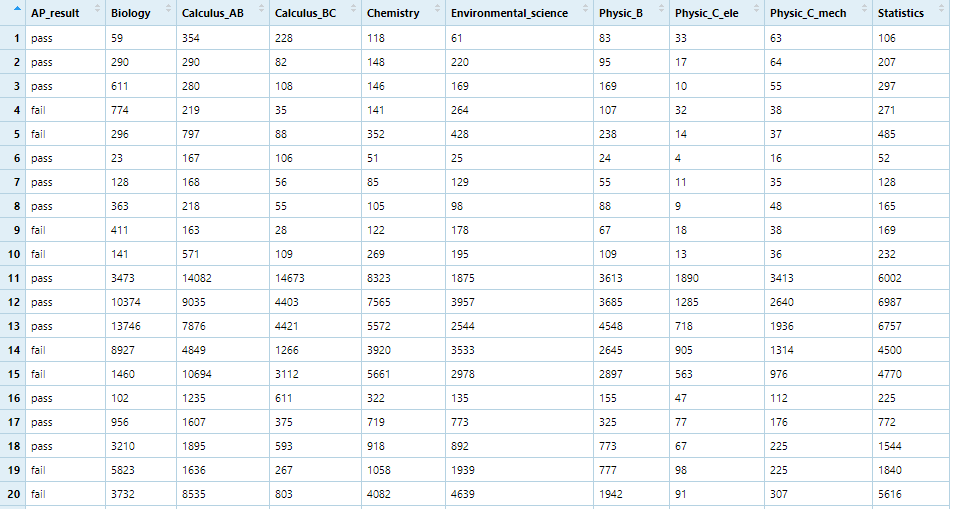
The main purpose of this report is to answer the research question based on the dataset. Although this dataset is enough to answer the question, it still lacks off some information to analyze deeper. The deeper analysis requires more details about students who taking the AP test. The information about the students, such as age, school, gender, and specific details, can support the future and deeper analysis. Furthermore, the analysis can be used to other kinds of exams that might find which exam is good for each student’s categories. The more information of students, the deeper and more accurate analysis. Age and gender of students might impact the results of the exams. For example, 10-12 year-old female students pass the science tests more than male.

According to the College Board, the summary and score contribution reports is public and available in both PDF and Excel formats. The reports include detailed data separated into categories that can be used to evaluate relative performance during a single school year or over a span of years. Furthermore, the College Board support students, teachers, or analysts who need more information or specific information. They can request data based on information collected, derived and generated from student and institutional participation in College Board programs. Data can be requested for most College Board K¬20 programs including, but not limited to, student, item-level test, and customized aggregate data not publicly available.

The analysis processes in this analysis include data preparation, mining and modeling procedures, and mensuration the model.

**Data preparation**

This process is to gather, filter, clean, and transform the raw data in to the proper format. According to the technique part, the datasets after preparation were arranged in vertical columns by the categories of exams; “AP result”, “Biology”, “Calculus AB”, “Calculus BC”, “Chemistry”, “Environmental Science”, “Physics B”, “Physics C: Electricity and Magnetism”, “Physics C: Mechanics”, and “Statistics”. Horizontal rows were arranged by the number of AP score for students from different ethnic groups. Furthermore, AP score of students was contributed to two groups, pass and fail, based on the score. If the score is three or more than three, students passed the tests; otherwise, student who get the score less than three failed the test.



Then, separating the number of student in each ethnic group (every 5 rows) to be groups that is easy to transform the data to the percentage format and prepare for the k-fold cross validation technique. As a result, the values in each fold (total 10 folds) were the percentage of students in each ethnic group.

**Data mining and modeling**

This process is to create the training and test data in each ethnic group or fold, then use the data to build the classification decision tree model using k-fold cross validation. All the ten models were created predictors and check with the test data. To create a model, using the ‘Rpart’ function for the models from the ‘Rpart’ package in R language can create a classification model of decision tree. The function has many options to modify the model, this research uses the function of classification model.

**Mensuration the model**

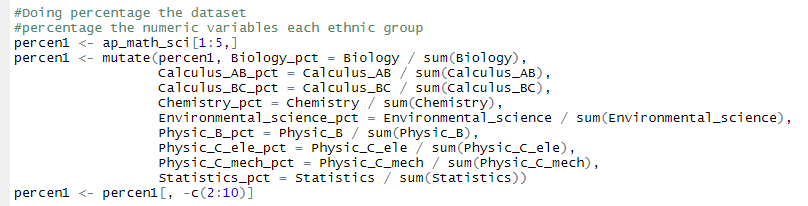
To measure the performance of the models, a confusion matrix is a clean and unambiguous way to present the prediction results of a classifier (Brownlee, 2014). For all ten models, the accuracy, recall, precision, and F1 score were averaged that can present the overall performance of the models. On the other hand, the diagrams of decision tree models were analyzed which exams between math or science categories occurs at the first splitting of the tree. If the science exam happens at the first splitting that means the science exam is the most critical exam for students who have only one chance to take.

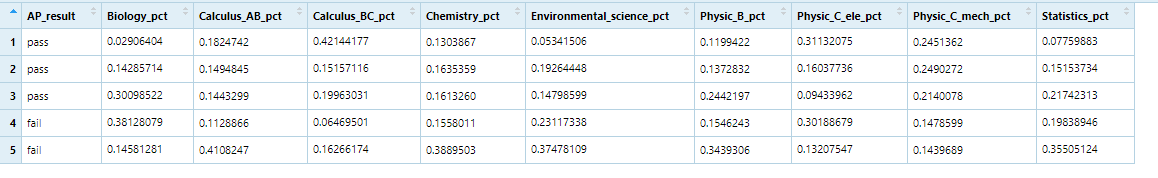
# **Model**

The purpose of the machine leaning model, a classification decision tree model, is to answer the research question that which exam between math and science categories of AP test has the most probability to take for students who have only one chance. According to Analytics Vidhya (2016), the structure of a decision tree model has five main parts; Root node, Splitting, Decision node, Leaf or Terminal node, and Branch or Sub tree. The root node is the most numerous groups, and the first splitting at the node is the most critical factor of the tree.

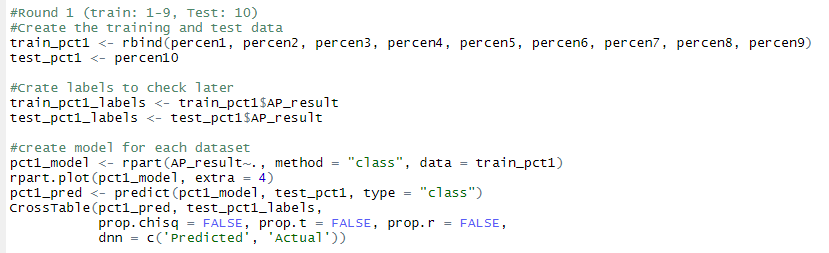
The model in this research uses the ‘rpart’ command to build a decision tree model. The command can modify the algorithm of splitting of the model, and the default algorithm of the command is Gini. One more reason to use this command, the plot of the ‘rpart’ package is easier than the common plot command in R, and the plot can modify and add more details on the diagram.

After the processes of data preparation, the dataset was separated to be ten folds, and the values of each fold were the proportion of students taking each exam. The programming steps for transforming the dataset and the example of each fold or ethnic group were shown below:



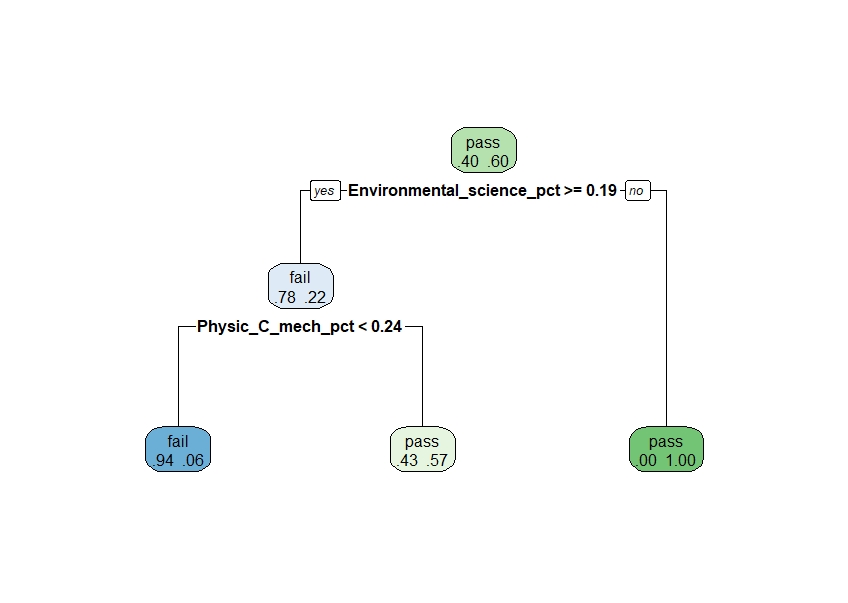


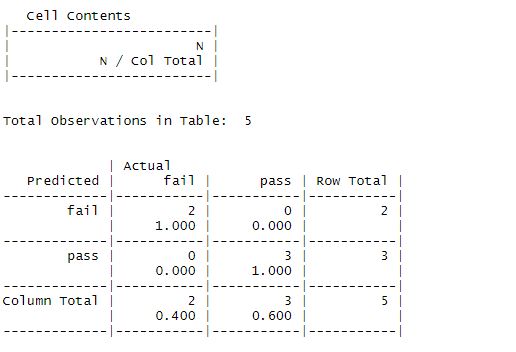
The 10 folds were combined to be training and test data followed the concept of k-fold cross validation. According to Brownlee (2018), the general procedure of k-fold cross validation is to build the model with 9 folds of training data and test with one fold. So, the number of doing k-fold cross validation depends on the number of fold that is 10 rounds. Each round contains the processes of creating training and test data, recording the labels of both data for a confusion matrix, building the model for this round, plotting the diagram of the tree, generating a predictor, and creating a confusion matrix. The programming steps for these processes were shown below:



Due to the small size of the dataset, building the decision tree models using k-fold cross validation can reduce the bias of the data, and enhance the performance of the model. The results of the ten models were analyzed into two parts. First the average of the performance of the ten models was calculated by using a confusion matrix to find the accuracy, recall, precision, and F1 score of the models. Second, the diagrams of the ten models was reviewed which exam mostly occurs at the first splitting.

The example of diagram and the confusion matrix of the model were shown below:





# **Results**

The results of the models were separated to be two parts. First, the performance of the model, such as accuracy, recall, precision, and F1 score, was averaged from the ten models of 10 folds, and then presented as a confusion matrix. According to Narkhede (2018), a confusion matrix consists of four different combinations of predicted and actual values that can measure the accuracy, recall, and precision of the model.

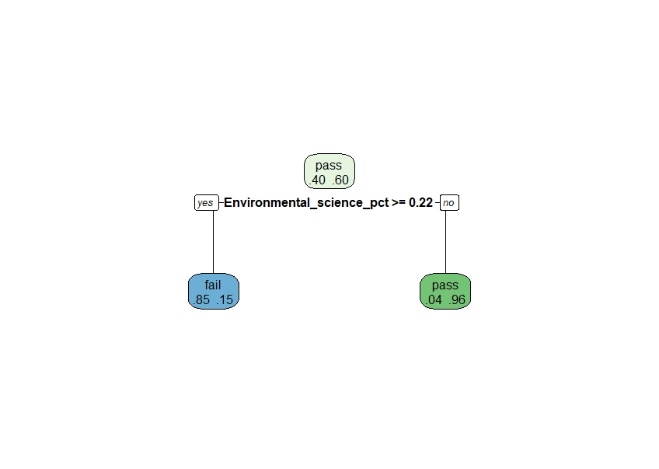
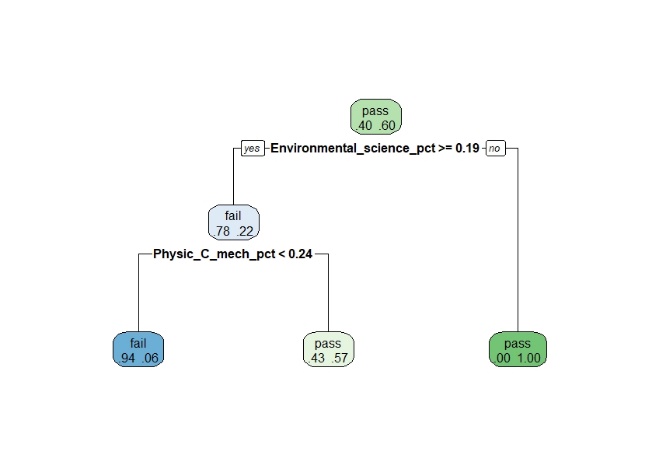
* Accuracy: Overall, how often is the classifier correct?
* Recall: For all positive cases, how often is the classifier correct?
* Precision: When model predicted correctly, how often is it correct?
* F1 score: A balanced mean between precision and recall

The comparison between the model using k-fold cross validation with data preparation and the regular model was shown below

|  |  |  |
| --- | --- | --- |
|  | Regular model | The model (k-fold cross validation) |
| Accuracy | 76% | 88% |
| Recall | 100% | 90% |
| Precision | 66% | 80% |
| F1 score | 79% | 85% |

Although some values of the regular model is higher than the model using k-fold cross validation, the regular model has more bias that means if the test data was changed, the performance of the regular model could drop. Therefore, the results of the model using k-fold cross validation is reasonable because it has more accuracy and less bias of the data.

On the other hand, the diagrams of the tem models have the similar results. The majority of the diagram shows the Environmental Science is the first splitting of the tree.

According to the trees, most of the models illustrate that the exam of environment science in the science category was the most critical exam. This means science exams have more probability than mathematics exams to be successful based on the both models.

The processes of data preparation and the cross validation technique mainly influenced the results of the model. Data preparation is to clean and transform from the raw data to be the proper data for the models. According to Geller (2019), using the right feature scaling method can significantly increase the accuracy of the models. The common errors of building the model are the data format before building the model because the wrong input cannot build the right model. The limitation the data, such as size, bias, missing values can influence the result and performance of the model.

# **Limitation**

The dataset was a summary report in 2013 from the College Board website that included the number of students who took the AP exams and the number of the students in each AP score. In the new version of the summary report on the website, the report was separated to many formats and more information; however, the personal information of students does not include in any report. The personal details of students, such as age, gender, and school name, can extend and improve the research and analysis more meaningful, accurate, and deeper. Therefore, the research in this report was based on the dataset that was public and available for everyone to access the data.

# **Conclusion**

This study case shows that how data mining can be used in the research of AP test. This report firstly introduce the related information and knowledge about AP tests and the processes of analysis using the machine learning model to answer the research question of the project. The processes included data preparation, model, evaluation, and the result of the research based on the dataset from the College Board website in 2013. Data preparation contains the cleaning and transforming steps the raw data to be the proper data for building the model. The model was used in this research is a classification model of decision tree. Due to the small size of the dataset, using k-fold cross validation can reduce the bias of the data and improve the performance of the model. Furthermore, the performance of the model was described in a confusion matrix that can calculate the accuracy, recall, precision, and F1 score of the model.

The results of the research illustrate the student having only one chance to take should take the science exams based on the diagrams of the decision tree models. The accuracy, recall, precision, and F1 score of the model are 88%, 90%, 80%, and 85%, respectively. The overall of the performance of the model is good based on the dataset. Therefore, the students should decide to take the exams in science category that have more probability to pass the AP test than the mathematic category.

# **References**

Analytics Vidhya (2016) Tree Based Algorithms: A Complete Tutorial from Scratch (in R & Python) Retrieved from <https://www.analyticsvidhya.com/blog/2016/04/tree-based-algorithms-complete-tutorial-scratch-in-python/>

Brownlee, Jason (2014) Machine Learning Mastery: Classification Accuracy is Not Enough: More Performance Measures You Can Use Retrieved from <https://machinelearningmastery.com/classification-accuracy-is-not-enough-more-performance-measures-you-can-use/>

Brownlee, Jason (2018) Machine Learning Mastery: A Gentle Introduction to k-fold Cross-Validation Retrieved from <https://machinelearningmastery.com/k-fold-cross-validation/>

Chakure, Afroz (2019) Toward Data Science: Decision Tree Classification Retrieved from <https://towardsdatascience.com/decision-tree-classification-de64fc4d5aac>

Geller Shay (2019) Toward Data Science: Normalization vs Standardization — Quantitative analysis Retrieved from <https://towardsdatascience.com/normalization-vs-standardization-quantitative-analysis-a91e8a79cebf>

Magee, John F. (1964) Harvard Business Review: Decision Trees for Decision Making Retrieved from <https://hbr.org/1964/07/decision-trees-for-decision-making>

Pearlman, Shana (2020) Talend: What is Data Preparation? Retrieved from https://www.talend.com/resources/what-is-data-preparation/

Sunil, Ray (2018) Analytics Vidhya: Improve Your Model Performance using Cross Validation (in Python and R) Retrieved from <https://www.analyticsvidhya.com/blog/2018/05/improve-model-performance-cross-validation-in-python-r/>

The College Board (n.d.) AP Central: Course and Exam Pages Retrieved from https://apcentral.collegeboard.org/courses