Identifying Underage Alcohol Abuse Risk

Sofie Kupiec and Alyssa Unell

May 9, 2022

1 Introduction

Alcohol abuse is a serious problem regardless of age. However, young people [1] are significantly more likely to abuse alcohol than adults. The National Institute on Alcohol Abuse and Alcoholism (NIAAA) defines binge drinking to be a "pattern of drinking that brings blood alcohol concentration (BAC) to 0.08%" [AA]. While young people drink less frequently than adults, a US-based survey shows that more than 90% of alcoholic drinks consumed by young people are consumed through binge-drinking behavior.[AA]

This is concerning as drinking, especially at higher or binge-drinking levels, can impair judgment and lead to a host of undesired situations both in society, such as disruptive behavior, and on an individual level, such as personal harm. In the United States, it is estimated that each year excessive drinking is responsible for over 3,900 deaths and costs the country $24 billion dollars[2][DP]. It is essential to combat underage drinking as it is associated with a myriad of dangerous consequences and there is a strong correlation between early initiation of drinking and the development of alcohol use disorders later in life. Fortunately, underage drinking, and specifically underage alcohol abuse, is preventable; however, the longer alcohol abuse goes unnoticed, the harder it is to undo. Thus, it is imperative to have methods in place to identify warning signs for underage alcohol abuse and implement a proactive prevention plan as early as possible.

Current methodologies for establishing students who are at risk for abusing alcohol rely on a counselor or other prominent community member to watch over students, look out for typical warning signs such as changes in mood, and exercise their best judgement to determine if a student requires additional screening and subsequent intervention or not. Given that the student-to-school-counselor ration in the United States is, on average, 415-to-1[3], this method leads to many students falling through the cracks, not being identified as 'at-risk,' and thus suffering the unfortunate consequences of underage alcohol abuse. [Ass] We hypothesize that there are objective signs (both from a student's home demographics as well as information about their school performance) that can be utilized in addition to human

[1] People ages 12 to 20 are categorized as "young people."

[2] Statistics from 2010.

[3] Statistic from the 2020-2021 school year.

judgement to predict the risk of excessive underage drinking for a given student. With access to Kaggle's "Student Alcohol Consumption" dataset [Lea], it was our goal to identify these key risk factors and subsequently build an interpretable, interactive tool that can be used by school counselors to help access alcohol over-consumption risk in students that potentially would have been overlooked in the current method.

Our Approach

As opposed to a "black box" solution, we decided to approach this problem from the lens of interpretable artificial intelligence– we not only want to utilize modern artificial intelligence tools to help us identify which students are at risk of alcohol over-consumption but we want to do so in a way such that we know what factors the algorithm is using to predict the likelihood of over-consumption. Through this we achieve two things: (i) From a machine learning perspective, we can avoid inherent bias in the classifier from the data if we can see what features are deemed important; and (ii) from a practical application perspective, we can now have more information sociologically about what may drive students to over consume alcohol. We hope that the interpretability of our solution will allow counselors to seamlessly interact with out model and use it as a tool to help them assess risk in students. To achieve this, we utilized existing subpackages from Interpretable AI on our Kaggle dataset. Below we detail the process in which we explored various learners, the results they produced, and which method we chose to use to create our tool.

Data Pre-Processing

The Kaggle "Student Alcohol Consumption" dataset contains information about 395 students across 33 features. Some examples of these features are sex, family relationship, and absences[4]. The data was compiled math classes from two Portuguese secondary schools. It is important to note the small scope of this dataset– there are not many student data points, and as such, any bias in the dataset is amplified and must be addressed in pre-processing in order for our classifier to generalize to populations beyond this data set. However, we believe that this dataset will allow us to create a framework that can subsequently be expanded upon when we are able to access a larger dataset.

We started our pre-processing by changing all binary variables to a 1 or 0 representation. We also modified "Mjob" and "Fjob" to represent employed (1) vs unemployed (0) as opposed to their job and modified "guardian" to be mother/father (1) vs other (0). Additionally, we made the choice to drop the "reason" features as we could not come up with a numerical representation of the feature and we were unsure how IAI handles one-hot encodings. It would be an area of future study to establish if this feature (which was reason the student chose their school) would change the output of the questionnaire.

[4]See Appendix 1.1 for complete list of features.

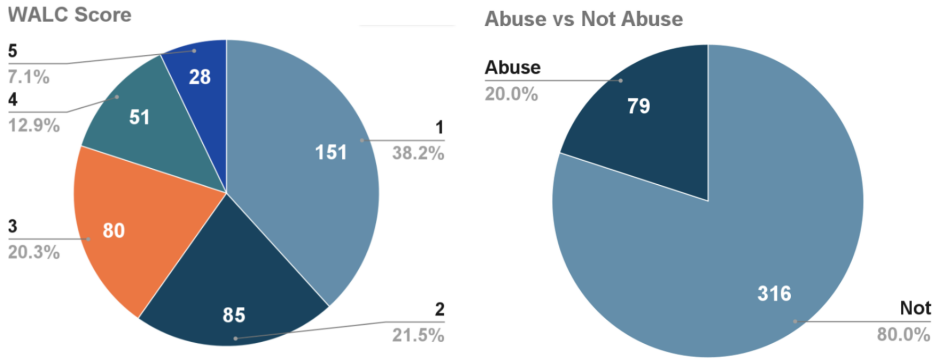


Figure 1: Students classified as abusing alcohol make up 20% of our data when categorizing a Walc score of 4 or 5 as abusive.

We are looking to predict a weekend alcohol score ("Walc") of a 4 or 5. Weekend alcohol score and weekday alcohol score ("Dalc") were both features in the dataset. We made the assumption that a student would be most likely to abuse alcohol on the weekend, and that any student abusing alcohol on the weekdays would be very likely to abuse on the weekends, as well. So, for the output of our model, we are looking at the "Walc" score. In order not to have any features involving alcohol detract from the other features, we made the decision to also drop "Dalc" from the features.

In regards to our output, we originally chose a score of 3, 4, or 5 to indicate a risk for over-consumption. However, as shown in Figure 1, a Walc score of 3, 4, or 5 made up nearly 50% of the dataset. This was suboptimal since we are trying to classify at-risk students and our classifier could just guess and produce a decent accuracy. As such, we changed the threshold to a 4 or 5 on the scale and this consisted of 20% of our data. So, we then changed "Walc" to be 1 if the score was a 4 or 5 and 0 otherwise. This was our target variable to predict.

Sex Bias in Data

It is important to note the discrepancy between males and females and how sex affects Walc score. We can see in Figure 2's breakdown of Weekend Alcohol score for both men and women that men are more likely to self-report excessive drinking than women. And, Figure 3 shows that not only are women equally as likely to binge drink than men, but in the United States, they are slightly more likely to abuse alcohol than men [AA]. This leads us to believe that women are less likely to self-report high Walc scores and thus this bias will affect how our model views the relationship between sex and alcohol abuse. Thus, as we explore potential models in the following sections, we look at the models performance on the data with and without sex.

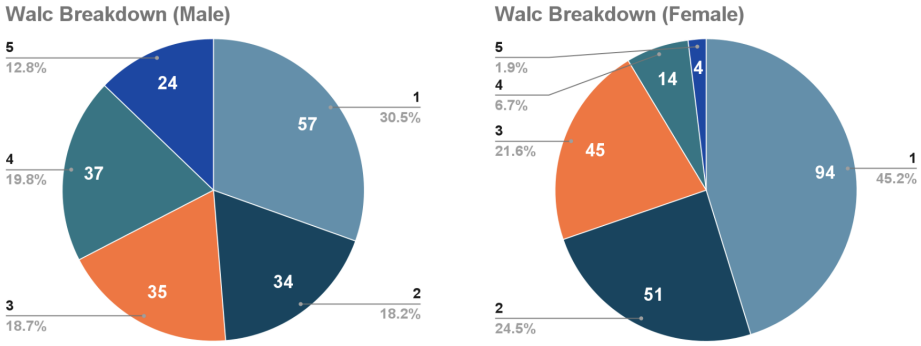


Figure 2: Substantially less women categorize themselves as having a higher Walc score than men.

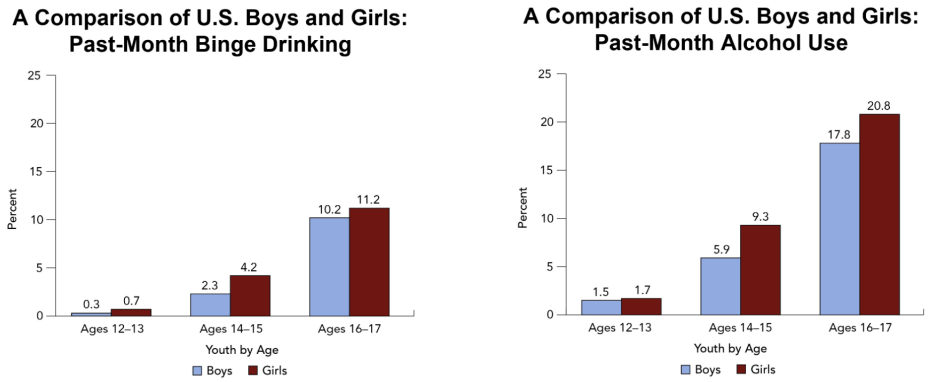


Figure 3: One study argued that females are slightly more likely to binge drink than males.[AA]

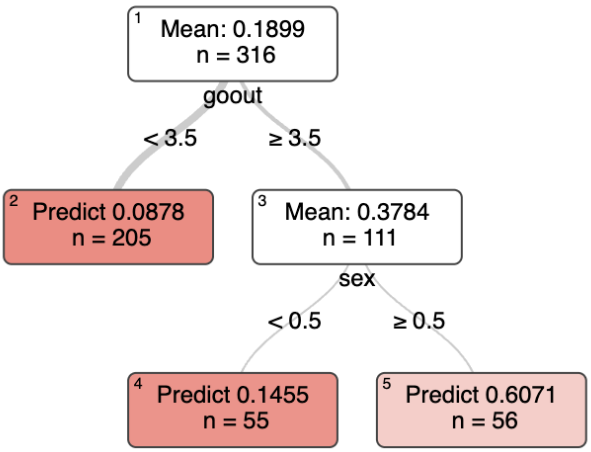
Models

Optimal Regression Tree

We began our analysis be exploring a linear regression optimal decision tree implementation to predict underage alcohol abuse probability. To do this, we used IAI's Optimal Regression Tree (ORT) learner [BD19]. The code we used can be found in the Appendix.

ORT with Sex as a Feature

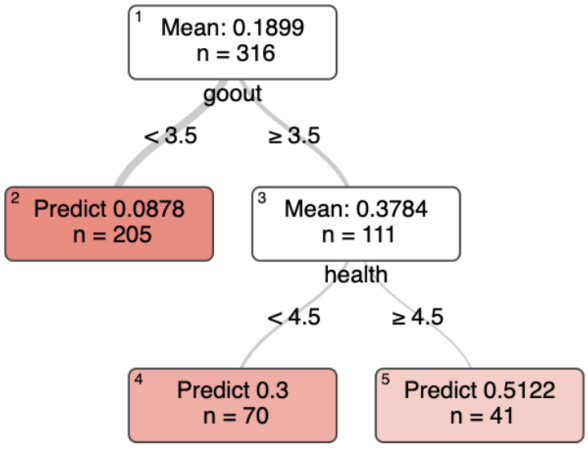
We first implemented the tree with sex as a feature. Below is the tree that was outputted.



The accuracy of this tree, as calculated by the mean square error, is .246 train accuracy and .380 test accuracy. The two features that the model deems important are the going out rating and sex of the student. Thus our model believes that men who go out frequently have a 60% chance of abusing alcohol while women who go out only have a 15% probability of abusing alcohol.

ORT without Sex as a Feature

Since we have already established the inherent bias in our dataset, we then explore the impact of removing sex as a feature from the data. Below is the tree that was outputted when removing sex from the feature set.



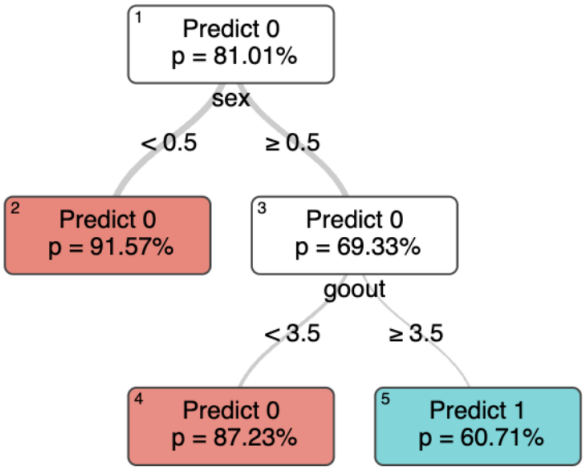
When we remove sex, our tree weights the majority of its regression calculation on the going out feature. The tree also splits on health status of the student. The accuracy, again as calculated by the mean square error, is .149 train accuracy and .216 test accuracy. This is a fairly low accuracy, and as such, we wanted to examine other methods for predicting risk of underage alcohol abuse.

Optimal Classification Tree

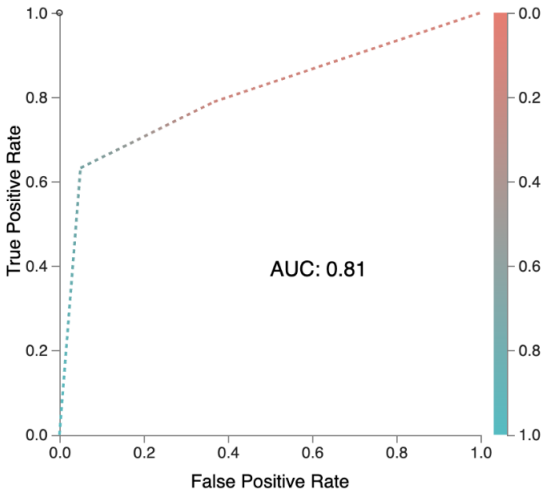
We then changed this problem to one of classification: whether a student will (1) or will not (0) abuse alcohol. We believe that this change of regression to classification will be beneficial for our problem statement. We can change the threshold of classification to customize our policy based on how much we prioritize false positives vs. false negatives. Additionally, classification would simplify the problem for counselors, alerting them when a student tests 'at risk' as opposed to outputting different probabilities of risk. To do this, we used IAI's Optimal Classification Tree (OCT) learner [BD17]. As before, the code can be found in the appendix.

OCT with Sex as a Feature

We ran our OCT on the dataset including the sex feature and received a very simple decision tree, as shown below, which utilized only 2 features: whether a student goes out and what gender that student is.



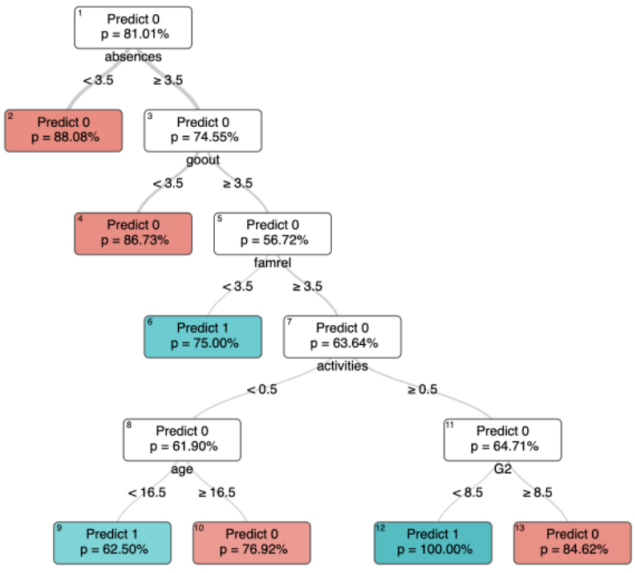
The train accuracy, as calculated by misclassification, for this model is .848 for training accuracy and .873 for testing accuracy. The AUC for this model is .81, which is misleadingly high.



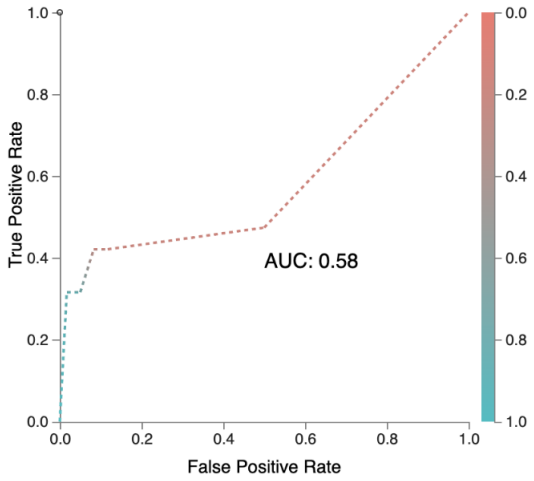
Upon first glace, an AUC of .81 and .873 accuracy makes us believe that this model is effective, but this could be attributed to the distribution of our data. 80% of our students are not categorized at risk while 20% are, so if the model classifies 100% of students as no risk, it will achieve a high accuracy of 80% while not actually providing insight into the problem. We know this model is not as accurate as it seems since, as we have stated previously, females are even more likely to abuse alcohol than males. Thus, a model that classifies every female as not at risk will not necessarily be accurate on a different dataset.

OCT without Sex as a Feature

To address this bias, we dropped sex as a feature in the tree. And, when we did this, the tree became much more informative regarding the other factors that may be important for underage drinking, as shown below.



Dropping sex as a feature makes the assumption that the other features indicative of underage alcohol abuse manifest the same in men as in women. We recognize that this may be an oversimplification, but we believe that for the purposes of this application it will suffice and result in a high accuracy on a larger testing set than provided. With this new representation, again calculated with misclassification, our train accuracy is .861 and our test accuracy is .797. In this optimal classification tree, we see many different features used in the tree to classify risk. Some key features are absences, family relationship, and grades. This makes intuitive sense and supports our hypothesis that objective features can be extracted from a student's demographics to aid in predicting risk.



It is worth noting that our AUC for this model (see above) is fairly low at .58. Further research will need to be performed to establish the optimal policy for finding the classification threshold for eliminating false positives while minimizing false negatives. It is impractical to have a policy that alerts the majority of students that they are at risk while it is also unhelpful if the majority of at-risk students are not alerted.

2 Results

Model	Train Acc	Test Acc	AUC
ORT w/ Sex	.246	.380	N/A
ORT w/o Sex	.149	.216	N/A
OCT w/ Sex	.848	.873	.81
OCT w/o Sex	.861	.797	.58

Despite not necessarily producing the highest accuracy with this given dataset, we believe the final model (OCT without sex) will produce the highest accuracy on other datasets. As such, we used this model to create a tool to help school counselors assess a given student's alcohol abuse risk. This tool is a questionnaire, built using IAI's write questionnaire function, that

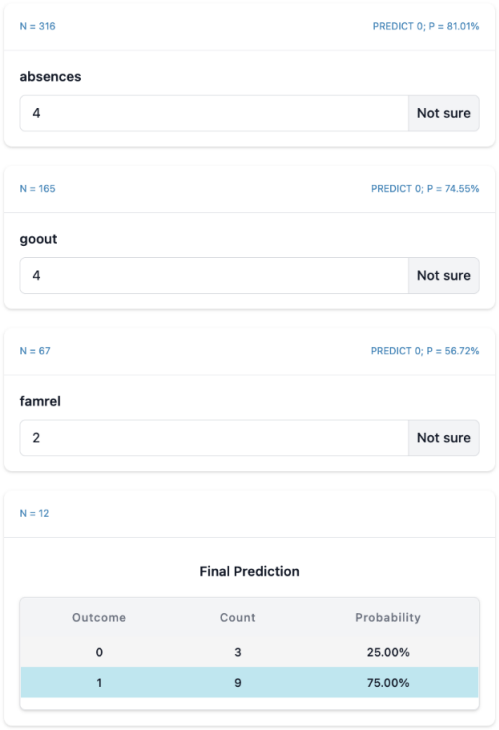


Figure 4: Our tool predicts this student to have a 75% risk of alcohol abuse.

prompts the counselor to input the significant variables required, from our chosen decision tree, to predict a student's risk. This allows the counselor to have an interpretable tool to aid his/her support process and hopefully identify student's otherwise not considered.

Conclusion and Future Work

There are many directions in which this project can be expanded for optimal societal impact. First, expanding the dataset beyond the 395 students in Portugal would improve robustness of the model to other student populations. Additionally, there are other demographic features that can be included into the model to improve accuracy, such as a parent's relationship with alcohol.

Another area to explore is the accuracy of a similar classification model implemented with a black box neural network. This would be helpful for us to directly compare if the interpretability of the IAI model affects the accuracy of our classification.

Finally, to increase accessibility for students and educators, we would like to automate our questionnaire. A lot of the social demographic information required for the decision tree is filled out routinely by students and school administrators. Absences, grades, and other academic features are automatically gathered from the school database. As such, risk can automatically be calculated and students can be flagged without any additional work on behalf of the administrators.

Appendix

2.1 Kaggle Student Alcohol Consumption Dataset

variable	description	representation
school	student's school	'GP' or 'MS'
sex	student's sex	'F' or 'M'
age	student's age	15 to 22
address	student's home address type	'U' - urban or 'R' - rural
famsize	family size	'LE3' - ≤ 3 or 'GT3' - ≥ 3)
Pstatus	parent's cohabitation status	'T' - together or 'A' - apart
Medu	mother's education	0 to 4 (most)
Fedu	father's education	0 to 4 (most)
Mjob	mother's job	'teacher', 'health', 'services', 'at home' or 'other'
Fjob	father's job	'teacher', 'health', 'services', 'at home' or 'other'
reason	reason to choose this school	close to 'home', 'reputation', 'course' or 'other')
guardian	student's guardian	'mother', 'father' or 'other'
traveltime	home to school travel time	1 to 4 (most)
studytime	weekly study time	1 to 4 (most)
failures	number of past class failures	1,2,3, or 4+
schoolsup	extra educational support	yes or no
famsup	family educational support	yes or no
paid	extra paid classes within the Math	yes or no
activities	extra-curricular activities	yes or no
nursery	attended nursery school	yes or no
higher	wants to take higher education	yes or no
internet	Internet access at home	yes or no
romantic	with a romantic relationship	yes or no
famrel	quality of family relationships	1 (very bed) to 5 (excellent)
freetime	free time after school	1 (very low) to 5 (very high)
goout	going out with friends	1 (very low) to 5 (very high)
Dalc	workday alcohol consumption	1 (very low) to 5 (very high)
Walc	weekend alcohol consumption	1 (very low) to 5 (very high)
health	current health status	1 (very low) to 5 (very high)
absences	number of school absences	0 to 93
G1	first period grade	0 to 20
G2	second period grade	0 to 20
G3	final grade	0 to 20

2.2 ORT Code with Sex as a Feature

```
X = df[:, 1:(end - 1)]
y = df[:, end]
(train_X, train_y), (test_X, test_y) = IAI.split_data(:regression, X, y, seed
    =1, train_proportion =0.8)

grid = IAI.GridSearch(IAI.OptimalTreeRegressor(random_seed=1,),max_depth
    =1:30,minbucket=5:60)
IAI.fit!(grid, train_X, train_y)
IAI.get_learner(grid)

IAI.variable_importance(IAI.get_learner(grid))

println("MSE Train error: ", IAI.score(grid, train_X, train_y, criterion =:
    mse))
println("NMSE Test error: ", IAI.score(grid, test_X, test_y, criterion =:mse)
    )
```

2.3 ORT Code without Sex as a Feature

```
df = df[:, Not([:sex])]
X = df[:, 1:(end - 1)]
y = df[:, end]
(train_X, train_y), (test_X, test_y) = IAI.split_data(:regression, X, y, seed
    =1, train_proportion =0.8)

grid = IAI.GridSearch(IAI.OptimalTreeRegressor(random_seed=1,),max_depth
    =1:30,minbucket=5:60)
IAI.fit!(grid, train_X, train_y)
IAI.get_learner(grid)

IAI.variable_importance(IAI.get_learner(grid))

println("MSE Train error: ", IAI.score(grid, train_X, train_y, criterion =:
    mse))
println("NMSE Test error: ", IAI.score(grid, test_X, test_y, criterion =:mse)
    )
```

2.4 ORT Code with Sex as a Feature

```
X = df[:, 1:(end - 1)]
y = df[:, end]
(train_X, train_y), (test_X, test_y) = IAI.split_data(:regression, X, y, seed
    =1, train_proportion =0.8)

grid = IAI.GridSearch(IAI.OptimalTreeClassifier(random_seed=1,),max_depth
    =1:30,minbucket=5:60)
IAI.fit!(grid, train_X, train_y)
IAI.get_learner(grid)

IAI.variable_importance(IAI.get_learner(grid))

println("MSE Train error: ", IAI.score(grid, train_X, train_y, criterion =:
    misclassification))
println("NMSE Test error: ", IAI.score(grid, test_X, test_y, criterion =:
    misclassification))
```

2.5 ORT Code without Sex as a Feature

```
df = df[:, Not([:sex])]
X = df[:, 1:(end - 1)]
y = df[:, end]
(train_X, train_y), (test_X, test_y) = IAI.split_data(:regression, X, y, seed
    =1, train_proportion =0.8)

grid = IAI.GridSearch(IAI.OptimalTreeClassifier(random_seed=1,),max_depth
    =1:30,minbucket=5:60)
IAI.fit!(grid, train_X, train_y)
IAI.get_learner(grid)

IAI.variable_importance(IAI.get_learner(grid))

println("MSE Train error: ", IAI.score(grid, train_X, train_y, criterion =:
    misclassification))
println("NMSE Test error: ", IAI.score(grid, test_X, test_y, criterion =:
    misclassification))
```

References

[BD17] Dimitris Bertsimas and Jack Dunn. “Optimal classification trees”. In: *Machine Learning* 106.7 (2017), pp. 1039–1082.

[BD19] Dimitris Bertsimas and Jack Dunn. *Machine learning under a modern optimization lens*. Dynamic Ideas LLC, 2019.

[AA] National Institute on Alcohol Abuse and Alcoholism. *Underage Drinking*. URL: https://www.niaaa.nih.gov/publications/brochures-and-fact-sheets/underage-drinking. (accessed: 05.09.2022).

[Ass] American School Counselor Association. *School Counselor Roles Ratios*. URL: https://www.schoolcounselor.org/About-School-Counseling/School-Counselor-Roles-Ratios. (accessed: 05.09.2022).

[DP] Centers for Disease Control and Prevention. *Underage Drinking*. URL: https://www.cdc.gov/alcohol/fact-sheets/underage-drinking. (accessed: 05.09.2022).

[Lea] UCI Machine Learning. *Student Alcohol Consumption*. URL: https://www.kaggle.com/datasets/uciml/student-alcohol-consumption?resource=download. (accessed: 05.09.2022).