CSC105M Final Project

Fernandez, Ryan Austin

Poblete, Clarisse Felicia M.

Dataset Description

- Student alcohol consumption dataset
- UCI machine learning repository
- 650 instances

Attr #	Attribute
1	school
2	sex
3	age
4	address
5	famsize
6	Pstatus
7	Medu
8	Fedu
9	Mjob
10	Fjob
11	reason

Attr #	Attribute
12	guardian
13	traveltime
14	studytime
15	failures
16	schoolsup
17	famsup
18	paid
19	activities
20	nursery
21	higher
22	internet

Attr #	Attribute
23	romantic
24	famrel
25	freetime
26	gout
27	Dalc
28	Walc
29	health
30	absences
31	G1
31	G2
32	G3

Data Preprocessing

- No missing values
- Discrete
- Normalization for Regression and Neural Networks

Data Preprocessing

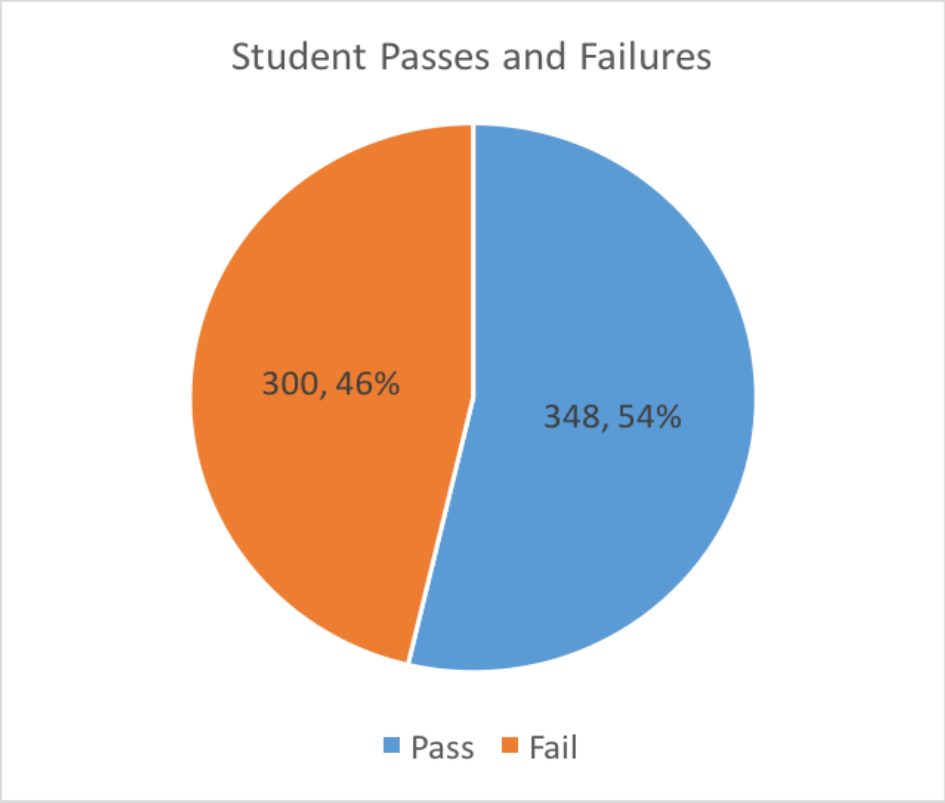
- Binary: 1 or 0
- Nominal: n values -> n – 1 attributes
- Ordinal: 1 to n
- Final Grade:
  - If >= 12, Pass
  - Else, Fail
- Min/Max Standardization

Feature Selection

- Regression
  - Multicollinearity checks
  - Low correlation coefficients across the board
- Decision Trees
  - C4.5 Algorithm prunes the features
- Neural Networks
  - Neural Networks are robust to noise

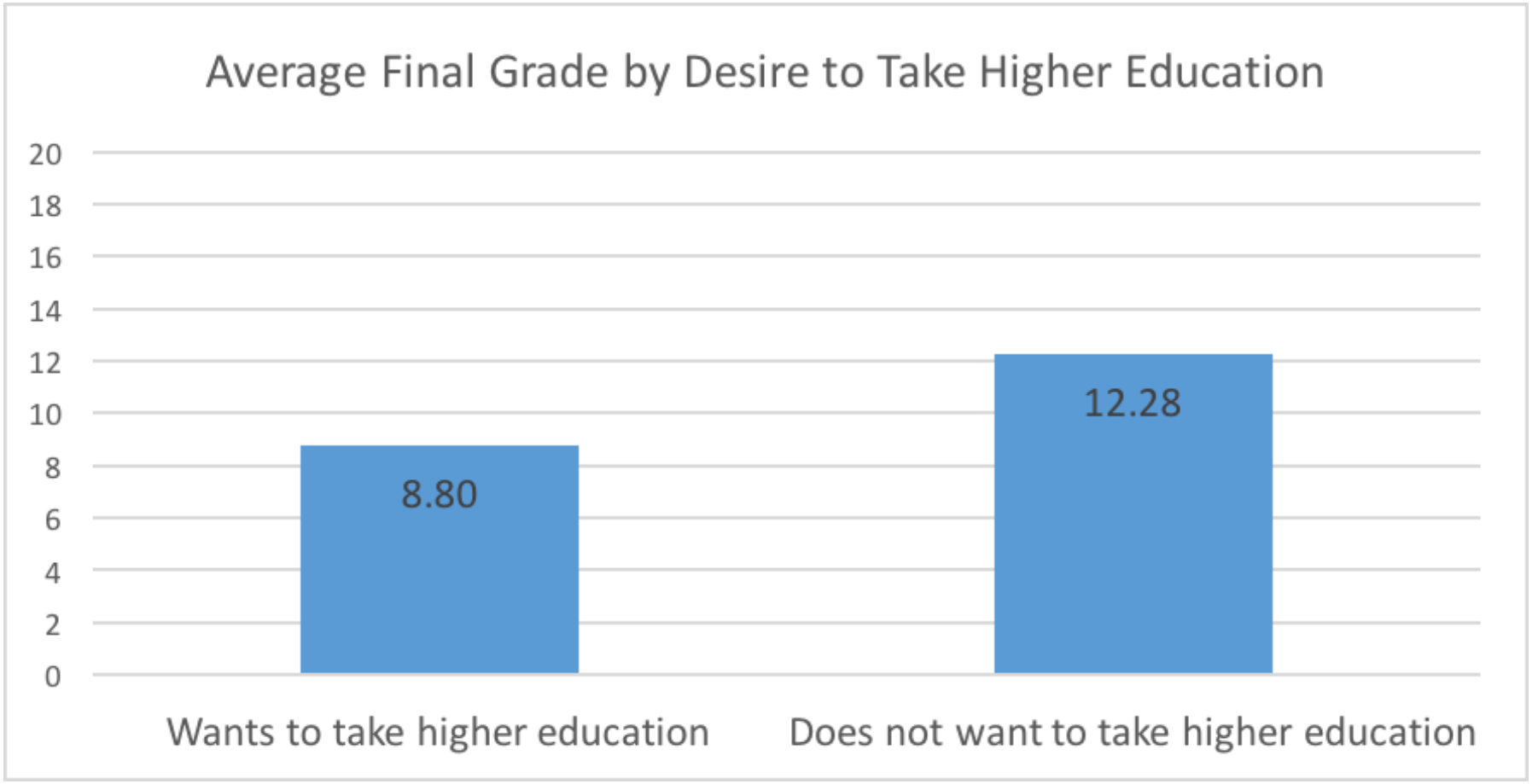
Visualization

Passes and Failures



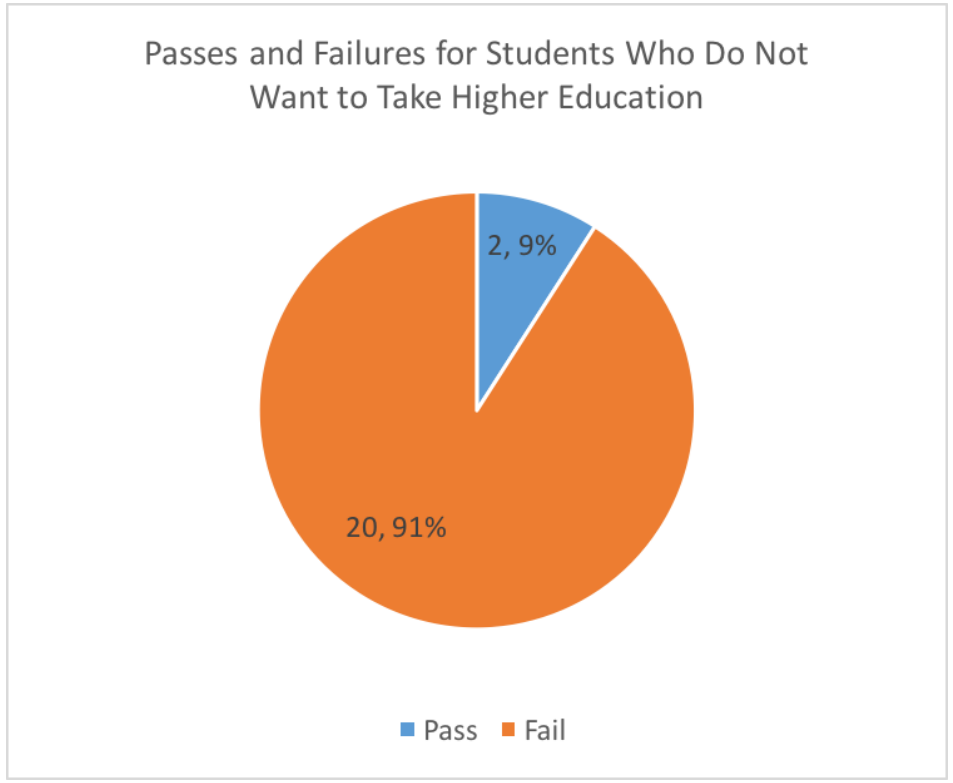
Visualization

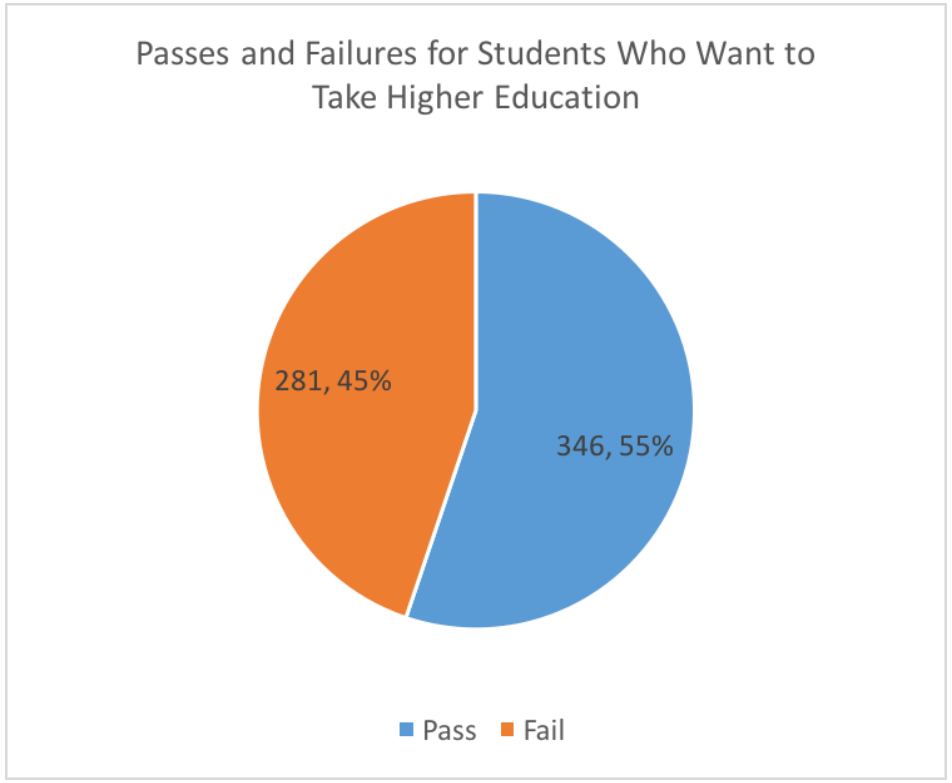
Desire to Take Higher Education



Visualization

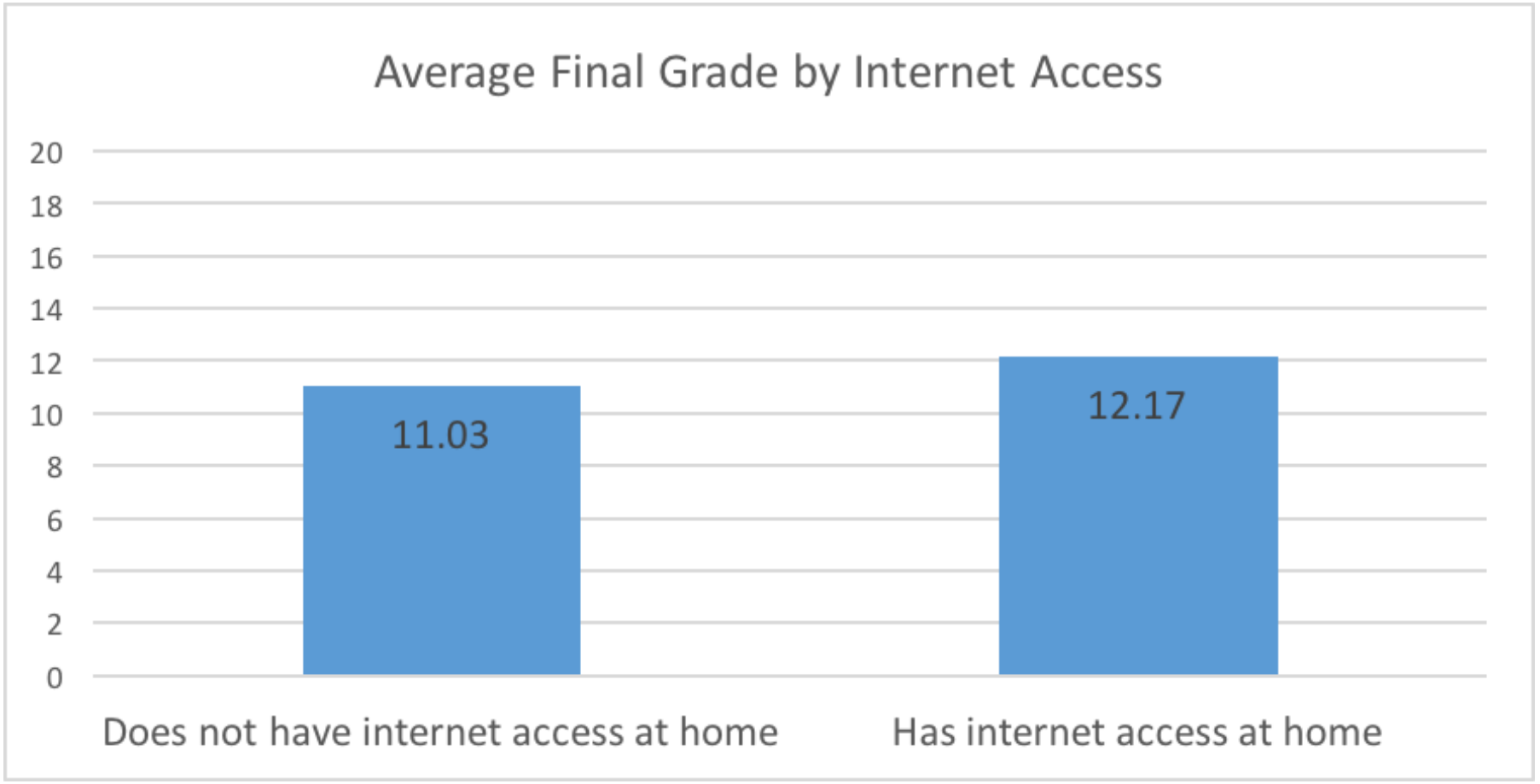
Desire to Take Higher Education





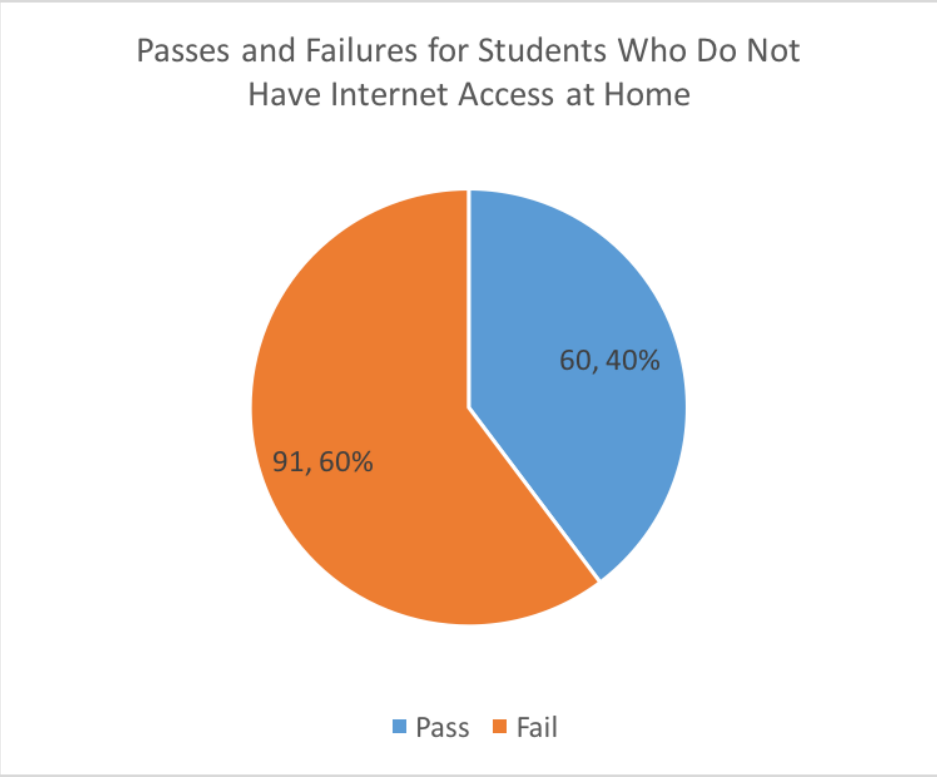
Visualization

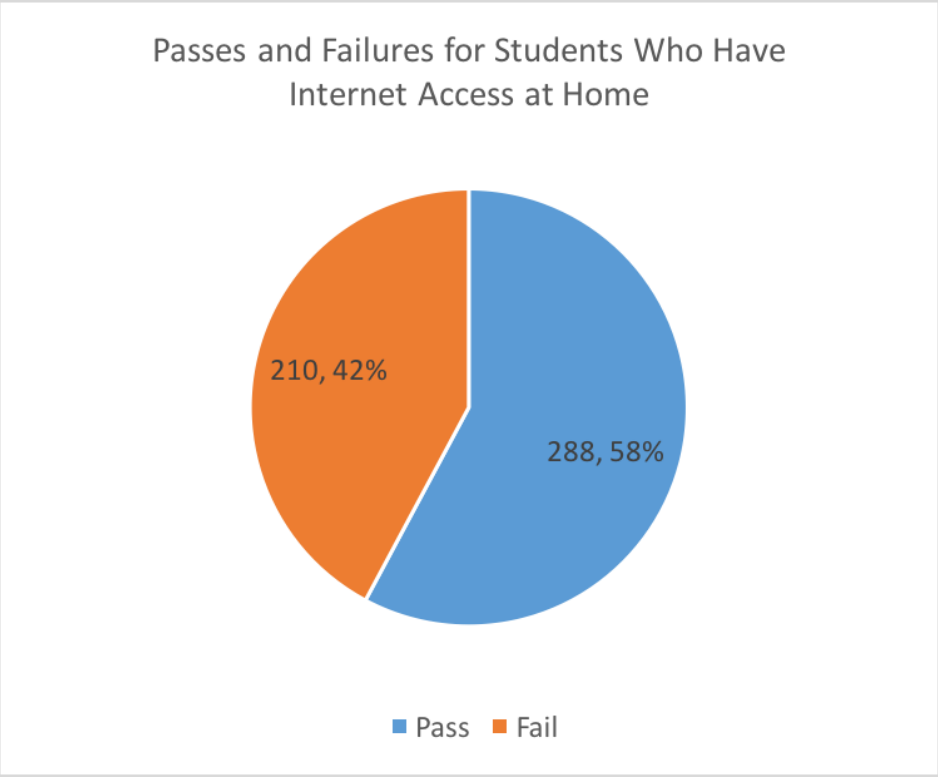
Internet Access



Visualization

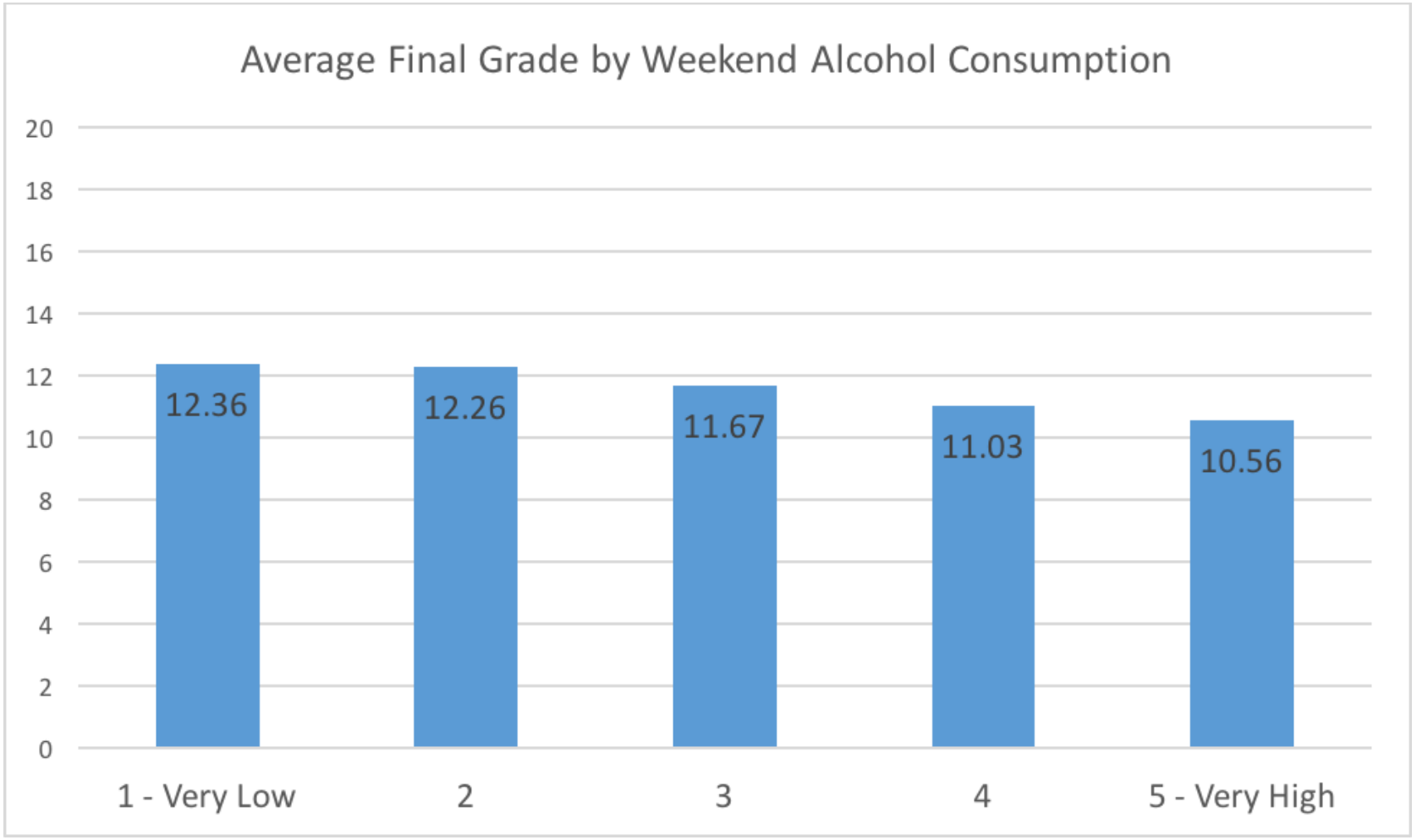
Internet Access





Visualization

Weekend Alcohol Consumption



Analytics

- Predict Pass or Fail
- Three techniques
  - Regression
  - Decision Trees
  - Neural Networks
- Bootstrap Aggregating
  - 80% of dataset with replacement

Regression

- Low correlations deem this unsuitable for the dataset

Decision Trees

- C4.5 Algorithm
- Using J48 Implementation in Weka
- Parse Decision Trees using Java
- Bagging via voting scheme

Neural Network

- Custom Implementation of Backpropagation Algorithm
- Sigmoid Hidden Layer and Output Layer neurons
- One Output neuron for Pass, one for fail
- If Pass, Pass Neuron's target is 0.9, Fail Neuron's is 0.1
- If Fail, Pass Neuron's target is 0.1, Fail Neuron's is 0.9

Interpretations, Findings, and Conclusions

Decision Tree

Actual/Prediction	Pass	Fail
Pass	333	15
Fail	50	251

Interpretations, Findings, and Conclusions

Decision Tree

- Classification Accuracy: 89.9846%
- Classification Error: 10.0154%
- Sensitivity: 95.6897%
- Specificity: 83.3887%

Interpretations, Findings, and Conclusions

Decision Tree

Correct Predictions	Wrong Predictions	Rule
158	31	failures = 0 ^ higher = yes ^ Mjob != home ^ Walc <= 3 ^ schoolsup = no ^ school = GP ^ internet = yes ^ age <= 18 -> Pass
139	24	higher = yes ^ failures = 0 ^ school = GP ^ nursery = yes ^ internet = yes ^ schoolsup = no ^ Dalc <= 1 -> Pass
110	24	failures = 0 ^ higher = yes ^ Mjob != home ^ Dalc <= 2 ^ Fjob != teach ^ absences <= 3 ^ health <= 4 -> Pass
88	5	failures > 0 ^ age <= 19 -> Fail
88	3	failures > 0 ^ Medu <= 3 ^ Fedu > 0 -> Fail

Interpretations, Findings, and Conclusions

Neural Networks

Actual/Prediction	Pass	Fail
Pass	329	19
Fail	27	274

Interpretations, Findings, and Conclusions

Neural Networks

- Classification Accuracy: 92.9122%
- Classification Error: 7.0878%
- Sensitivity: 94.5402%
- Specificity: 91.0299%

Interpretations, Findings, and Conclusions

Neural Networks

- Possible overfitting in NN
- Neural Networks performed better
- Success in building an analytic model
- Possible use of SVM in future studies

References

Amran, H. & Pagnotta, F. (2016). Using Data Mining to Predict Secondary School Alcohol Consumption. *University of Camerino.* doi: 10.13140/RG.2.1.1465.8328

Baesens, B. (2014). *Analytics in a Big Data World: The Essential Guide to Data Science and its Applications.* NJ: John Wiley & Sons.

Cortez, P. & Silva, A. (2008) Using Data Mining to Predict Secondary School Student Performance. In A. Brito and J. Teixeira Eds., Proceedings of 5th Future Business Technology Conference pp. 5-12, Porto, Portugal, EUROSIS, ISBN 978-9077381-39-7.

Mitchell, T. (1997). *Machine learning*. McGraw-Hill.

Stockburger, D.W. (n.d.) Multiple Regression With Categorical Variables. Retrieved July 27, 2016, from Psychological Statistics at Missouri State University: http://www.psychstat.missouristate.edu/multibook/mlt08m.html

Thank you!