CSC105M Final Project

Fernandez, Ryan Austin

Poblete, Clarisse Felicia M.

Dataset Description

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

Attr #	Attribute	Description
1	school	student's school (binary: 'GP' – Gabriel Pereira or 'MS' – Mousinho da Silveira)
2	sex	student's sex (binary: 'F' – female or 'M' – male)
3	age	student's age (numeric: from 15 to 22)
4	address	student's home address type (binary: 'U' – urban or 'R' – rural)
5	famsize	family size (binary: 'LE3' – less or equal to 3 or 'GT3' – greater than 3)
6	Pstatus	parent's cohabitation status (binary: 'T' – living together or 'A' – apart)
7	Medu	mother's education (numeric: 0 – none, 1 – primary education (4th grade), 2 5th to 9th grade, 3 secondary education or 4 higher education)
8	Fedu	father's education (numeric: 0 – none, 1 – primary education (4th grade), 2 5th to 9th grade, 3 secondary education or 4 higher education)
9	Mjob	mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
10	Fjob	father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')

Attr #	Attribute	Description
11	reason	reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other')
12	guardian	student's guardian (nominal: 'mother', 'father' or 'other')
13	traveltime	home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)
14	studytime	weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
15	failures	number of past class failures (numeric: n if 1<=n<3, else 4)
16	schoolsup	extra educational support (binary: yes or no)
17	famsup	family educational support (binary: yes or no)
18	paid	extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)
19	activities	extra-curricular activities (binary: yes or no)
20	nursery	attended nursery school (binary: yes or no)
21	higher	wants to take higher education (binary: yes or no)

Attr #	Attribute	Description
22	internet	Internet access at home (binary: yes or no)
23	romantic	with a romantic relationship (binary: yes or no)
24	famrel	quality of family relationships (numeric: from 1 – very bad to 5 – excellent)
25	freetime	free time after school (numeric: from 1 – very low to 5 – very high)
26	gout	going out with friends (numeric: from 1 – very low to 5 – very high)
27	Dalc	workday alcohol consumption (numeric: from 1 – very low to 5 – very high)
28	Walc	weekend alcohol consumption (numeric: from 1 – very low to 5 – very high)
29	health	current health status (numeric: from 1 – very bad to 5 – very good)
30	absences	number of school absences (numeric: from 0 to 93)
31	G1	first period grade (numeric: from 0 to 20)
31	G2	second period grade (numeric: from 0 to 20)
32	G3	final grade (numeric: from 0 to 20, output target)

Data Preprocessing

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

Data Preprocessing

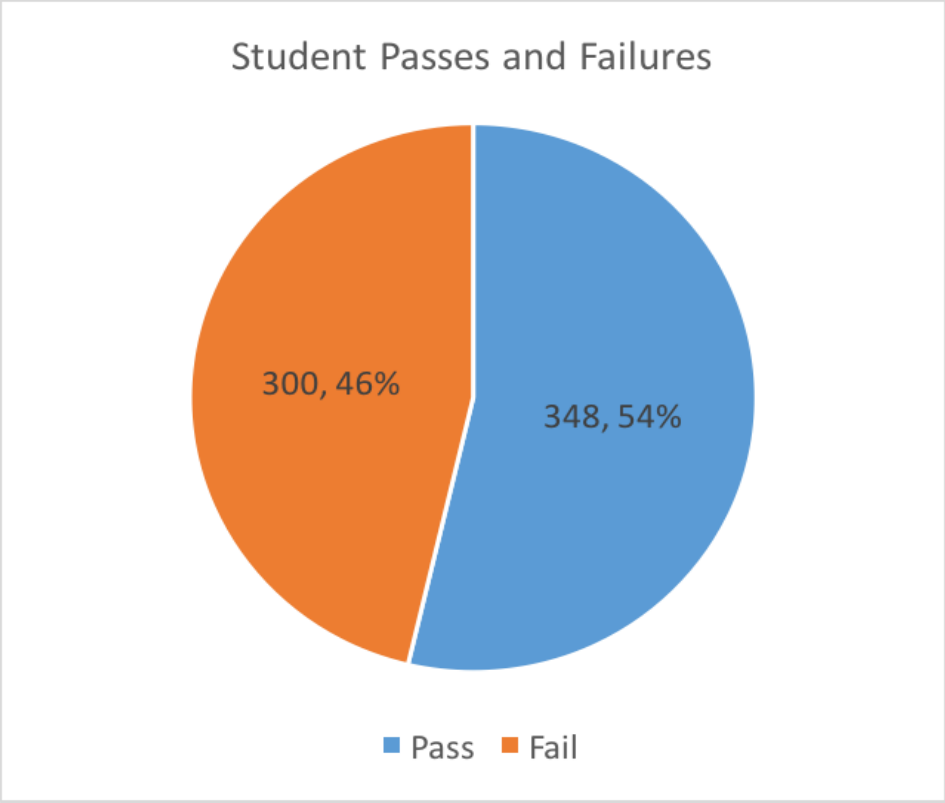
- Binary: 1 or 0
- Nominal: n values -> n – 1 attributes
- Ordinal: 1 to n
- Final Grade:
  - If >= 12, Pass
  - Else, Fail
- Min/Max Standardization

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

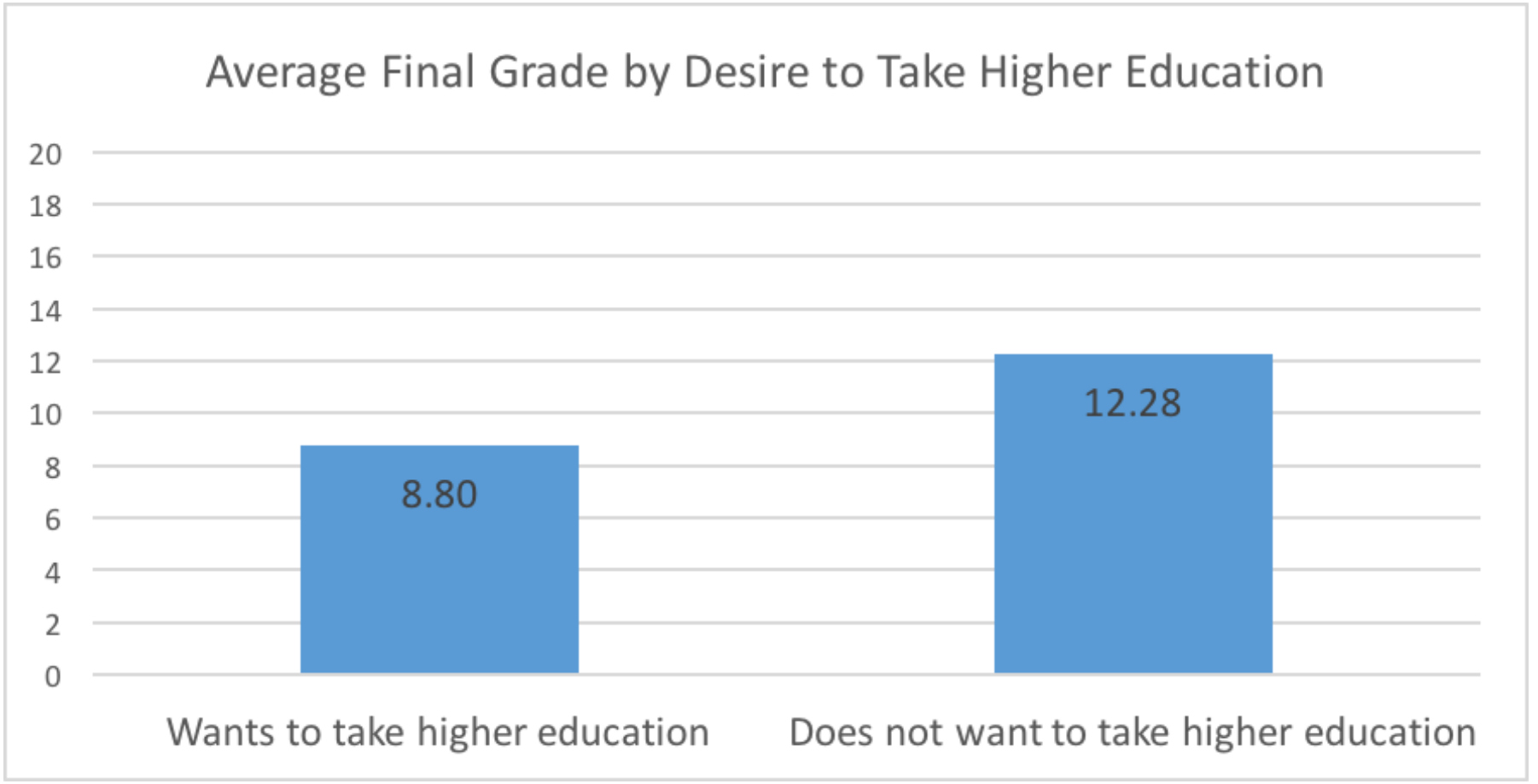
Visualization

Passes and Failures



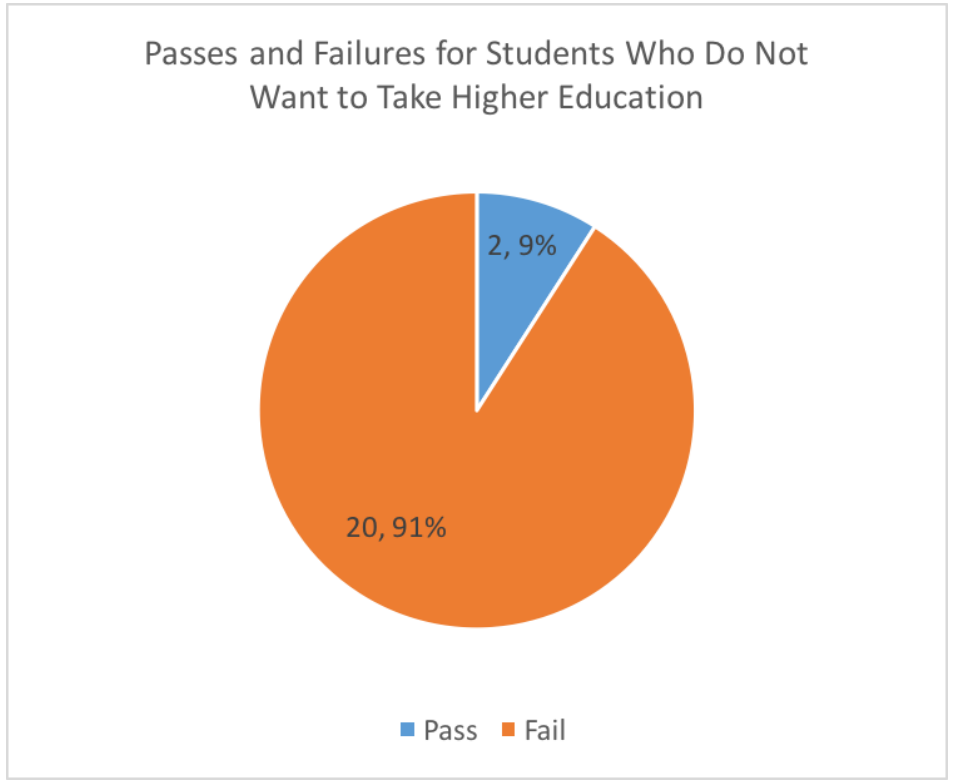
Visualization

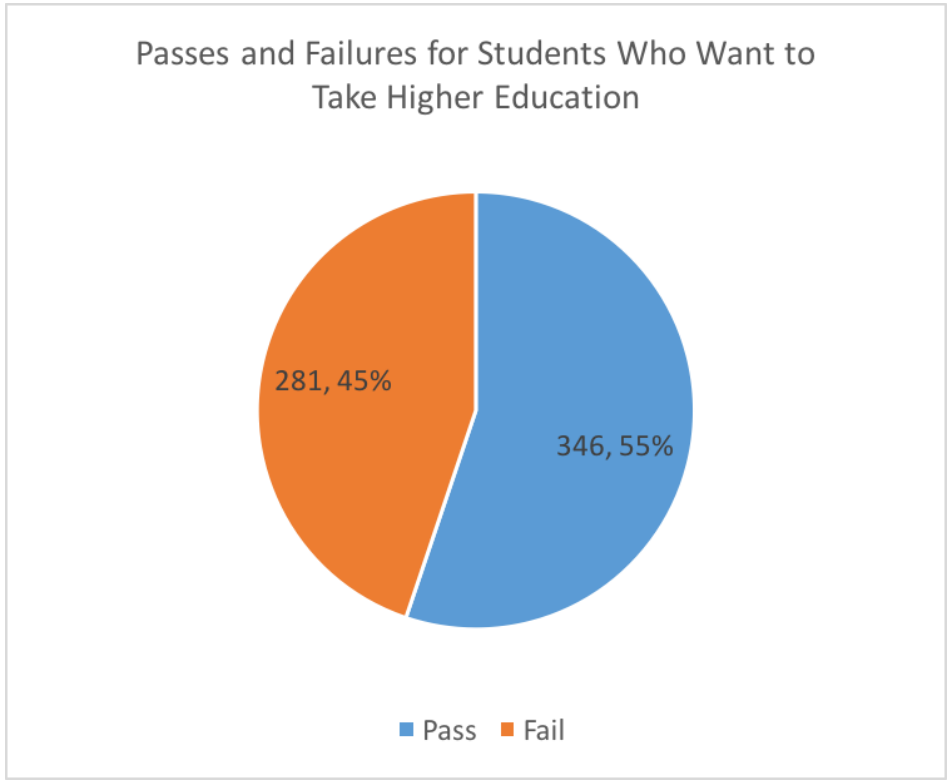
Desire to Take Higher Education



Visualization

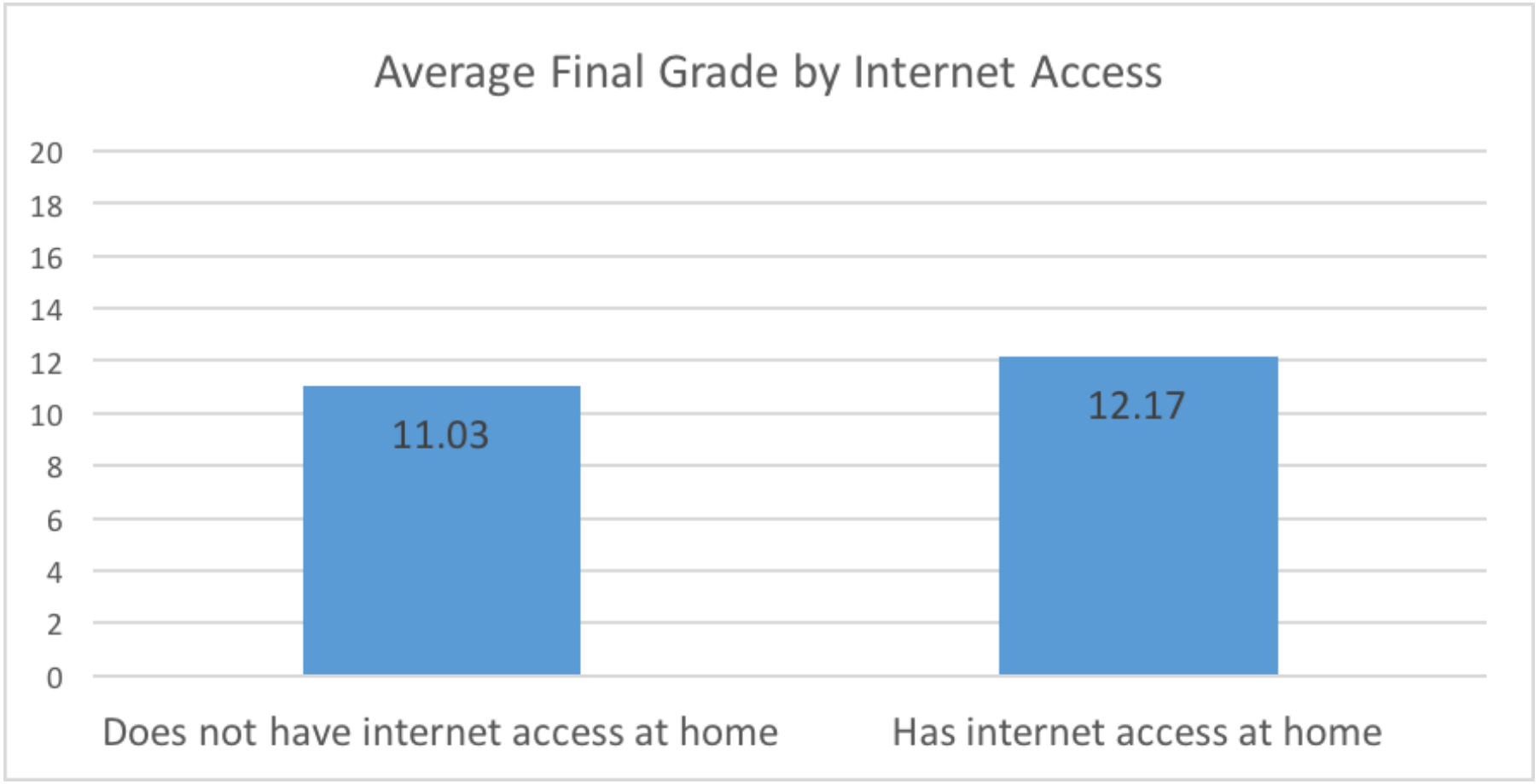
Desire to Take Higher Education





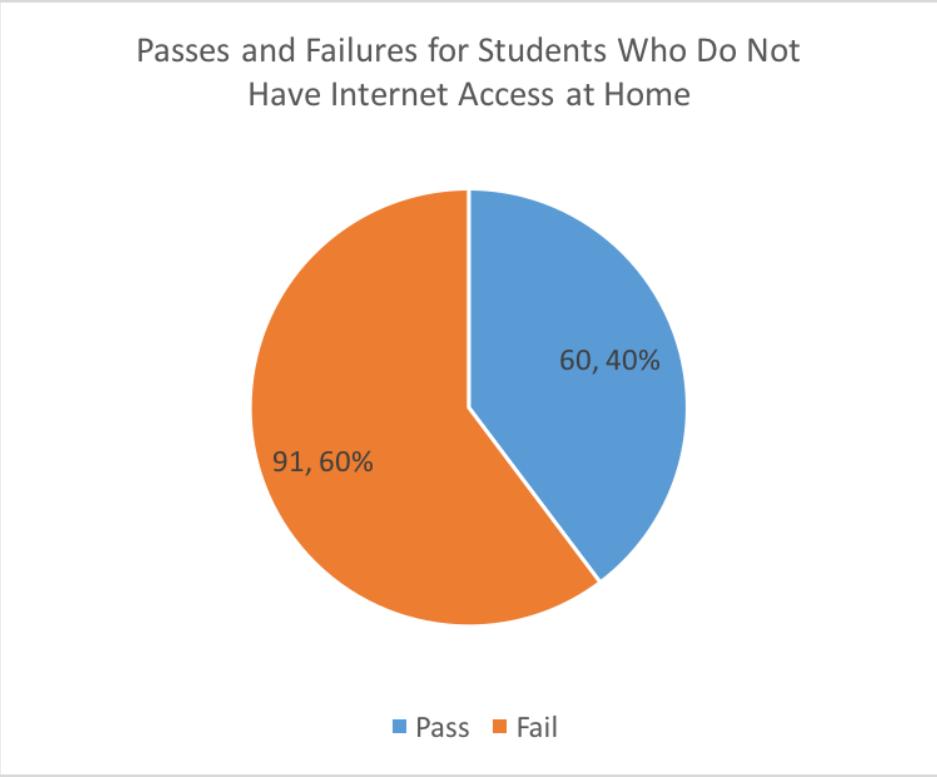
Visualization

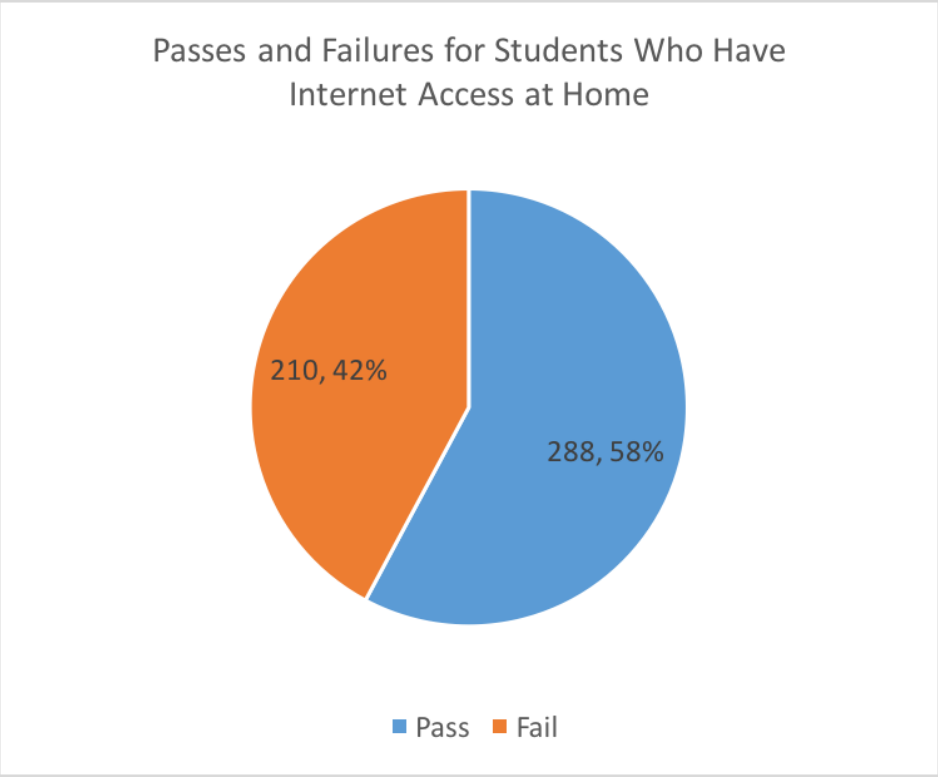
Internet Access



Visualization

Internet Access





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