Alcohol Consumption and Absences Effect on Student Failures

Gabriel Solomon Holland

Nathan Eloe, NWMSU

NORTHWEST MISSOURI STATE UNIVERSITY

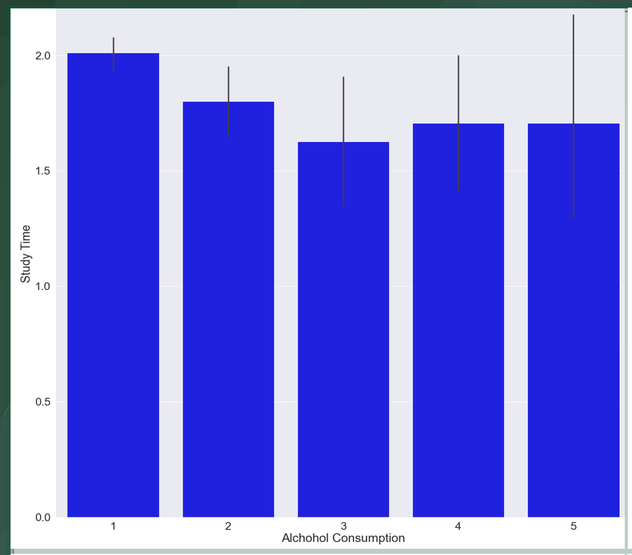
Introduction

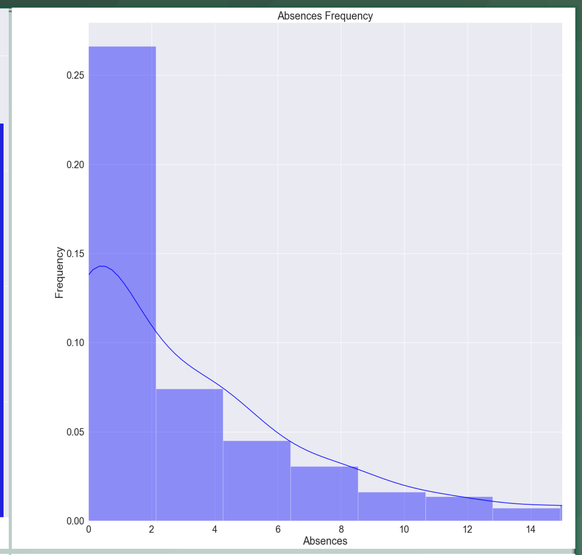
- Dataset was a collected at a High School in Portugal
- Goal was to find potential causes of a student's success or failure
- Target features were alcohol consumption and absences
- One foreseen issue: many of the features were self reported scales 1-5 which result in inconsistent data

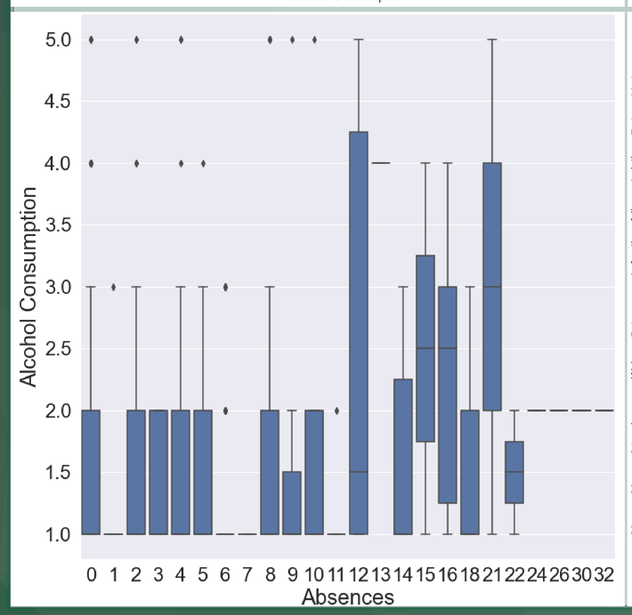
Materials/Methods

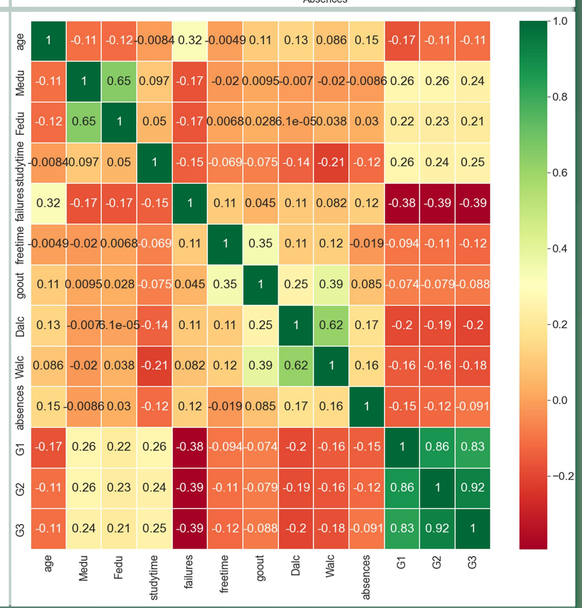
- Jupyter Notebook
- Matplotlib Python Library
- Seaborn Python Library
- Scikit-Learn Python Library
- Kaggle.com

Results









Conclusion

- Barely any trends/correlations in the dataset
- Initial hypothesis disproven in absences vs alcohol consumption box plot, across higher levels of absences alcohol consumption becomes more diverse but not bigger.
- Correlation matrix shows that there are no immediate correlations between failures and any features
- Machine learning algorithms achieved an 85% accuracy in predicting failure regardless of features used. Upon further analysis the model guessed 0-2 failures every single time, which made up ~85% of the dataset. It guessed the statistically most likely answer every time.
- I was unable to successfully predict a failure rate of students using this dataset.
- Possible reason for failure: self reported data, certain numbers could be exaggerated or understated. The scale of numbers was not clear

Sources

Scikit-Learn: https://scikit-learn.org/stable/
Seaborn: https://seaborn.pydata.org/
Matplotlib: https://matplotlib.org/
Kaggle: https://www.kaggle.com/uciml/student-alcohol-consumption

Acknowledgements

Special thanks to Nathan Eloe

Further Information

https://github.com/GabrielSolomonHolland/data-science-final