Risk and Protective Factors of Student Alcohol Consumption

Andreea G., Nikki N., Rishav R., Mridula S., Alexander T.

Introduction

- Data: UCI ML data on Student Alcohol Consumption
    - Portuguese secondary school
    - Size: 423 students
    - Age: 15-18+
    - 30 original features

- Included several demographic and behavioral variables

- Response variable: Weekend alcohol consumption (Walc)

Relevant Variables

`goout`

Going out with friends?
(Likert, 1-5)

`famrel`

Quality of family relationships?
(Likert, 1-5)

`female`

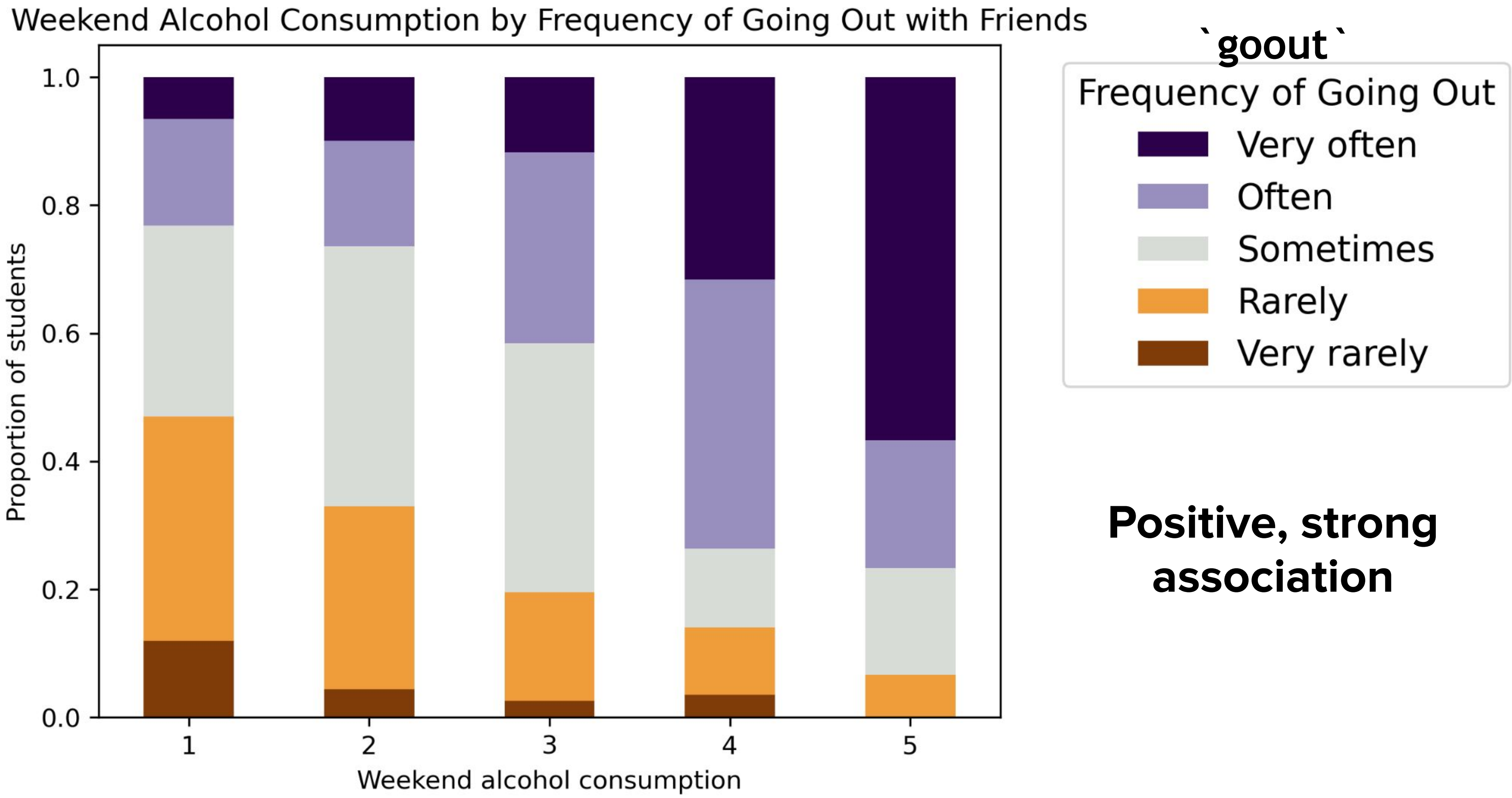
Female?
(binary, 1/0)

`Fjob_services`

Father in service industry?
(binary, 1/0)

`activities`

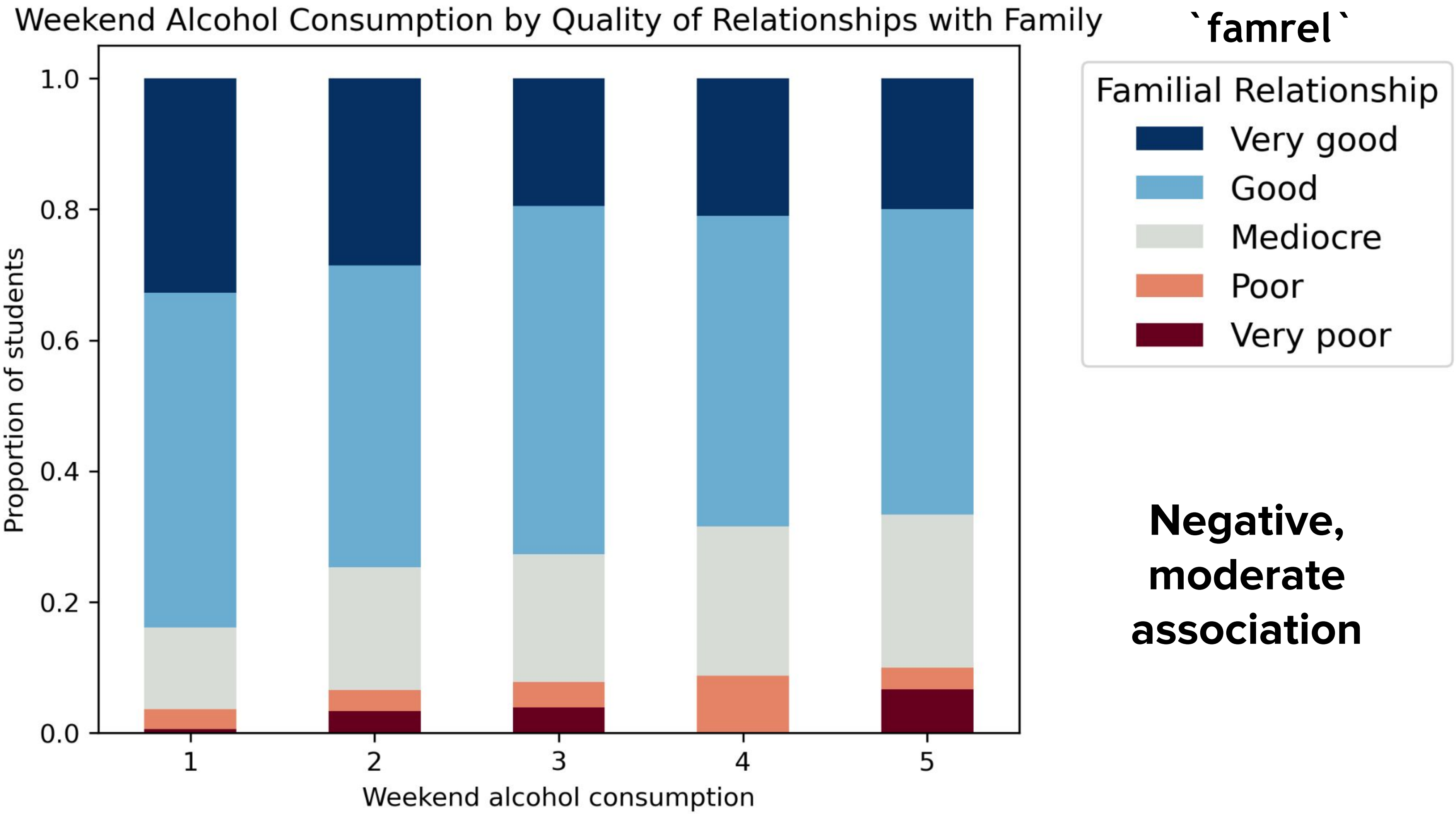
Extracurricular activities?
(binary, 1/0)



`goout`

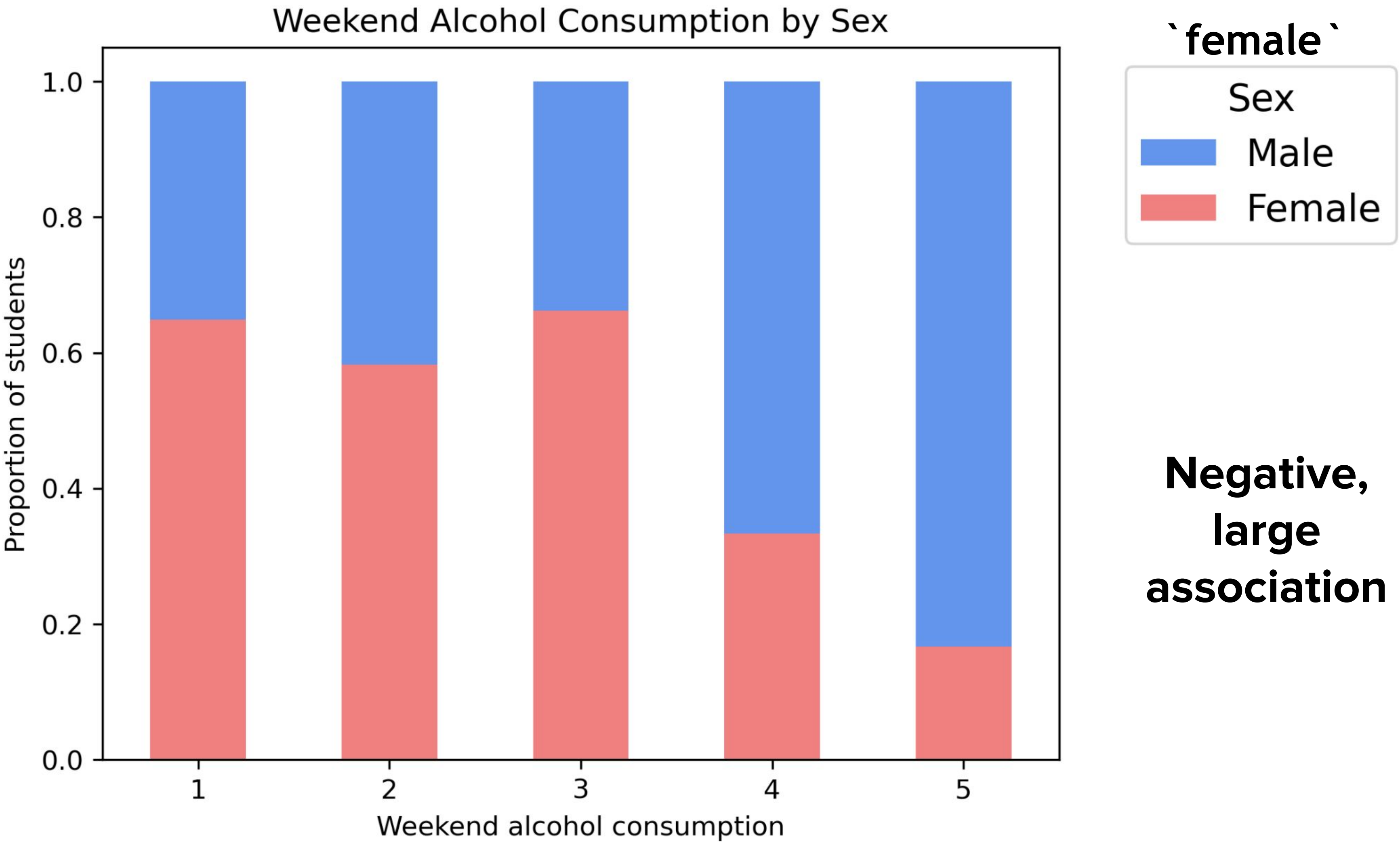
Positive, strong association

`famrel`



Negative, moderate association

`female`



Negative, large association

Feature	Hypothesized Importance	Hypothesized Effect	Calculated Importance	Calculated Effect
`goout`	Very high	Positive, large		
`famrel`	High	Negative, moderate		
`female`	Very high	Negative, large		

Data Format and Processing

- Binary variables reduced to 0, 1
- Parent job variables were one-hot encoded
- Likert scale responses and some categorical variables in ordinal form (1-5) but not linear
    - Ex: parent education and weekend alcohol consumption (Walc)

Feature Selection using Lasso

Lasso[5] regression uses L1 regularization, finding $\min_{\mathbf{w},b}\left(\frac{1}{N}\sum_{i=1}^{N}\left[f_{\mathbf{w},b}(\mathbf{x}_i)-y_i\right]^2+\alpha|\mathbf{w}|\right)$ where $\alpha \geq 0$.[6]

Lasso with α = 1:

```
array([ 0., -0.,  0.,  0., -0.,  0., -0., -0.,  0., -0., -0., -0.,  0.,
        0., -0.,  0.,  0.,  0.,  0., -0.,  0., -0., -0.,  0.])
```

Lasso with α = 0.01:

```
array([ 0.01209795, -0.01994296, -0.        ,  0.01600177, -0.09266678,
        0.        , -0.04173893, -0.        ,  0.        ,  0.        ,
       -0.00952809,  0.        ,  0.        ,  0.        , -0.0720974 ,
       -0.05019267,  0.14489994,  0.02028593,  0.0083867 , -0.25878453,
        0.        ,  0.        , -0.        ,  2.10094054])
```

Lasso regression requires linear data

- weekend alcohol consumption uses Likert scale
  - numeric: from 1 - very low to 5 - very high

Feature Selection only got us to ~45% Accuracy

- Train, test, split
- Permutation feature importance for logistic regression classifier and decision tree classifier
- The low accuracy score could be because the model had to predict 5 different scores instead of a binary yes or no

Top five features Logistic Regression Classifier:

	cols	importance
13	goout	0.104732
16	female	0.056782
18	famsize_GT3	0.030284
3	studytime	0.029022
11	famrel	0.027129

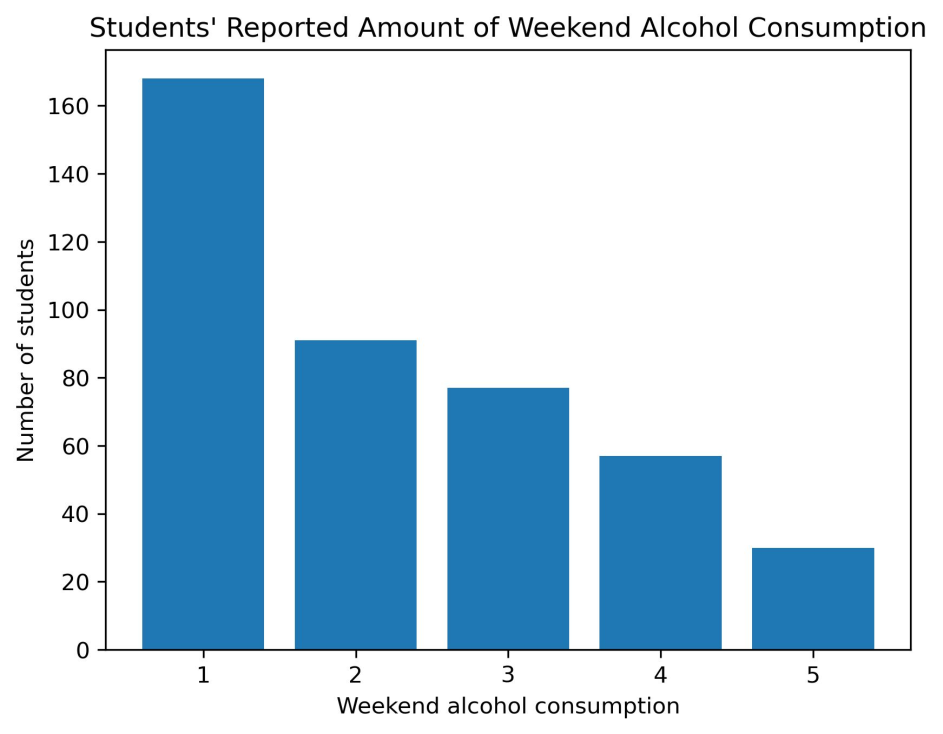
Top five features for the Decision Tree Classifier:

	cols	importance
13	goout	0.505994
11	famrel	0.252997
16	female	0.188013
20	age	0.174763
7	activities	0.146372

```
Logistic regression classifier with features ["goout", "female", "famsize_GT3"] earned an accuracy score of 0.453.
Decision tree classifier with features ["goout", "famrel", "age", "female"] earned an accuracy score of 0.425.
```

Binary classification may yield better predictions

- Compared low to average/high Walc
    - Binned 1-2 and 3-5
    - Yielded 60/40 split
- Data resampled to get 50/50 split



Five feature selection methods to identify key features:

By making the response variable binary, we were able to get the accuracy score up to around **72%**.

Feature Selection Methods*	Key Features Identified
KBest (threshold = 3)	failures, **goout**, absences
f_classif (threshold = 7)	**goout**, **female**, failures, Fjob_teacher, studytime, **famrel**, activities
mutual_info_classif (threshold = 5)	**goout**, higher, Fjob_other, **female**, Medu
VarianceThreshold (threshold = 0.3)	age, Medu, Fedu, traveltime, studytime, failures, **famrel**, freetime, **goout**, health, absences
permutation_importance (log reg)	**goout**, Fjob_services, **female**, Fjob_other, **famrel**
permutation_importance (decision tree)	**goout**, **female**, Medu, age, freetime

*Adapted from Halil Ertan's Medium article

Of six models, highest accuracy score was ~76.4%

Logistic Regression models:

```
permutation_importance model with features ["goout", "female", "famrel", "Fjob_services", "Fjob_other"] earned an accuracy score of 0.745.

f_classif model with features ["goout", "failures", "Fjob_teacher"] earned an accuracy score of 0.745.

educated_guesswork model with features ["goout", "female", "activities", "famrel", "Fjob_services"] earned an accuracy score of 0.764.
```

Decision Tree models:

```
permutation_importance model with features ["goout", "female", "Medu", "age", "health"] earned an accuracy score of 0.679.

f_classif model with features ["goout", "failures", "Fjob_teacher"] earned an accuracy score of 0.736.

educated_guesswork model with features ["goout", "female", "activities", "famrel", "Fjob_services"] earned an accuracy score of 0.632.
```

Feature	Hypothesized Importance	Hypothesized Effect	Calculated Importance	Calculated Coefficient
`goout`	Very high	Positive, large	0.146	0.860
`famrel`	High	Negative, moderate	0.027	–0.458
`female`	Very high	Negative, large	0.021	–0.847
`Fjob_services`	Medium	Positive, moderate	0.014	0.523
`activities`	None	Negligible, N/A	0.007	–0.475

Main takeaways

1. Response variable type matters → a small change can lead to a large accuracy boost
2. Utilize multiple different feature selection methods
3. Good to have some intuition when putting together a model

Questions?

Bibliography, Pt. 1

Bello, Camille. “Europe is home to the world’s heaviest drinkers. Which country drinks the most alcohol?” *Euronews*, January 31, 2024. https://www.euronews.com/health/2023/06/30/so-long-dry-january-which-country-drinks-the-most-alcohol-in-europe.

Cerqueira, Ana, Tania Gaspar, Fábio Botelho Guedes, Emmanuelle Godeau, and Margarida Gaspar De Matos. “Alcohol and tobacco use in Portuguese adolescents: The relationship with social factors, future expectations, physical and psychological symptoms.” *Children & Society* 36, no. 5 (February 11, 2022): 1010–25. https://doi.org/10.1111/chso.12552.

Deeken, Friederike, Tobias Banaschewski, Ulrike Kluge, and Michael A. Rapp. “Risk and protective factors for alcohol use disorders across the lifespan.” *Current Addiction Reports* 7, no. 3 (June 4, 2020): 245–51. https://doi.org/10.1007/s40429-020-00313-z.

Erol, Almila, and Victor M Karpyak. “Sex and gender-related differences in alcohol use and its consequences: Contemporary knowledge and future research considerations.” *Drug and Alcohol Dependence* 156 (September 5, 2015): 1–13. https://doi.org/10.1016/j.drugalcdep.2015.08.023.

Ertan, Halil. “Feature selection methods in SciKit Learn | Medium.” *Medium*, November 10, 2023. https://medium.com/@hertan06/which-features-to-use-in-your-model-350630a1e31c.

Bibliography, Pt. 2

Marques-Vidal, Pedro, and Carlos Matias Dias. “Trends and determinants of alcohol consumption in Portugal: results from the National Health Surveys 1995 to 1996 and 1998 to 1999.” *Alcoholism Clinical and Experimental Research* 29, no. 1 (January 1, 2005): 89–97. https://doi.org/10.1097/01.alc.0000150001.31722.d1.

NIAAA. “Risk factors: Varied Vulnerability to Alcohol-Related Harm.” National Institute on Alcohol Abuse and Alcoholism, February 27, 2024. https://www.niaaa.nih.gov/health-professionals-communities/core-resource-on-alcohol/risk-factors-varied-vulnerability-alcohol-related-harm.

Paixão, Maria Margarida, and Mélissa Mialon. “Help or hindrance? The alcohol industry and alcohol control in Portugal.” *International Journal of Environmental Research and Public Health* 16, no. 22 (November 18, 2019): 4554. https://doi.org/10.3390/ijerph16224554.

Regional Office for Europe. “Trends in Alcohol Consumption 2016: Portugal.” *WHO.Int*. World Health Organization, 2019. https://cdn.who.int/media/docs/librariesprovider2/country-sites/portugal/achp_fs_portugal.pdf.

Silvestre, Paulo, Jorge Oliveira, Hélder Trigo, Paulo Jorge Ferreira Lopes, and Nuno Colaço. “Risk factors of alcohol consumption among portuguese adolescents and young adults data from the Global-School Based Student Health Survey.” *International Journal of Drug Development and Research* 7, no. 4 (2014): 50–55. https://www.researchgate.net/publication/348944503_Risk_factors_of_alcohol_consumption_among_portuguese_adolescents_and_young_adults_data_from_the_Global-School_Based_Student_Health_Survey.

Statista. “Prevalence of alcohol consumption in Portugal 2022, by gender,” March 25, 2024. https://www.statista.com/statistics/1457946/portugal-alcohol-consumption-prevalence-by-gender/.

TPN/Lusa. “Portuguese consume 12 litres of pure alcohol annually.” *The Portugal News*, May 21, 2021. https://www.theportugalnews.com/news/2021-05-21/portuguese-consume-12-litres-of-pure-alcohol-annually/59951.