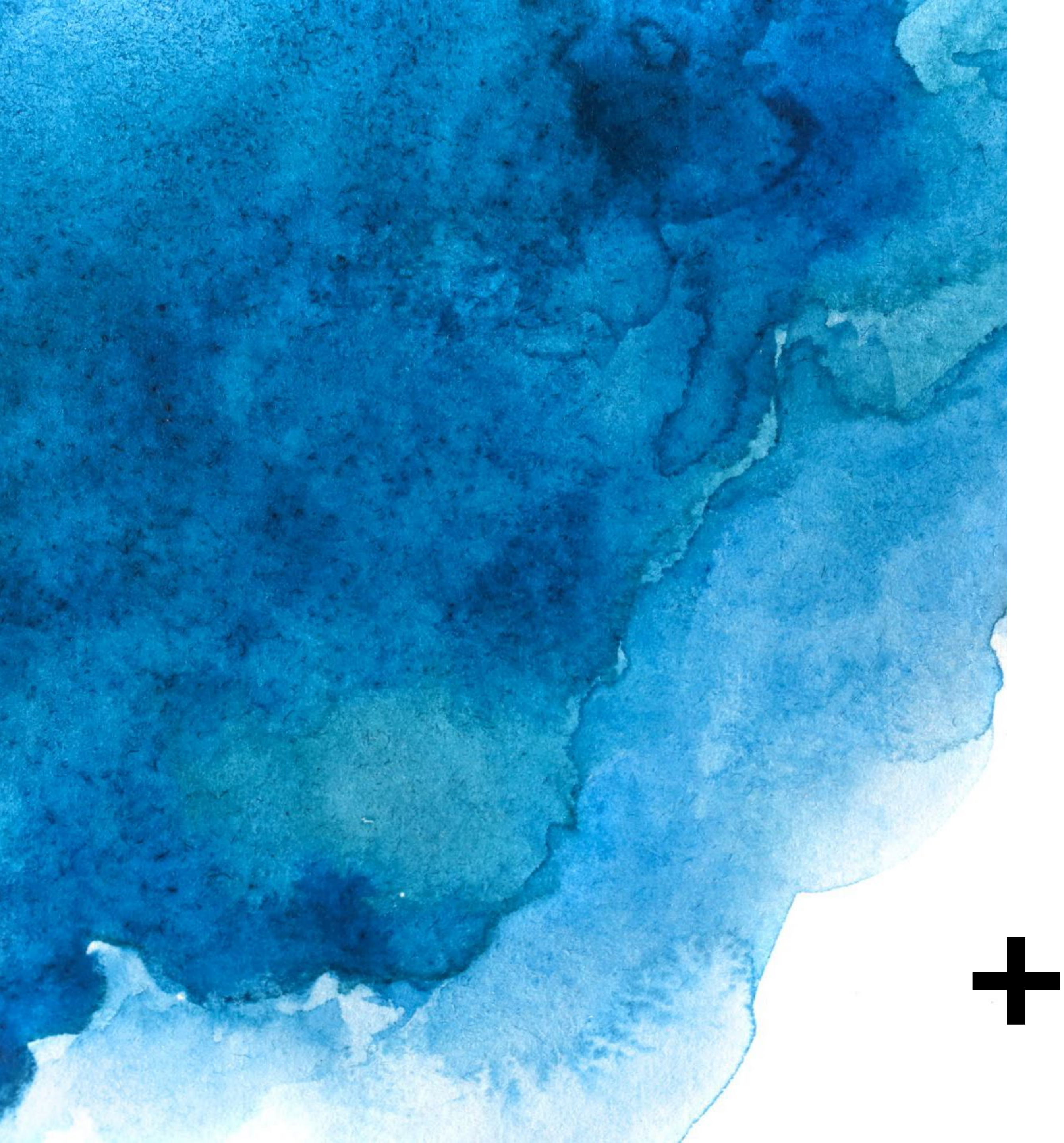


Capstone Project

Machine Learning Fundamentals

Jiawei Sun

2023/01/09

Table of Contents

- Overarching Analysis Question
- Exploring, Augmenting, and Visualizing the Data
- Classification Approaches
- Regression Approaches
- Conclusion

Overarching Analysis Question:

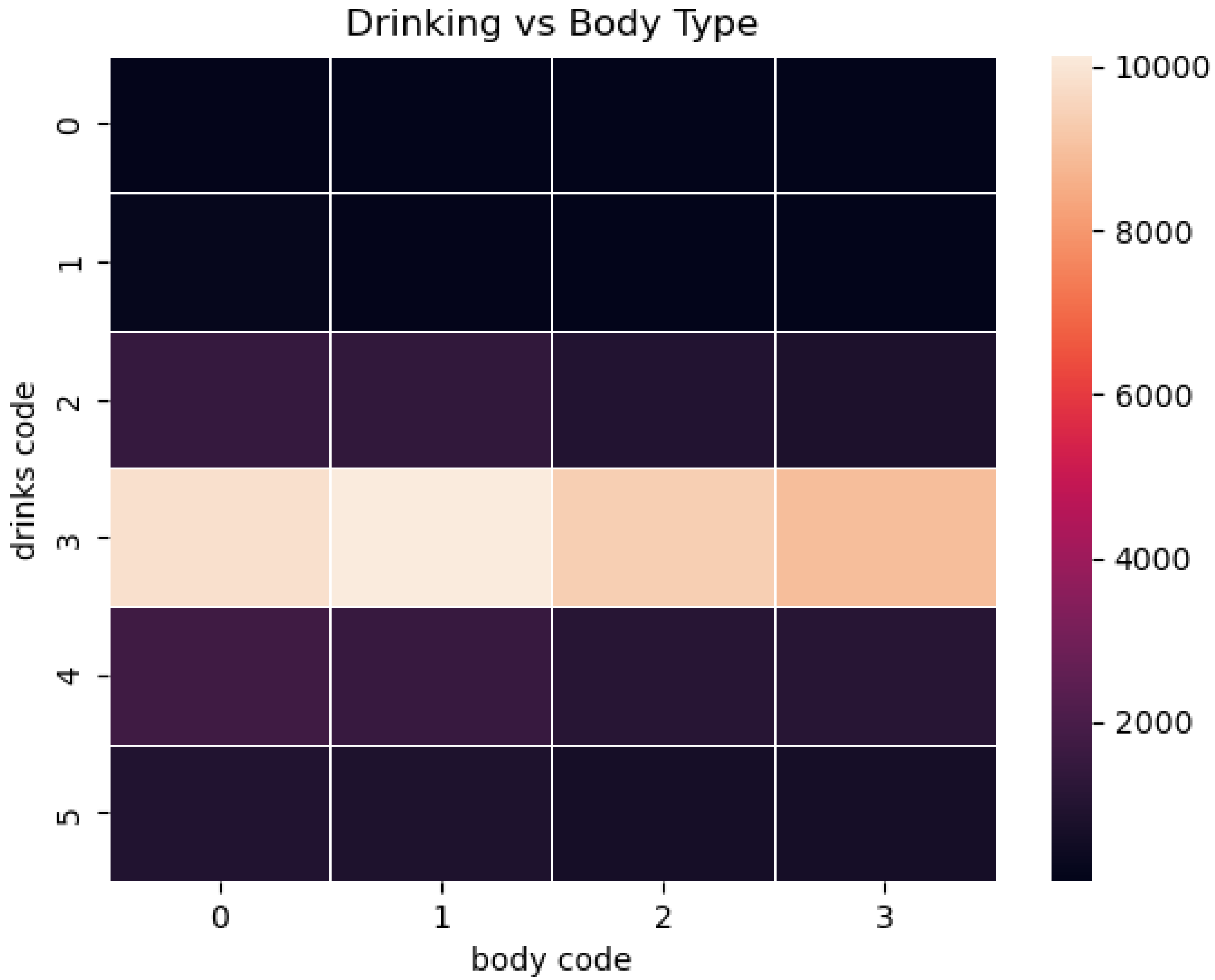
Is it possible to predict attributes of someone through their habits in smoking, drinking, and drugs, and their relationship status

Categorization of Body Type

- Logically, the attribute with the closest correlation to Smoking, Drinking, and Drugs would be the Body Type of the person.
- Mapping Body Types in the dataset
  - I generalized the body_type data into 4 linear groups: **below average** : 0, **average** : 1, **above average** : 2, **ideal** : 3
  - "overweight", "used up", "full figured", "skinny", "a little extra", "thin", "curvy" → 0
  - "average" → 1
  - "fit" → 2
  - "althetic", "jacked" → 3
- This sort of categorization, which incorporates many body types in "below average", is meant to account for human psychology, where people would rather not associate themselves with a "below average" body type and are more likely to place themselves more favorably.

*Disclaimer: These are just general body categorizations for the means of this exercise. They are in no way supposed to reflect a "correct" body type, not are they representative of my belief in how body types should be classified. This categorization also does not imply that any body type is inherently better than another.

Drinking vs. Body Type

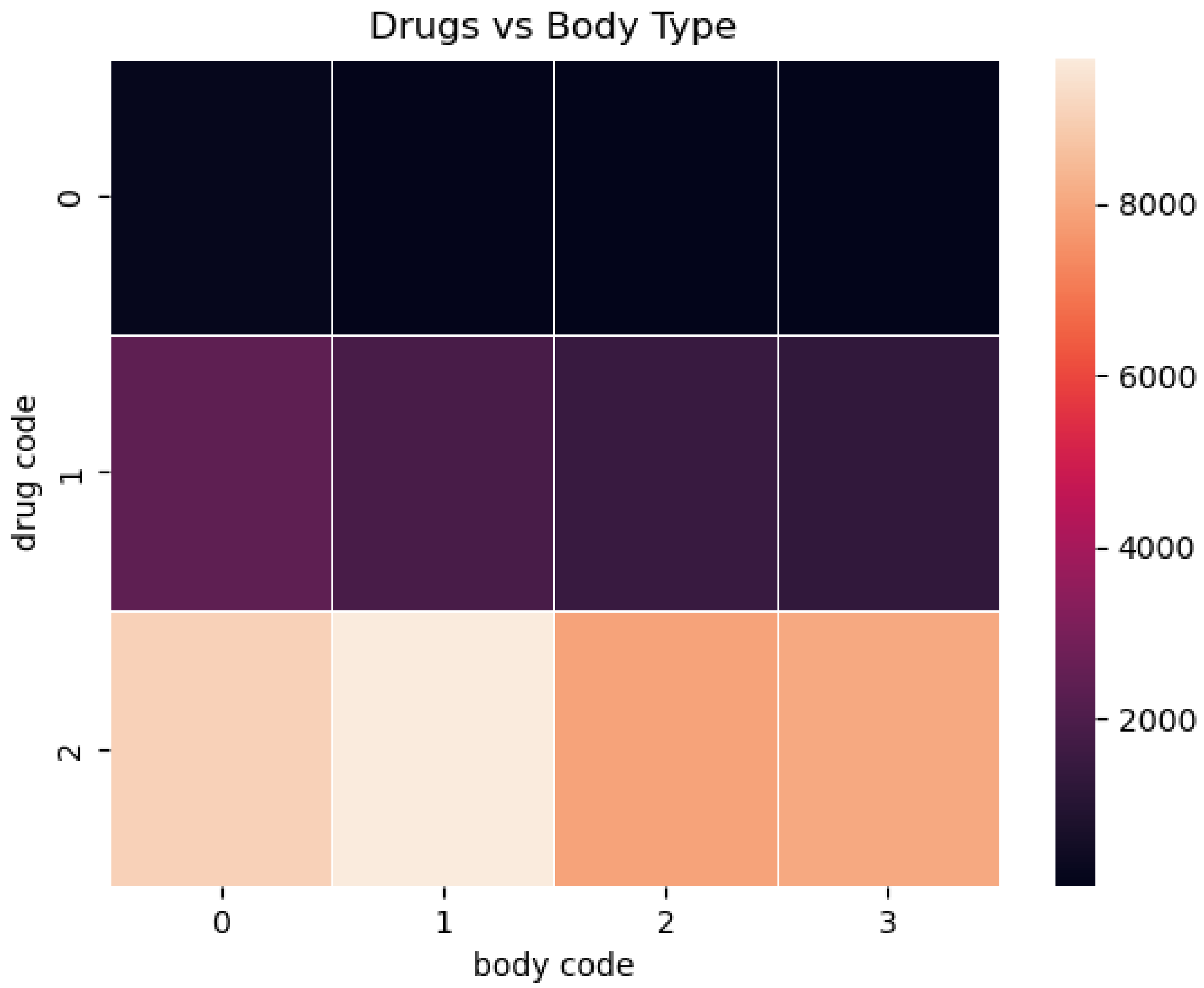


Legend:

Drinks Code:

- 'desperately’:0
- 'very often’:1
- 'often’:2
- 'socially’:3
- 'rarely’:4
- 'not at all':5

Drugs vs. Body Type

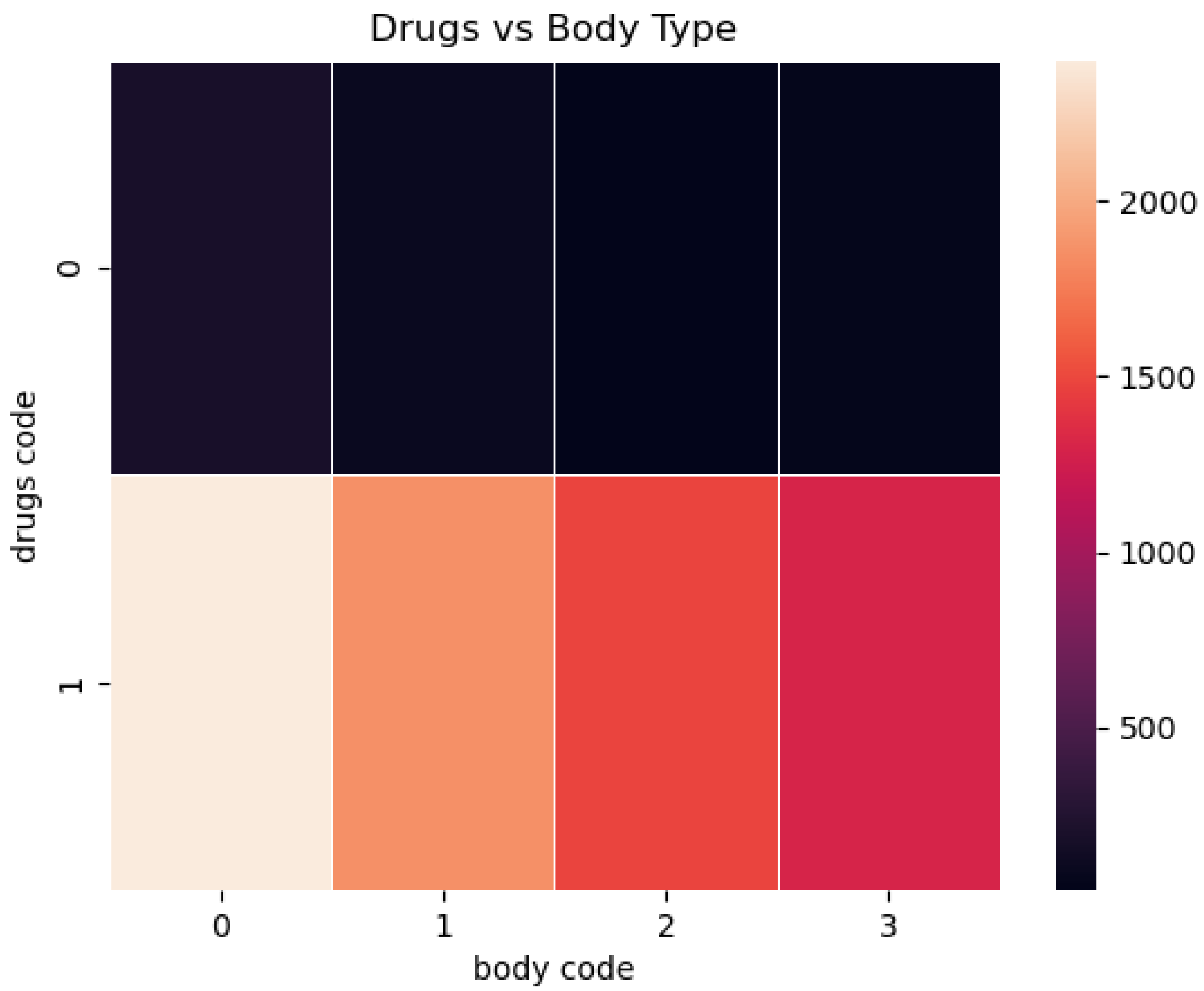


Legend:

Drugs Code:

- 'often’:0
- 'sometimes’:1
- 'never':2

Drugs vs. Body Type (Removed Never)



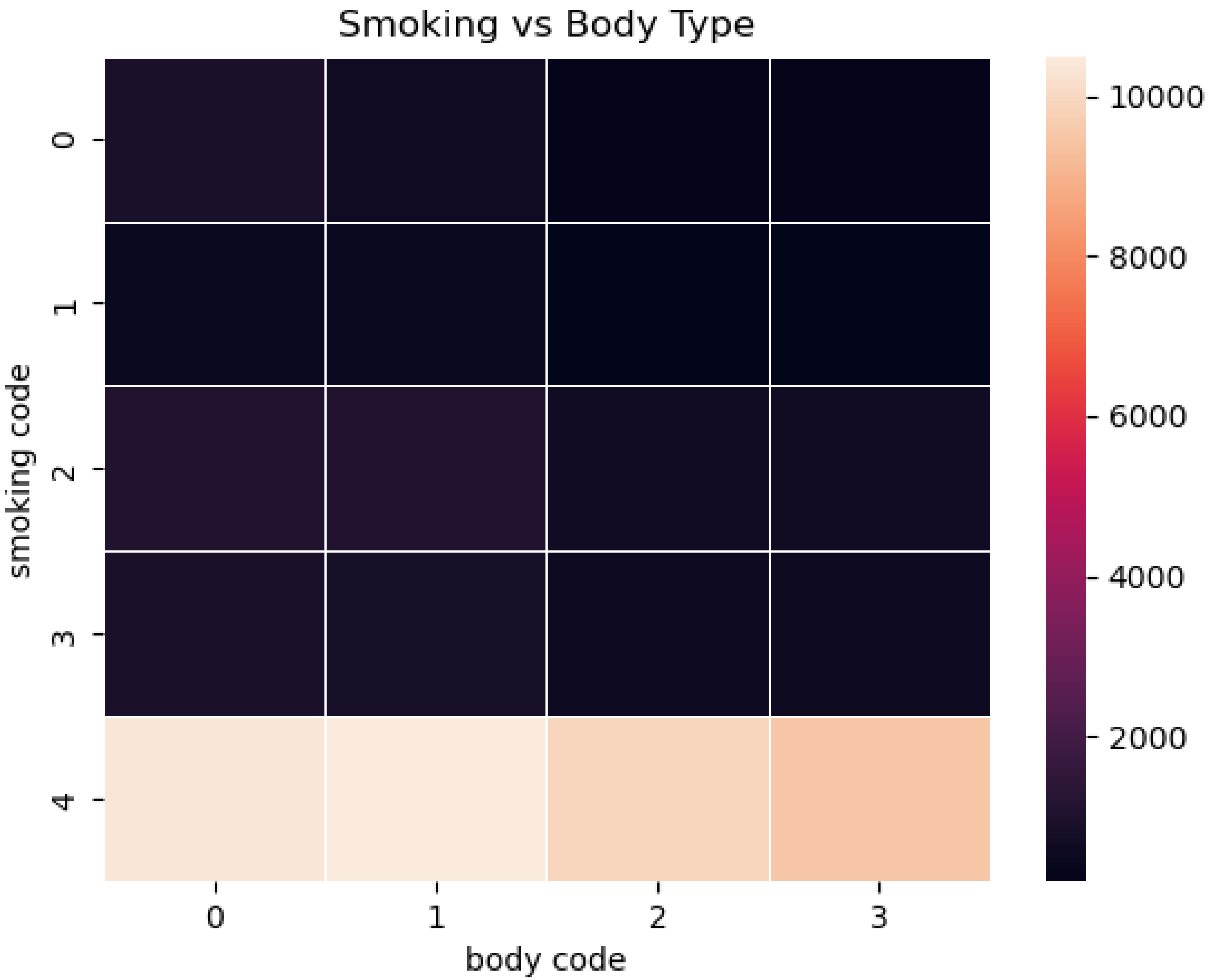
Legend:

Drugs Code:

- 'often':0
- 'sometimes':1

Since most of the data in this dataset had never done drugs, I removed that to see if the remaining answers would show any new relationships.

Smoking vs. Body Type

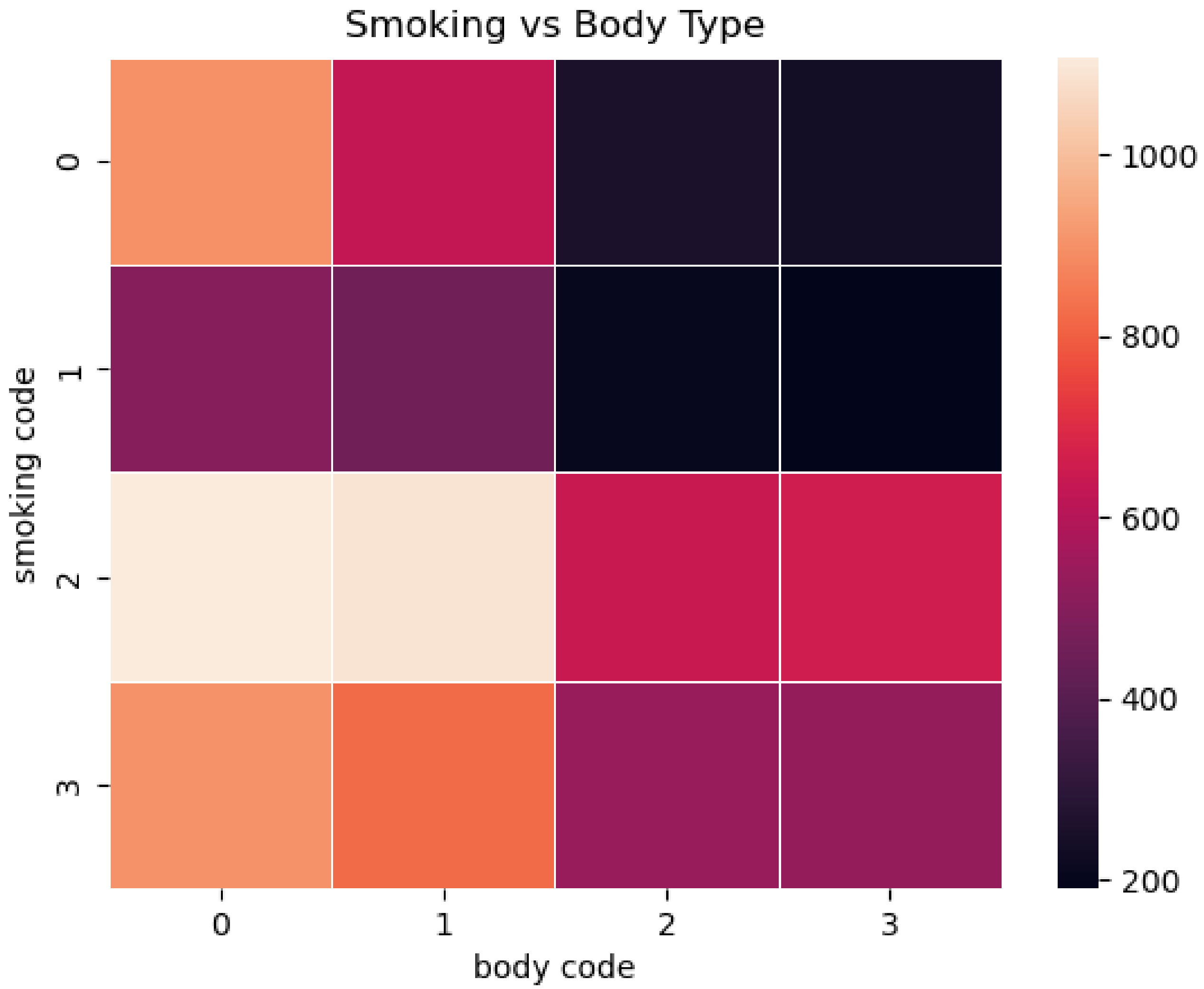


Legend:

Smoking Code:

- 'yes':0,
- sometimes':1
- 'when drinking':2
- 'trying to quit':3
- 'no':4

Smoking vs. Body Type (Removed No)



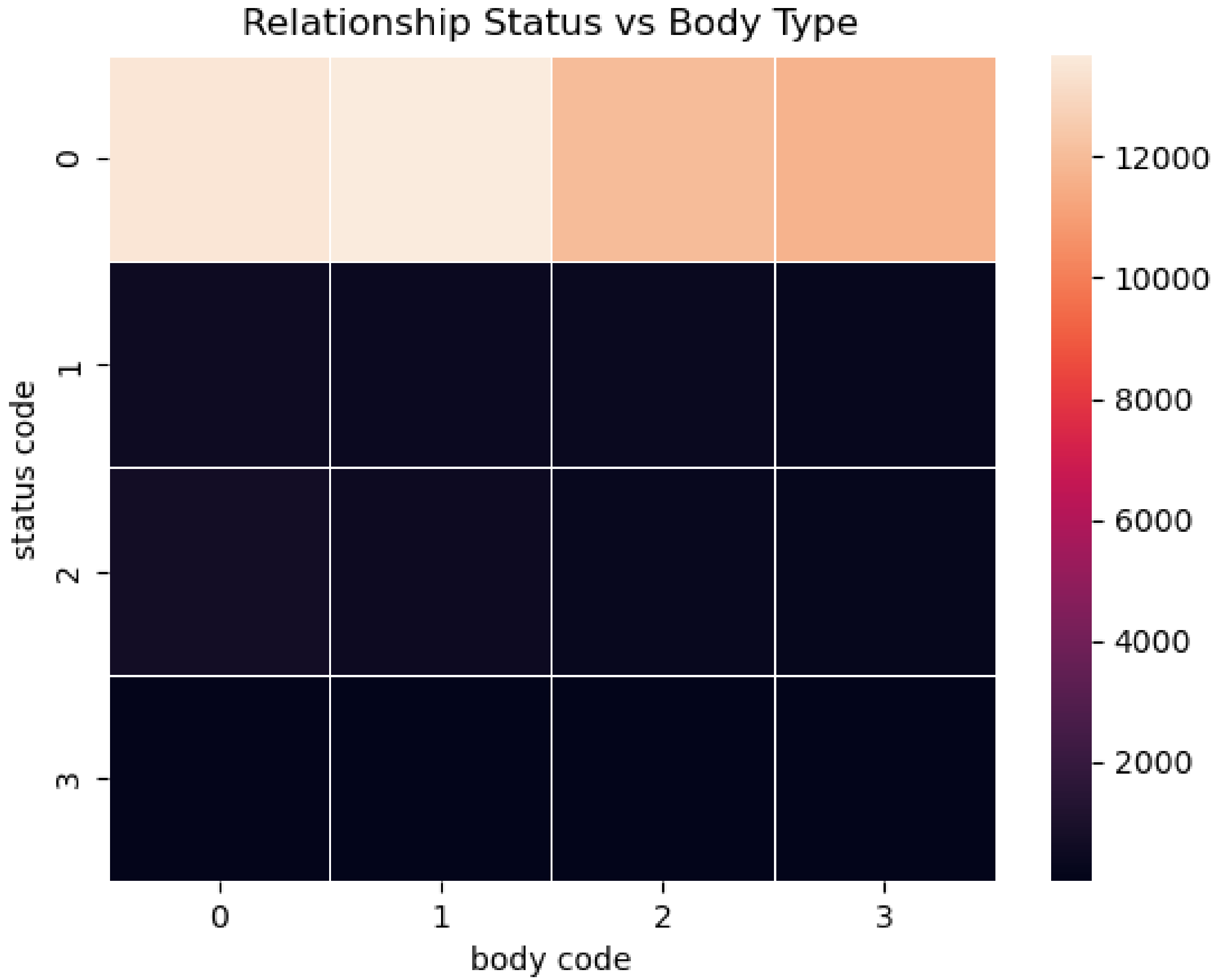
Legend:

Smoking Code:

- 'yes':0
- 'trying to quit':1
- sometimes':2
- 'when drinking':3

Since most of the data in this dataset does not smoke, I removed that to see if the remaining answers would show any new relationships.

Relationship Status vs. Body Type

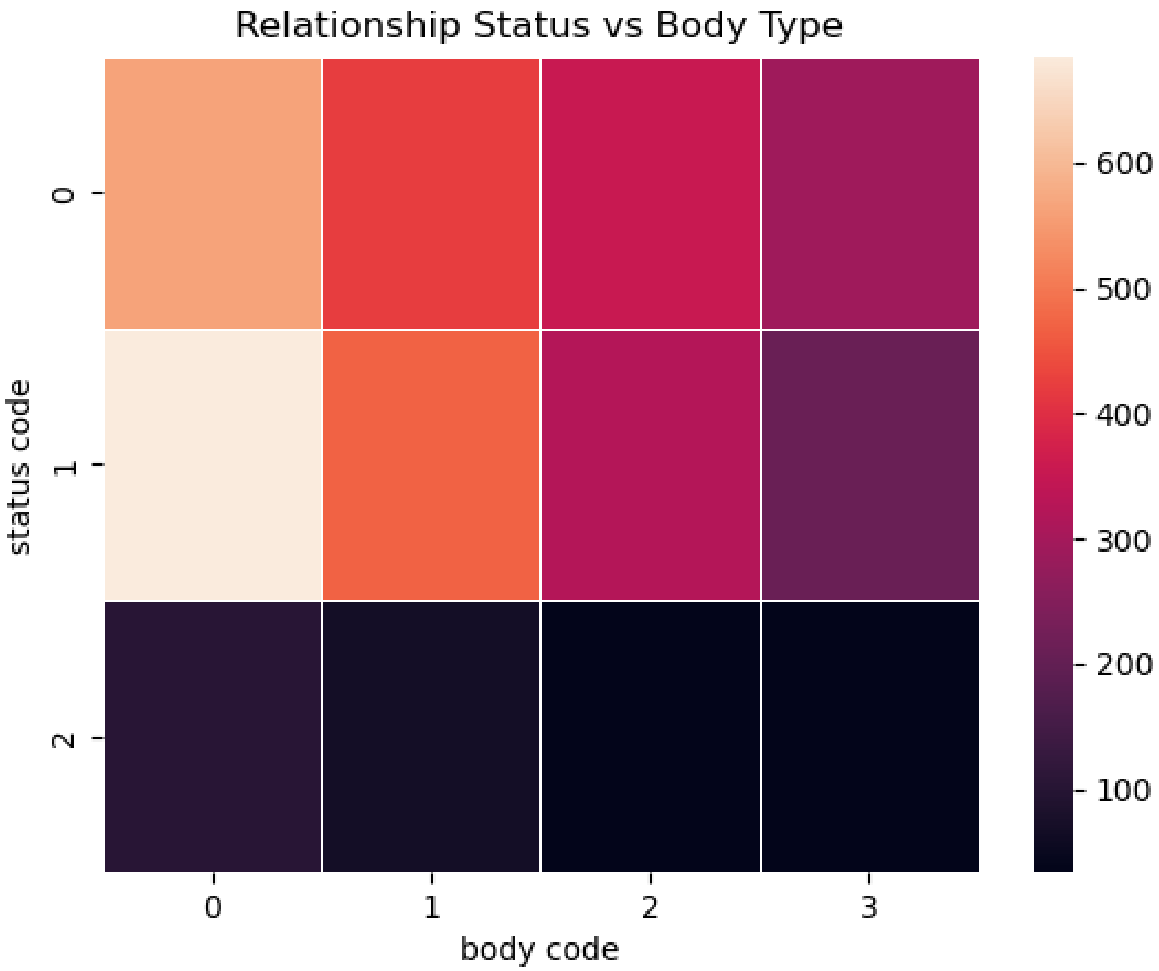


Legend:

Status Code:

- 'single’:0
- 'available’:1
- 'seeing someone’:2
- 'married':3

Relationship Status vs. Body Type (Removed Single)



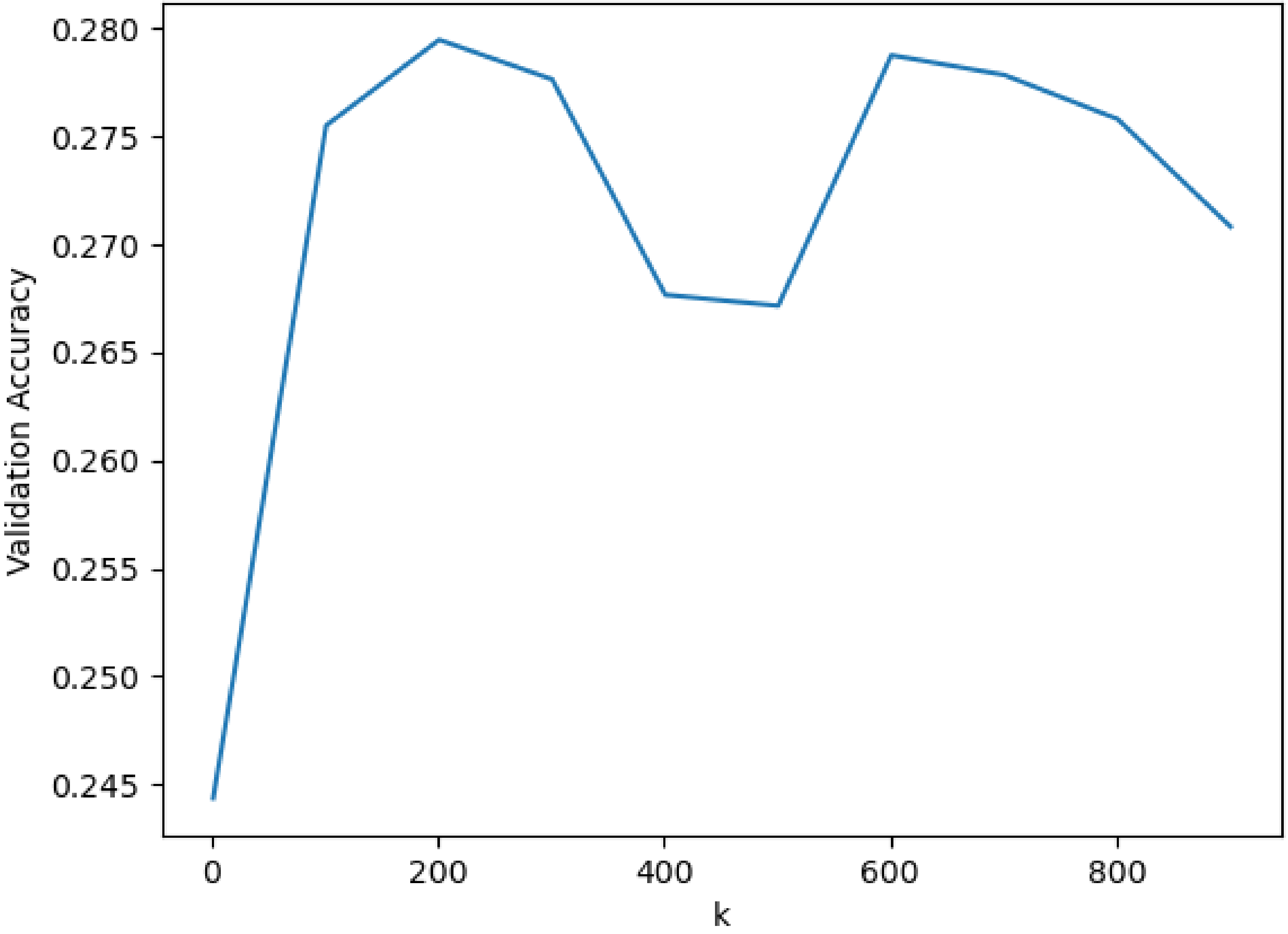
Legend:

Status Code:

- 'available’:0
- 'seeing someone’:1
- 'married’:2

Since most of the data in this dataset are single, I removed that to see if the remaining answers would show any new relationships.

Finding K, K Neighbors Classifier



1-1000

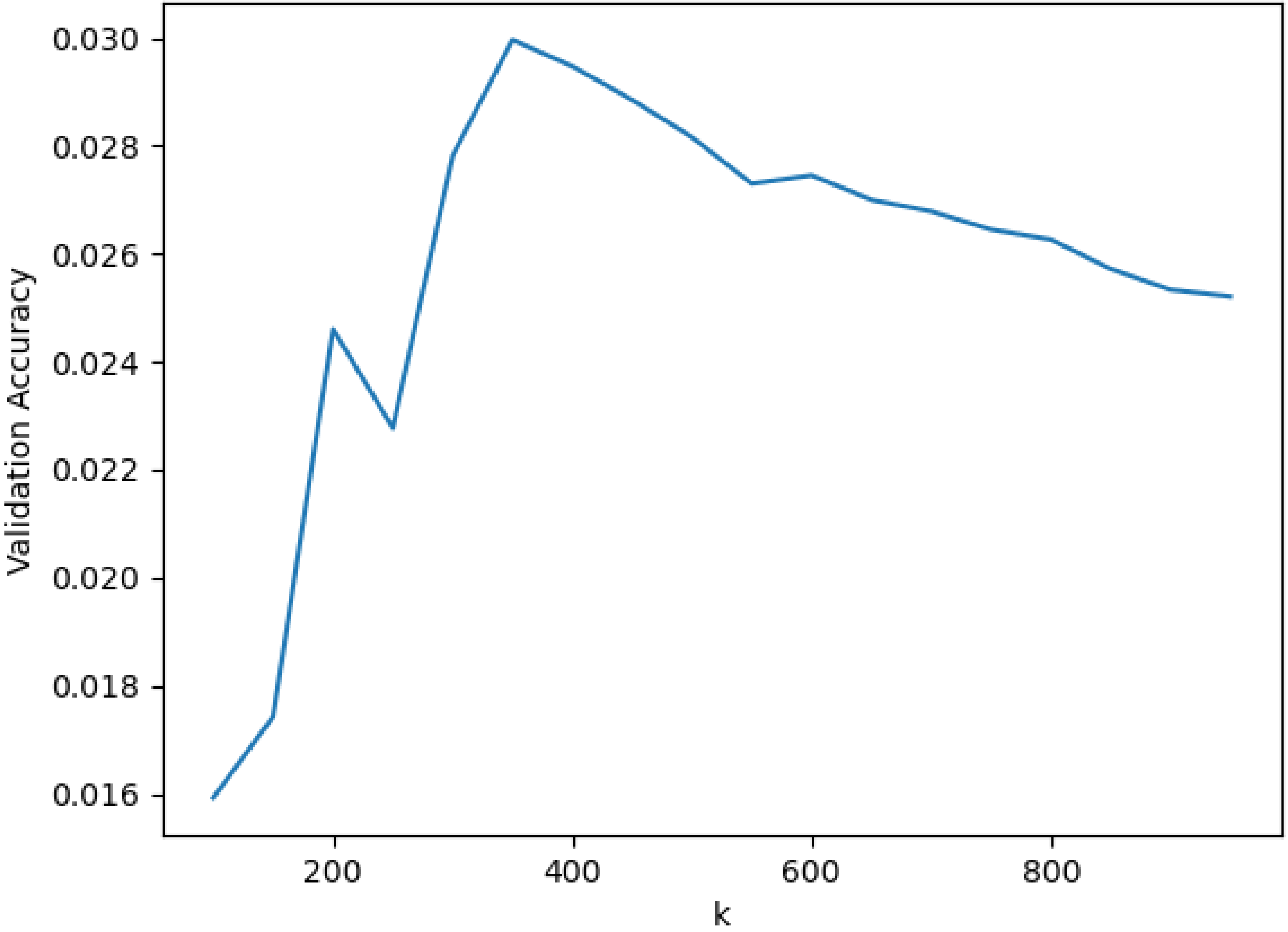
Steps: 100

Best Validation Accuracy:

~0.279

k = 200

Finding K, K Neighbors Regressor



100-1000

Steps: 50

Best Validation Accuracy:

~0.030

k = 350

Random Forest Classifier

Validation Accuracy: 0.30179262578936644

Feature Importances:

[0.19701189 0.34104799 0.32211338 0.13982674]

[Status, Smoking, Drinking, Drugs]

Body Type Conclusion

So, it seems that the correlation between body type, and somebody’s relationship status and their drinking, smoking, and drug habits is not very strong.

The best validation accuracy achieved through a Random Forest Classifier, which had an accuracy of 0.3018.

From the Random Forest Classifier, it seems that the most relevant aspect to someone’s body type is their smoking habits, followed closely by their drinking habits.

Comparisons with Income Classification

After the Body Type showed weak correlations, I was curious about how these habits can affect people's incomes.

I divided the incomes into >50k, and <= 50k

Random Forest Classifier

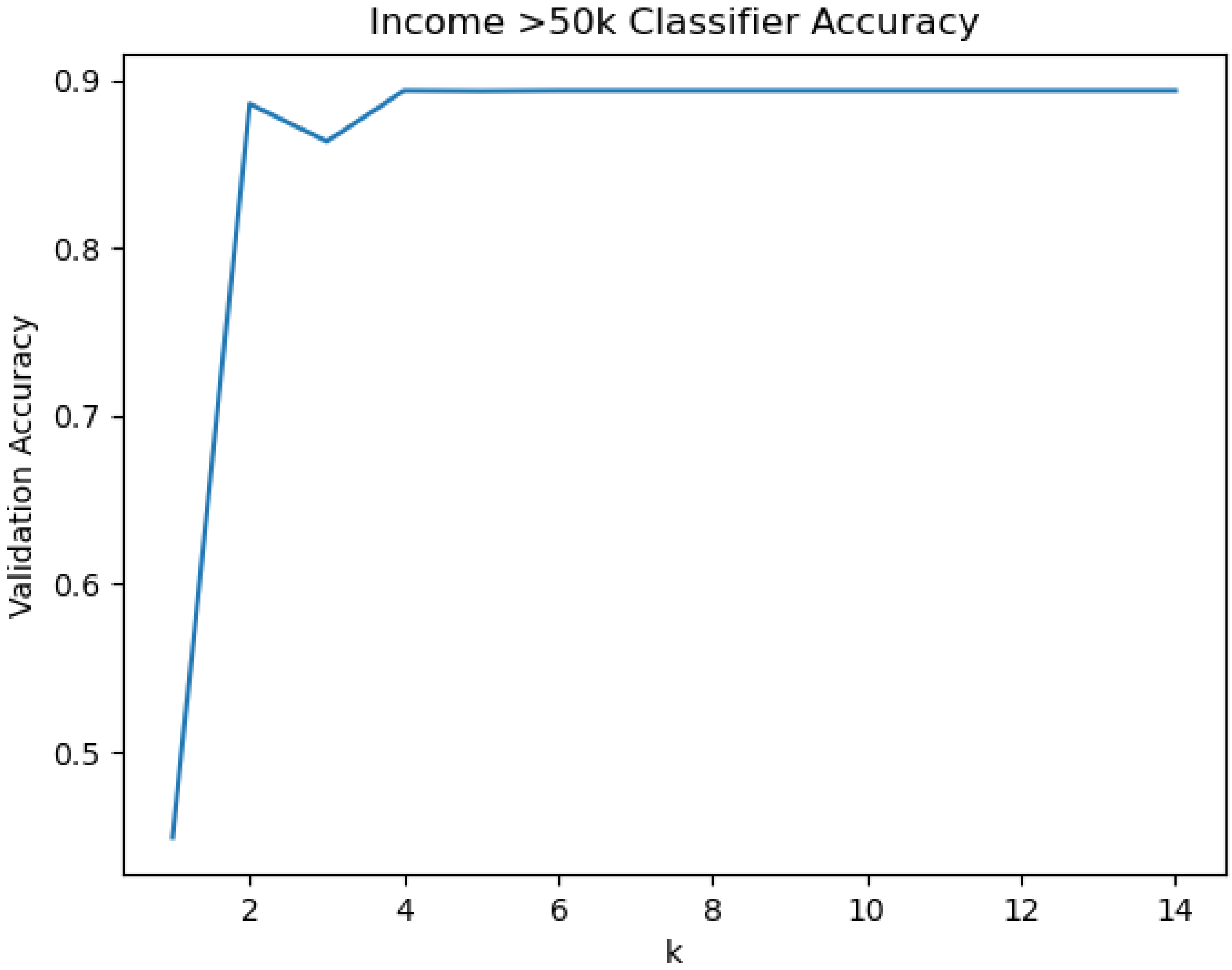
Validation Accuracy: 0.8933446295773322

Feature Importances:

[0.19381837 0.29494015 0.32636322 0.18487826]

[Status, Smoking, Drinking, Drugs]

Finding K, K Neighbors Classifier



1-15

Steps: 1

Best Validation Accuracy:

0.8937211710439612

k = 6

Income Classification Conclusion

So, it seems that the correlation between income above 50k, and somebody’s relationship status and their drinking, smoking, and drug habits is not very strong.

The best validation accuracy achieved through a K Neighbors Classifier with k=6, which had an accuracy of 0.8937.

From the Random Forest Classifier, it seems that the most relevant aspect to someone’s body type is their drinking habits, followed closely by their smoking habits, which is like the importances of body type classification, but flipped in order.

Conclusion

While there was quite a weak correlation between the intended classification – body type– and the chosen features of this analysis (relationship status, drinking habits, smoking habits, and drug use), I found a stronger correlation between the income classifications and the chosen features.

There could be multiple reasons for this, one of them being that income is an objective datapoint, unlike the subjective classification of body type. Furthermore, through the analysis of the data, I found that many of the incomes were unreported, so the total number of data points with income reported was significantly lower than that of body types, which could have affected the validation accuracy of the machine learning models.