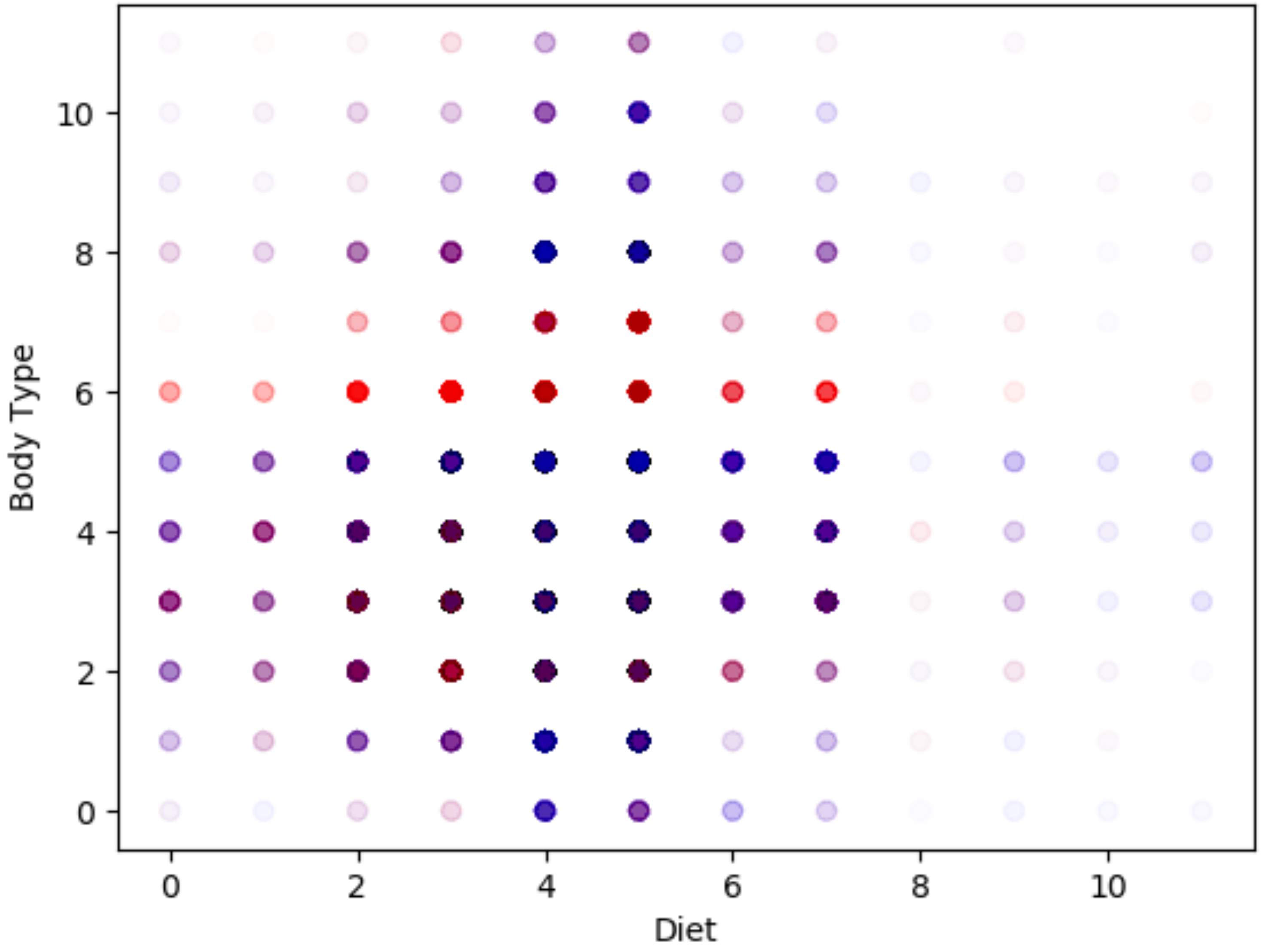
Capstone Project
OKCupid Dataset

Machine Learning Fundamentals
Andres Londono
2 December 2018

Table of Contents

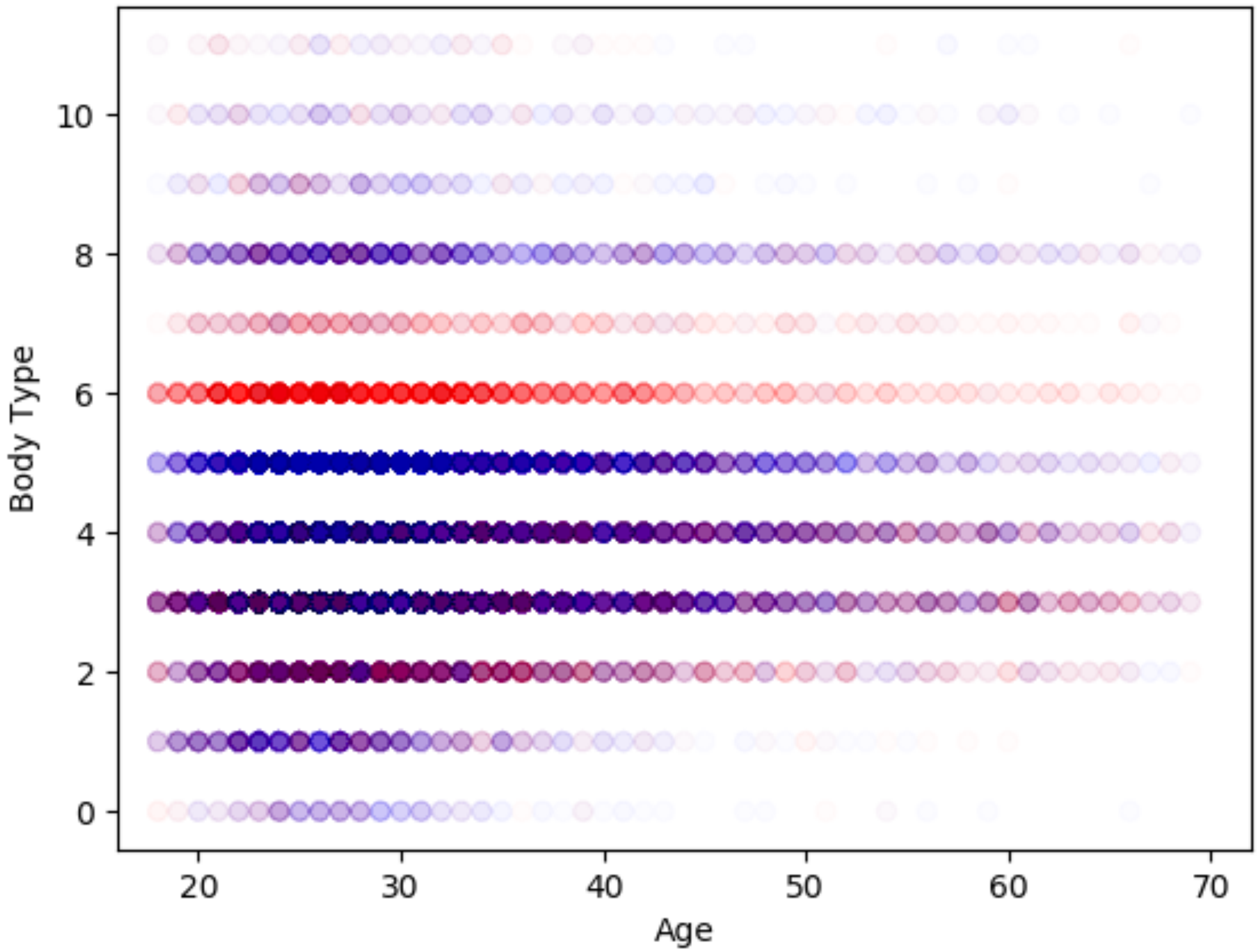
- Exploration of the dataset
- Question to answer
- Augmenting the dataset
- Classification approaches
- Regression approaches
- Conclusions - next steps

Exploration of the dataset



- Graph of diet vs body type.
- Diet and body type were mapped to values.
- The red dots are females, the blue are males.
- There are two clear body types almost exclusive to women (values 6 and 7). These are "curvy" and "full figured".
- The diets with values 8 and over are halal and kosher. They were put at the end of the mapping on purpose, to not create a hole in the middle of the data.

Exploration of the dataset



- Graph of age vs body type.
- Body type is mapped to values.
- The red dots are females, the blue are males.
- The two body types almost exclusive to women (values 6 and 7) are "curvy" and "full figured".
- Mostly people under 50 are represented in the dataset.
- Here I started to see areas where mostly men were represented, but other areas seems to be more represented by women.

Question to answer

- Based on the previous graphs, females and males seem to congregate on different areas of the graphs, when considering diet, body type, and age.

- Could I predict the gender of the person based on diet, body type, and age?

Augmenting the data

- As most of the data in this dataset is categorical, in order to create some features, we need numerical data. The **diet** and **body type** categories were mapped to numerical data.

- In this case what values to assign to each category could affect the results.

- For example, it was noticed that in **diet**, the “halal" and “kosher" categories had very few responses compared to the other categories, so putting these less chosen categories in the middle of the numeric range, as was done initially, would divide the data skewing the model results. For this reason it was decided to move these categories to the ends of the numerical range.

Augmenting the data

- As mentioned before, diet and body type were mapped to values. These were the mappings done.

```
diet_mapping = {"strictly vegan": 0,
                "vegan": 0,
                "mostly vegan": 1,
                "strictly vegetarian": 2,
                "vegetarian": 2,
                "mostly vegetarian": 3,
                "strictly anything": 4,
                "anything": 4,
                "mostly anything": 5,
                "strictly other": 6,
                "other": 6,
                "mostly other": 7,
                "strictly kosher": 8,
                "kosher": 8,
                "mostly kosher": 9,
                "strictly halal": 10,
                "halal": 10,
                "mostly halal": 11}
```

```
body_type_mapping = {"used up": 0,
                     "skinny": 1,
                     "thin": 2,
                     "average": 3,
                     "fit": 4,
                     "athletic": 5,
                     "curvy": 6,
                     "full figured": 7,
                     "a little extra": 8,
                     "jacked": 9,
                     "overweight": 10,
                     "rather not say": 11}
```

Augmenting the data

- Also removed the NaN from the features to classify.

```
wanted_features = ["diet_code", "body_type_code", "sex_code", "age"]
rows_to_cluster = df.dropna(subset = wanted_features)
```

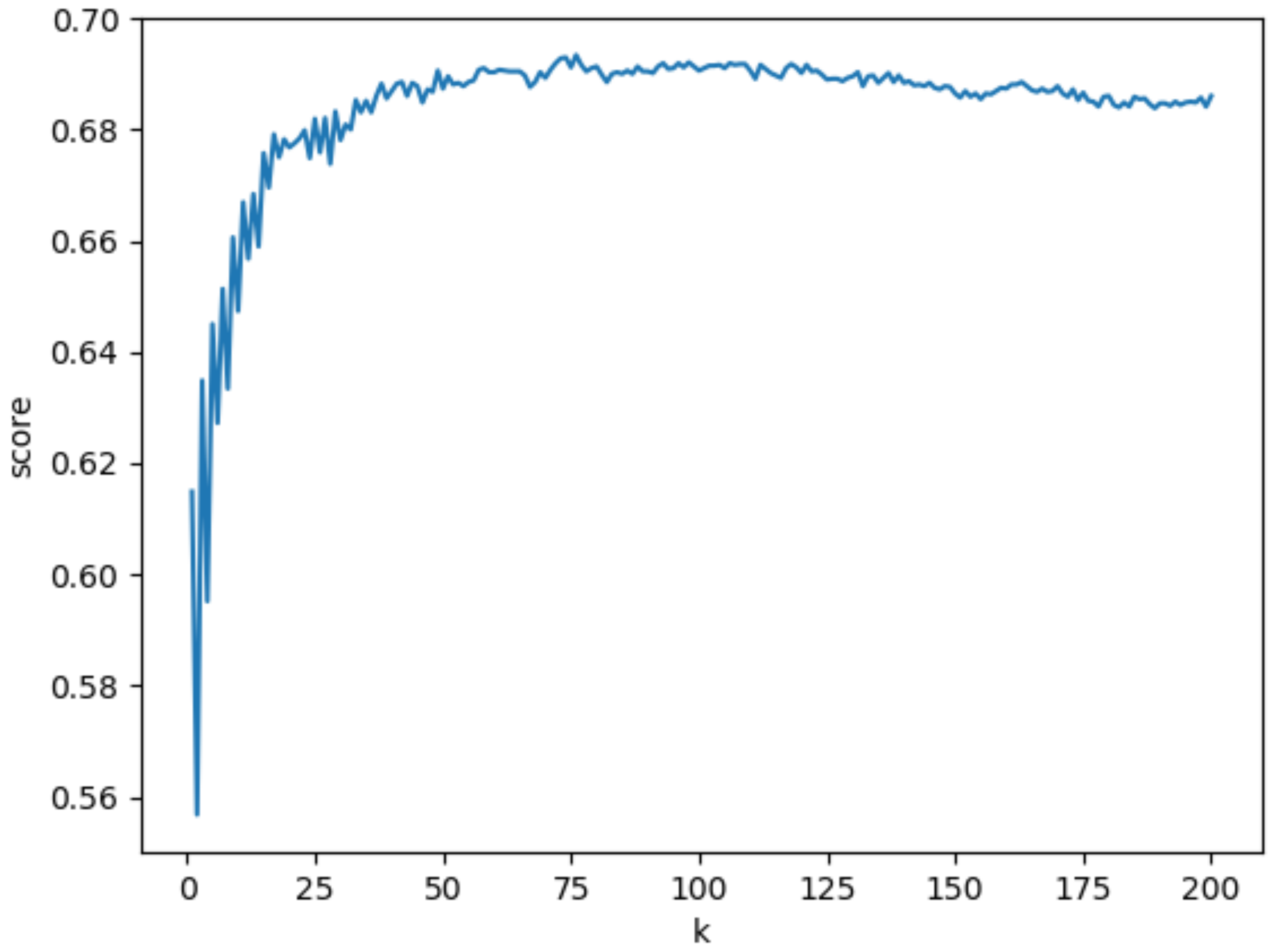
Classification approaches

- The two classification approaches used were:

1. K-Nearest Neighbour Classification
2. Suppor Vector Machine (SVC)

Classification approaches

- K-Nearest Neighbour Classification used to predict gender

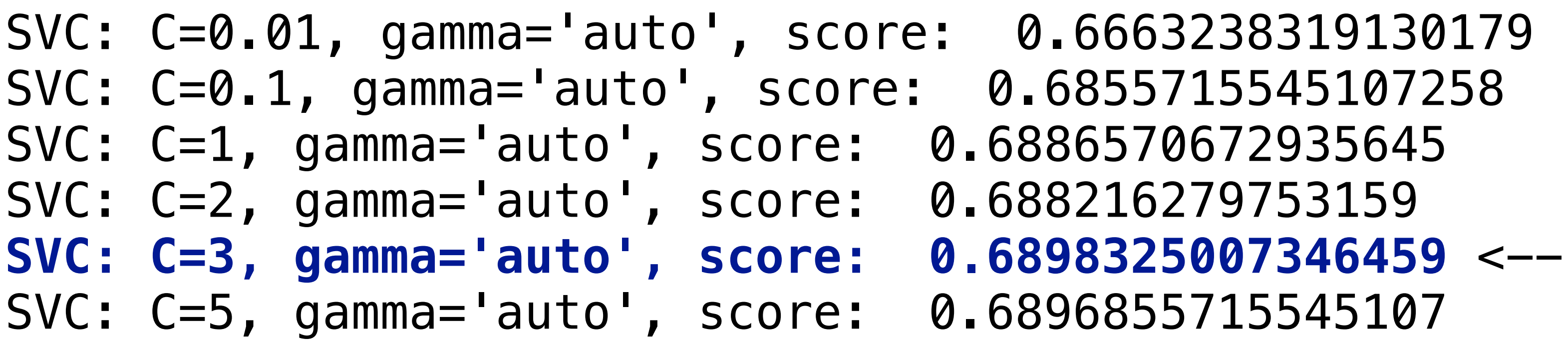


- Comparison of k-neighbours vs classification score, to find the best k-value to use.
- The best k-value found was k=76, with a score of 0.6934
- The proportion of males vs females in the initial data was m=0.597688, f=0.402312. So our result is a bit better than random chance.

Classification approaches

- Support Vector Machine (SVC) used to predict gender.

- SVC was used with the **radial bias function (rbf) kernel**, and with different values for C and gamma. The best results were found with gamma=‘auto’. Some examples of the different C values used:



```
SVC: C=0.01, gamma='auto', score:  0.6663238319130179
SVC: C=0.1, gamma='auto', score:  0.6855715545107258
SVC: C=1, gamma='auto', score:  0.6886570672935645
SVC: C=2, gamma='auto', score:  0.688216279753159
SVC: C=3, gamma='auto', score:  0.6898325007346459 <--
SVC: C=5, gamma='auto', score:  0.6896855715545107
```

- The best score found was 0.6898, which again is bit better than random chance, according to proportion of males and females in the original dataset.

Regression approaches

- **Multiple Linear Regression** was tried on our features, trying to answer the proposed question, but the results were not good.

- **K-Nearest Neighbour Regressor** was also tried also tried and the score (R^2) was 0.108, which is very poor.

- As the value to predict in our question — gender — has only two possible options in this data, a regression works better with a range of values as prediction, which doesn’t work with our proposed question.

Conclusions - Next steps

- We tried two different classification methods, K-Nearest Neighbour and Support Vector Machines, to try an answer the question, can we predict gender based on diet, body type, and age?

- Based on these two classification methods, we were able to predict gender a bit better than random alone. The distribution of males and females in the dataset was about 60% males - 40% females. The classification using K-Nearest Neighbour had an accuracy score of 69.34%, while the SVC had a score of about 69.0%.

- When mapping categorical data to numerical data, the way the categories are organised can have a big influence on the result. Some data like the one that was picked for the question above and their different categories didn't allow for a straightforward conversion to numeric data. Particularly diet can be mapped in many different ways.

- Mapping body types was also problematic, as some of the categories included could be interpreted in different ways. Maybe weight, instead of body type, would have been a better feature to correlate with diet.

- For our particular question, regression was not a good approach. Our question was basically a classification question, which didn't work well with a regression. A regression works better with a range of values as possible answer.

- It would be interesting to try other questions maybe based on religiousness, which might have some influence in diet. For example the diets halal and kosher are expected to be based on religion.

- It is easy to fall in the temptation of manipulating the data, for example by removing outliers, or mapping the data in a particular way, to get the results we want.