**Problem statement:** can demographic and habits data be a good predictor of the body type?

**Data Source:** Kaggle

**Organization:** Okcupid

The data has 59946 rows and 31 columns

The data has demographic information of customers like age, sex,location, job income and habits information such as drinks, drugs and smoke, also there is text data that where customers give  brief information about them.

Here is the brief description of the columns.

Except the age, income and height, all the columns are categorical variables with overlapping values hence needs to be grouped into appropriate categories

**Tools used:**

**Python:** Data processing and modeling

|  |  |
| --- | --- |
| **Column** | **Description** |
| age | age of the customer |
| status | status of the customer |
| sex | sex of the customer |
| orientation | orientation of the customer |
| body\_type | body type of the customer |
| diet | diet of the customer |
| drinks | customer drinking information |
| drugs | information if the customer consumes drugs |
| education | education of the customer |
| ethnicity | ethnicity of the customer |
| height | height of the customer |
| income | income of the customer |
| job | job of the customer |
| last\_online | last online date and time of the customer |
| location | location of the customer |
| offspring | offspring of the customer |
| pets | pets of the customer |
| religion | religion of the customer |
| sign | sign of the customer |
| smokes | information if the customer smokes |
| speaks | language of the customer |
| essay0 | introduction of the customer |
| essay1 | introduction of the customer |
| essay2 | introduction of the customer |
| essay3 | introduction of the customer |
| essay4 | introduction of the customer |
| essay5 | introduction of the customer |
| essay6 | introduction of the customer |
| essay7 | introduction of the customer |
| essay8 | introduction of the customer |
| essay9 | introduction of the customer |

Count and percentage of missing values in the data:

|  |  |  |
| --- | --- | --- |
| **column** | **obs** | **Percent** |
| offspring | 35561 | 59.30% |
| diet | 24395 | 40.70% |
| religion | 20226 | 33.70% |
| pets | 19921 | 33.20% |
| essay8 | 19225 | 32.10% |
| drugs | 14080 | 23.50% |
| essay6 | 13771 | 23.00% |
| essay9 | 12603 | 21.00% |
| essay7 | 12451 | 20.80% |
| essay3 | 11476 | 19.10% |
| sign | 11056 | 18.40% |
| essay5 | 10850 | 18.10% |
| essay4 | 10537 | 17.60% |
| essay2 | 9638 | 16.10% |
| job | 8198 | 13.70% |
| essay1 | 7572 | 12.60% |
| education | 6628 | 11.10% |
| ethnicity | 5680 | 9.50% |
| smokes | 5512 | 9.20% |
| essay0 | 5488 | 9.20% |
| body\_type | 5296 | 8.80% |
| drinks | 2985 | 5.00% |
| speaks | 50 | 0.10% |
| height | 3 | 0.00% |
| orientation | 0 | 0.00% |
| status | 0 | 0.00% |
| sex | 0 | 0.00% |
| income | 0 | 0.00% |
| last\_online | 0 | 0.00% |
| location | 0 | 0.00% |
| age | 0 | 0.00% |

Distribution of class variable without preprocessing:

|  |  |
| --- | --- |
| **body\_type** | **obs** |
| thin | 4711 |
| skinny | 1777 |
| a little extra | 2629 |
| average | 14652 |
| curvy | 3924 |
| full figured | 1009 |
| rather not say | 198 |
| used up | 355 |
| overweight | 444 |
| athletic | 11819 |
| fit | 12711 |
| jacked | 421 |

**Data processing and cleaning**: all the categorical columns having similar values have been grouped into one category. All the numerical variables remain as is except the income column where the –1 which stands for missing values has been treated to remove those values

**Status**before: **single**and **available**values are grouped into **single**category

|  |  |
| --- | --- |
| **Status** | **Obs** |
| single | 55697 |
| available | 1865 |
| seeing someone | 2064 |
| married | 310 |
| unknown | 10 |

**Status**after:

|  |  |
| --- | --- |
| **Status** | **Obs** |
| single | 57562 |
| seeingsomeone | 2064 |
| married | 310 |
| unknown | 10 |

**Sex**: no categorization of the values in this column

|  |  |
| --- | --- |
| **Sex** | **Obs** |
| m | 35829 |
| f | 24117 |

Categorization used for **Body Type**:

|  |  |  |
| --- | --- | --- |
| **body\_type** | **obs** | **new category** |
| thin | 4711 | thin |
| skinny | 1777 | thin |
| a little extra | 2629 | overweight |
| average | 14652 | overweight |
| curvy | 3924 | overweight |
| full figured | 1009 | overweight |
| rather not say | 198 | overweight |
| used up | 355 | overweight |
| overweight | 444 | overweight |
| athletic | 11819 | fit |
| fit | 12711 | fit |
| jacked | 421 | fit |

**Body type** after:

|  |  |
| --- | --- |
| **Body Type** | **Obs** |
| fit | 24951 |
| overweight | 23211 |
| thin | 6488 |

Categorization used for **Diet**:

|  |  |  |
| --- | --- | --- |
| **diet** | **obs** | **new category** |
| strictly other | 452 | anything |
| other | 331 | anything |
| mostly other | 1007 | anything |
| strictly anything | 5113 | anything |
| mostly anything | 16585 | anything |
| anything | 6183 | anything |
| strictly halal | 18 | kosher/halal |
| mostly halal | 48 | kosher/halal |
| halal | 11 | kosher/halal |
| strictly kosher | 18 | kosher/halal |
| mostly kosher | 86 | kosher/halal |
| kosher | 11 | kosher/halal |
| vegan | 136 | vegan |
| strictly vegan | 228 | vegan |
| mostly vegan | 338 | vegan |
| vegetarian | 667 | vegeterian |
| strictly vegetarian | 875 | vegeterian |
| mostly vegetarian | 3444 | vegeterian |

**Diet**after:

|  |  |
| --- | --- |
| **Diet** | **Obs** |
| anything | 29671 |
| vegetarian | 4986 |
| vegan | 702 |
| kosher/halal | 192 |

Categorization used for **Drinks**:

|  |  |  |
| --- | --- | --- |
| **drinks** | **obs** | **new category** |
| socially | 41780 | yes |
| often | 5164 | yes |
| very often | 471 | yes |
| desperately | 322 | yes |
| not at all | 3267 | no |
| rarely | 5957 | no |

**Drinks**after:

|  |  |
| --- | --- |
| **Drinks** | **Obs** |
| yes | 47737 |
| no | 9224 |

Categorization used for **Drugs**:

|  |  |  |
| --- | --- | --- |
| **Drugs** | **obs** | **new category** |
| never | 37724 | no |
| sometimes | 7732 | yes |
| often | 410 | yes |

**Drugs**after:

|  |  |
| --- | --- |
| **Drugs** | **Obs** |
| no | 37724 |
| yes | 8142 |

Categorization used for **Education**:

|  |  |  |
| --- | --- | --- |
| **education** | **obs** | **new category** |
| college/university | 801 | graduated |
| graduated from college/university | 23959 | graduated |
| graduated from high school | 1428 | graduated |
| graduated from law school | 1122 | graduated |
| graduated from masters program | 8961 | graduated |
| graduated from med school | 446 | graduated |
| graduated from ph.d program | 1272 | graduated |
| graduated from space camp | 657 | graduated |
| graduated from two-year college | 1531 | graduated |
| high school | 96 | graduated |
| law school | 19 | graduated |
| masters program | 136 | graduated |
| med school | 11 | graduated |
| ph.d program | 26 | graduated |
| space camp | 58 | graduated |
| two-year college | 222 | graduated |
| dropped out of college/university | 995 | dropped out |
| dropped out of high school | 102 | dropped out |
| dropped out of law school | 18 | dropped out |
| dropped out of masters program | 140 | dropped out |
| dropped out of med school | 12 | dropped out |
| dropped out of ph.d program | 127 | dropped out |
| dropped out of space camp | 523 | dropped out |
| dropped out of two-year college | 191 | dropped out |
| working on college/university | 5712 | studying |
| working on high school | 87 | studying |
| working on law school | 269 | studying |
| working on masters program | 1683 | studying |
| working on med school | 212 | studying |
| working on ph.d program | 983 | studying |
| working on space camp | 445 | studying |
| working on two-year college | 1074 | studying |

**Education**after:

|  |  |
| --- | --- |
| **Education** | **Obs** |
| graduated | 40745 |
| studying | 10465 |
| dropped out | 2108 |

**Ethnicity**: Ethnicity column was split with space as separator and the first part of the string was used to get the ethnicity value. Similar values are grouped into one category.

|  |  |  |
| --- | --- | --- |
| **ethnicity** | **obs** | **new category** |
| black | 2,008 | black |
| black, | 1,063 | black |
| indian | 1,077 | indian |
| indian, | 119 | indian |
| other | 1,706 | other |
| middle | 811 | other |
| pacific | 717 | other |
| native | 709 | other |
| hispanic | 4,379 | hispanic |
| asian | 6,134 | asian |
| asian, | 2,071 | asian |
| white, | 641 | white |
| white | 32,831 | white |

**Ethnicity**after:

|  |  |
| --- | --- |
| **Ethnicity** | **Obs** |
| white | 33472 |
| asian | 8205 |
| hispanic | 4379 |
| other | 3943 |
| black | 3071 |
| indian | 1196 |

Categorization used for **Job**:

|  |  |  |
| --- | --- | --- |
| **job** | **obs** | **new category** |
| executive / management | 2373 | Business Management |
| sales / marketing / biz dev | 4391 | Business Management |
| hospitality / travel | 1364 | Business Management |
| banking / financial / real estate | 2266 | Business Management |
| transportation | 366 | Business Management |
| construction / craftsmanship | 1021 | Business Management |
| computer / hardware / software | 4709 | Technology |
| science / tech / engineering | 4848 | Technology |
| entertainment / media | 2250 | Entertainment |
| artistic / musical / writer | 4439 | Entertainment |
| student | 4882 | Unemployed |
| retired | 250 | Unemployed |
| unemployed | 273 | Unemployed |
| education / academia | 3513 | Education |
| medicine / health | 3680 | Medical |
| law / legal services | 1381 | law |
| other | 7589 | other |
| rather not say | 436 | other |
| political / government | 708 | govt |
| clerical / administrative | 805 | govt |
| military | 204 | govt |

**Job**after:

|  |  |
| --- | --- |
| **Job** | **Obs** |
| business management/development | 11781 |
| technology | 9557 |
| other | 8025 |
| entertainment | 6689 |
| unemployed | 5405 |
| healthcare | 3680 |
| education | 3513 |
| government | 1717 |
| law | 1381 |

Categorization for **Smokes**:

|  |  |  |
| --- | --- | --- |
| **smokes** | **obs** | **new category** |
| no | 43896 | no |
| sometimes | 3787 | yes |
| when drinking | 3040 | yes |
| yes | 2231 | yes |
| trying to quit | 1480 | yes |

**Smokes**after:

|  |  |
| --- | --- |
| **Smokes** | **Obs** |
| no | 43896 |
| yes | 10538 |

**Data cleaning:**

The raw data was subset to consider only the columns that we are going to use for the model which are:

age, sex, height, income, status, body\_type, diet, drinks,drugs,education, ethnicity, job,smokes. The data used for model has 52318 rows and 13 columns

* Removed the null values from the class variable body\_type. Distribution of the class variable:

|  |  |
| --- | --- |
| **body\_type** | **obs** |
| overweight | 22381 |
| thin | 6193 |
| fit | 23744 |

* Removed the null values from the drinks column which has less than 5% missing values
* Treated all the categorical variables with the most occuring value
* Treated the –1 values in the income column with median values by each job type
* Removed the pets and offspring columns since they had missing values more than 30%
* Rest of the columns were removed because of less relevancy to the class variable and no variance in the values

**Modeling**:

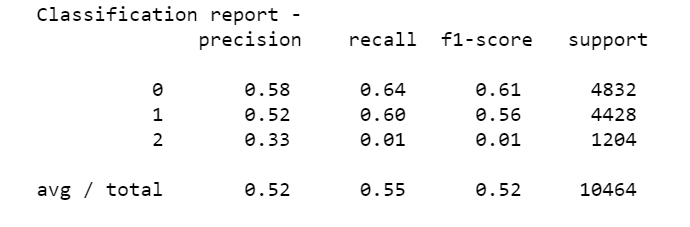
* Converted the class variable using label encoder
* Converted all the categorical variables into binary
* Split the data into training and test using 80-20 split

**Decision Tree**: ran the decision tree in python in loop for depth from 1 to 20 to get the depth with the best accuracy.The accuracy is 55% with max depth of 7

**Random forest classifier**: ran the random forest classifier in python with the 1000 as estimators which gave the accuracy of 50%

**K nearest neighbors**: ran the KNN in loop to get the neighbors with highest accuracy from 1 to 100. The nearest neighbors were 88 and with the maximum accuracy of 54%

Decision tree had the best accuracy hence ran the tree keeping the max depth parameter as 7  on the training data. Ran the confusion matrix on the test data and below is the classification report. The precision and recall for the **Thin**category are the lowest due to uneven distribution of the observations.



**Factors for improvement of accuracy:**

* The data used for analysis was similar to a survey data because there is no way to validate if the customers entered the correct information
* The categorical variables if grouped differently might increase the accuracy
* The availability of the columns such as weight, BMI might increase the accuracy
* The missing value treatment if done differently might lead to an increase in the accuracy