TARTU ÜLIKOOL
arvutiteaduse instituut

KAGGLE EDIBILITY OF MUSHROOMS (B6)

Andrei Tambovtsev B11058
Semjon Kravtšenko B83331

First there was an idea:

- Find intersting dataset
- Apply gathered knowledge
- Make something usefull

№1

We found **mushrooms.csv** from **kaggle**

(https://www.kaggle.com/uciml/mushroom-classification?select=mushrooms.csv)

No missing values
No duplicates
23 columns and 8124 unique rows
Each row = mushroom
Eeach column = mushroom's parameter

Columns provide information such as:

- shape and color of different parts of mushroom
- where mushroom may be found
- how big is the population
- odor
- bruices
- and main column - “class” (edible or poisonous)

All information is categorical.

"Hmm... it is good data set for project" - we thought

And decided to:

- Clear the data (we did not know that this is already clear)
- Train some models to predict edability
- Find strongest correlations between edability and other parameters
- Find intersting correlations among other attributes
- Write a programm for mushroom gatherers based on trained model

№2

1) Import data and check for duplicates and missing values
there was non of those

2) Use hot-encoding to be able to train ML models
map function and dummies turned our 23 columns into 114 columns

SVM

We did not even try SVM because encoded dataset is too high dimensional

3) Use apriori algorithm and visualized decision tree to find interesting connections between attributes
we found, that if a mushroom has no odor and no green spore prints - it is probably eadible.
while this rule may be useful for some novice shroomers, we had decided to not stop there, since those 2 attributes are in our opinion somewhat hard to measure.

3) Run Lasso and Ridge regression with best alphas and watch, which attributes had high coefficients
regression does not really help with classification, but model still search for useful corellations and try to give values as close as possible to 1 (edible) or 0 (poisonous)
Both regression models was agree on 'odor' and 'spore color' parameters, but Lasso prefers 'stalk color' as 3rd most valuable parameter, Ridge prefered 'ring type'

4) Run KNN and RandomForest models, trained only on parameters with high coefficients from Lasso and Ridge regression. This led us to model with 99.7%(on 3 valuable parameters from Lasso) and 99.5% (on 3 valuable parameters from Ridge) accuracy.

```
Based on 'odor', 'spore-print-color' and 'stalk-color-below-ring'
Knn prediction score:  0.9975381585425899
Random Forest prediction score:  0.9975381585425899
```

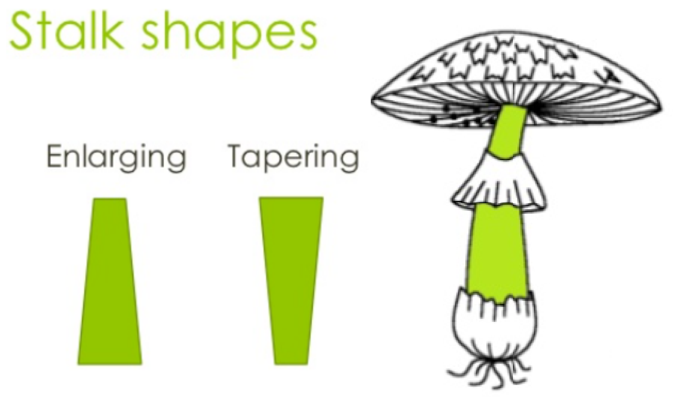
```
Based on 'odor', 'spore-print-color' and 'ring-type'
Knn prediction score:  0.9950763170851797
Random Forest prediction score:  0.9950763170851797
```

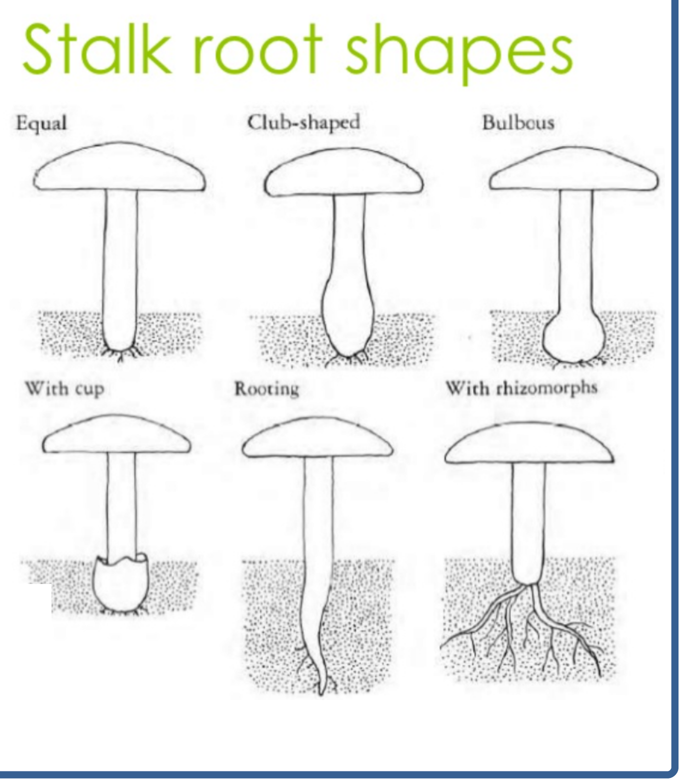
which was pretty good, but not very usefull for mushroom gatherers. Odor is too subjective and spore color too hard to notice

5) Return to original data set and drop columns with parameters that are hard to define.
such as 'odor', 'spore print color', 'population', etc. 13 columns left

6) Run Lasso and Ridge regression with parameters that are left after cleaning
models were agree on 3 parameters, that are very easy to define, that are:

Gill colors
brown, orange,
white, yellow

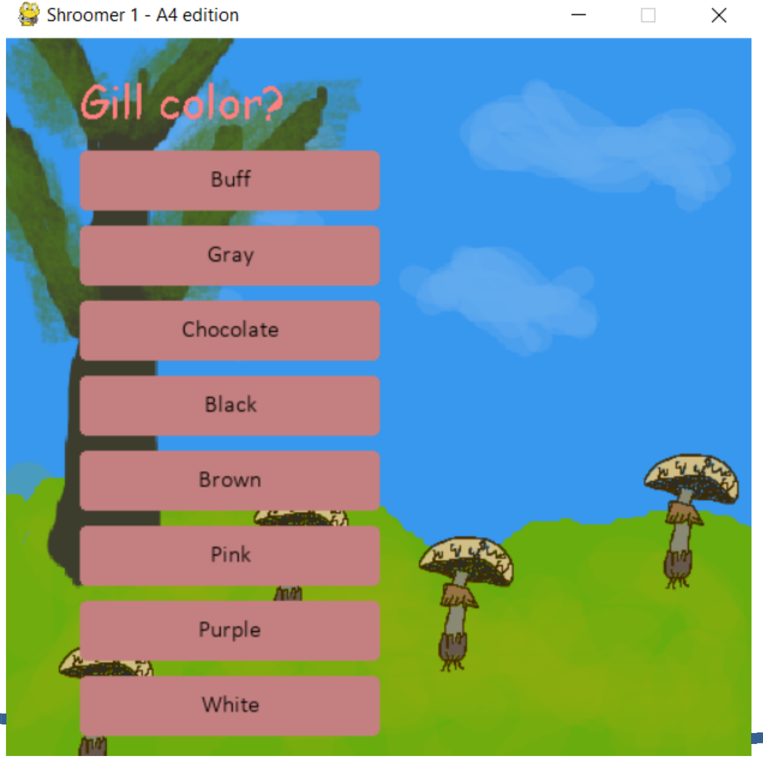




And prediction score reduces by less than 5%

```
Based on 'stalk-shape', 'gill-color' and 'stalk-root'
Knn prediction score:  0.9533932951757972
Random Forest prediction score:  0.9533932951757972
```

№3



All tests led us to developing simple but usefull programm writen on python with use of pygame Programm asks questions about mushroom parameter and then predicts edability of mushroom using trained model