Alex Jonauskas

Data Mining Final Project, 11/31/2022

1. Research objectives:
   1. We would like to discover if there is/are genetic mutations that increase your likelihood of having cancer.
   2. If so, what genetic mutations would do so?
   3. Can we come up with a reasonably accurate method of determining if a patient has cancer based on genetic mutations?
2. Key Results
   1. Our highest accuracies occur when these mutations are used often in our random forest:
      1. BRAF\_GRCh38\_7:140753336-140753336\_Missense-Mutation\_SNP\_A-A-T
      2. KRAS\_GRCh38\_12:25245350-25245350\_Missense-Mutation\_SNP\_C-C-A\_C-C-G\_C-C-T\_C-G-G\_C-A-A
      3. XYLT2\_GRCh38\_17:50356606-50356606\_Frame-Shift-Del\_DEL\_C-C—

Which would suggest that that these mutations may be more prevalent in cancer patients.

* 1. We can generate random forests to predict if a patient will have cancer with an average accuracy of 69% per forest, seeing a high of 76%

1. Summary of Methods (in order)
   1. Reduced the restriction of each decision tree from sqrt(n) many features to n/2 to n\*.75 many features, resulting in a rough improvement of about 10% accuracy.
   2. Played with the number bootstraps selected. I noticed ~3% improvement from 10 to 25 bootstraps and a <1% improvement from 25 to 50 but a much slower runtime
   3. Removed mutations that only 2, 3, 4, and 7 patients have resulting in a 3%, 2%, 2%, and -2% increase in accuracy, respectively
   4. Increased decision trees to depth 3 and got a sharp decrease in accuracy (about -15%), I would speculate that this is because there are only a few relevant mutations in this dataset
   5. Removed Samples with less than 2 and 5 mutations which lowered the accuracy 5% and 6% respectively.

ACC 0.703297

Sens 0.550000

Spec 0.823529

Prec 0.709677

Miss Rate 0.450000

FDR 0.290323

FOR 0.300000

For the numbers above BRAF was selected 18 times, KRAS 15 times, and XYLT2 7 times. The rest of the features were used at most 4 times (but mostly once). This is always the general structure of our forest when we end up with accuracies over 70%.

These numbers and the best numbers in general came from using n\*.75 features in each bootstrap, 25 bootstraps, and removing features that less than 3 patients had. The depth 3 decision tree and removing patients made the accuracy worse and were not used

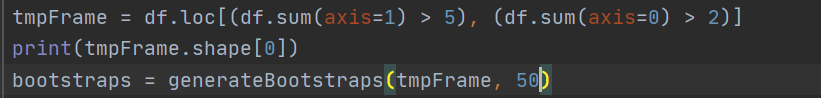
1. Discussion
   1. Test cancer patients for the 3 mutations above, recording what kind of cancer they have, and recording what percentage of them have each one as well as combinations of them
   2. Identify which cancer may be connected to each mutation and use our random forest to identify if someone may have those cancers with a higher accuracy.
   3. A quick google search seems to imply that the BRAF mutation has a correlation to melanoma and colon cancer while the KRAS mutation is correlated to lung cancer so the random forest

Deep Classification Screenshots:

The following 2 screenshots demonstrate how I can generate more or less bootstraps. In the 2nd screenshot I generate 50 of them.

Text

Description automatically generated



The following 7 screenshots demonstrate how I implement depth 3 decision trees into my bootstraps/forest classifier. To adjust it I have a variable called depth that I can change from 2 to 3.

Text

Description automatically generated

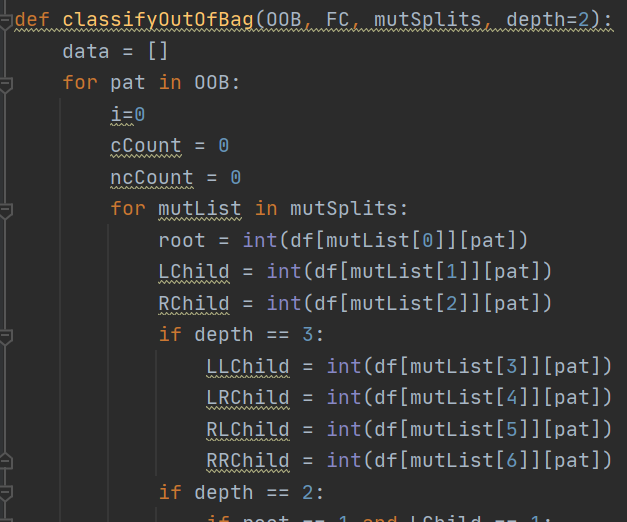
Text

Description automatically generated Text

Description automatically generated

Text

Description automatically generated

A screenshot of a computer

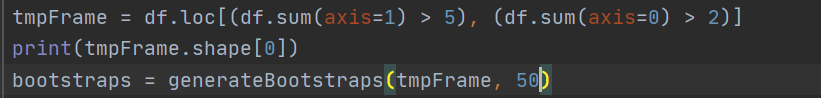
Description automatically generated with medium confidence

Text

Description automatically generated

Optimization Screenshots

The first line in the following screenshots demonstrates how I can restrict my dataset to only mutations with more than 2 patients and then patients with more than 5 mutations. This seems to be optimized when the 5 is actually -1 (keeping all samples in the dataset).



The following screenshot shows how I can adjust the number of mutations in each bootstrap. The first line in the for loop sets the bootstrap as a random sample of the data frame passed choosing each row (patient) with replacement. Then the commented line makes the bootstrap a sample of itself of size sqrt(number of columns). The line following shows the optimized choice of choosing 3/4s of the features each time.

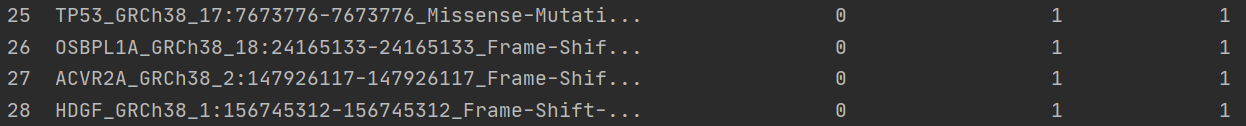
Text

Description automatically generated

Output of Program

Text

Description automatically generated



A screenshot of a computer

Description automatically generated with medium confidenceGraphical user interface, text

Description automatically generated