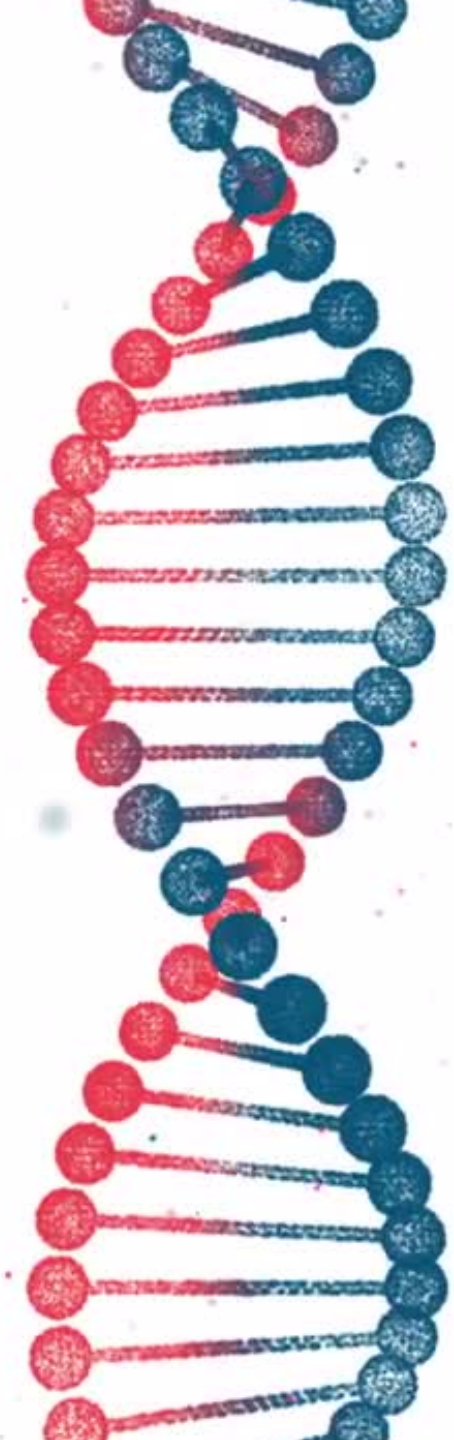
Genetic Variant Classification (Classification)

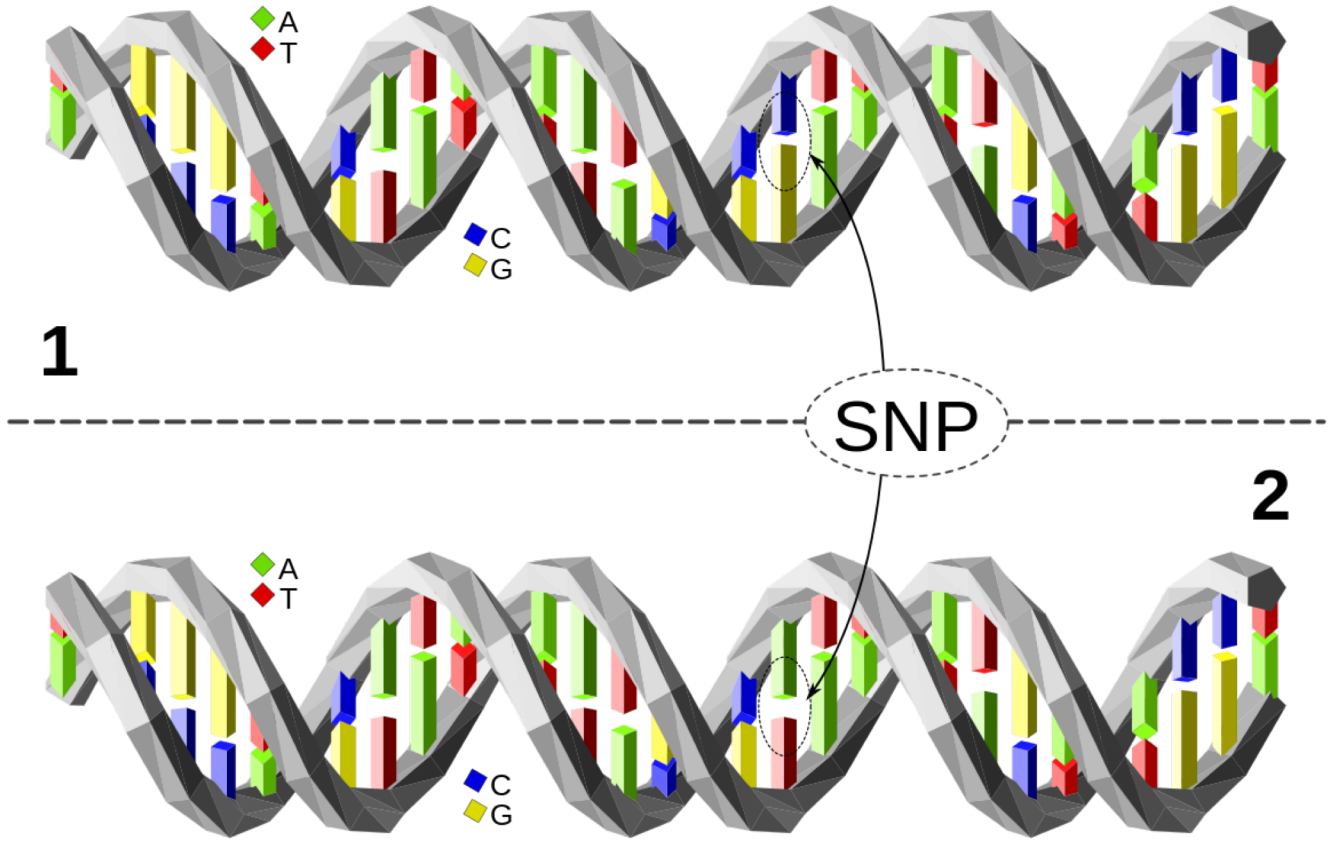
Nick Kinnaird

2/10/21



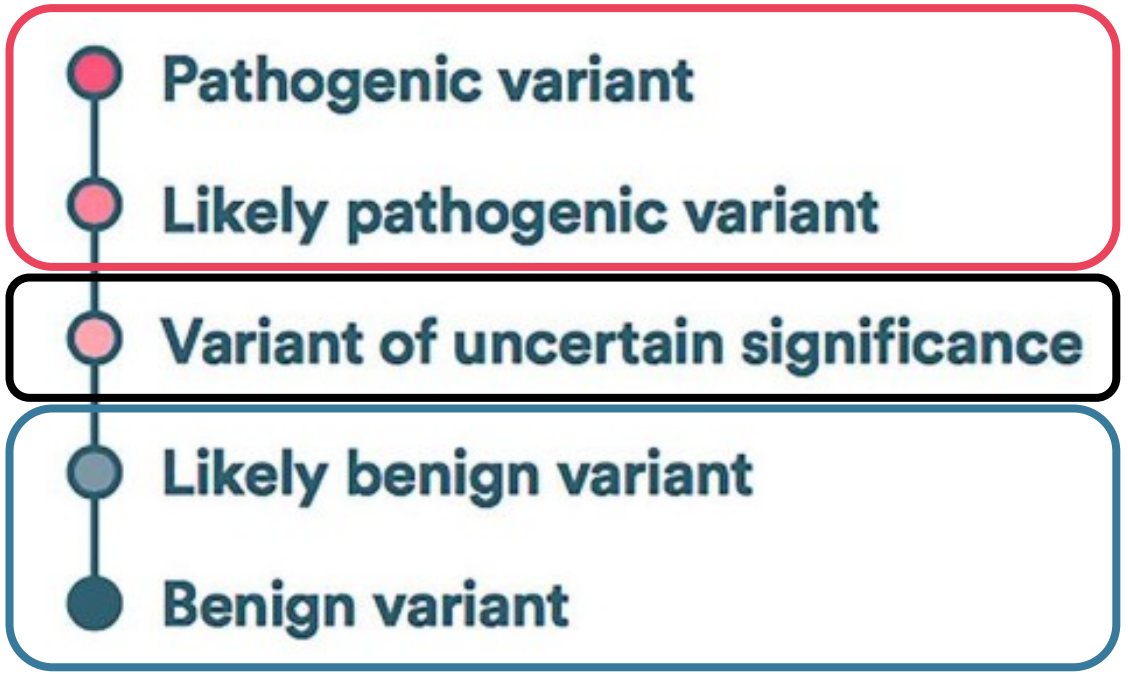
DNA Primer

- We all have DNA and genes
- Some of those genes are variants
- These variants might be common (eg. SNP) or rare (mutations)
- Variants might have little to no effect, or a negative effect (sometimes but far less often a positive effect)



Gene Classification

- Genetic variants are classified at labs (usually manually) in one of 5 different ways in terms of clinical significance
- These are split into 3 categories
- When results from separate labs are classified as different categories, they are said to be “conflicting”



Purpose

- By identifying which gene variants are likely to be classified in a conflicting manner or not:
  - Biologists and labs can better identify those variants which require further study and lab testing
  - Physicians can better respond to lab results for patient treatment

The Data

- Kaggle Dataset: https://www.kaggle.com/kevinarvai/clinvar-conflicting/
  - ClinVar – public archive of gene reports

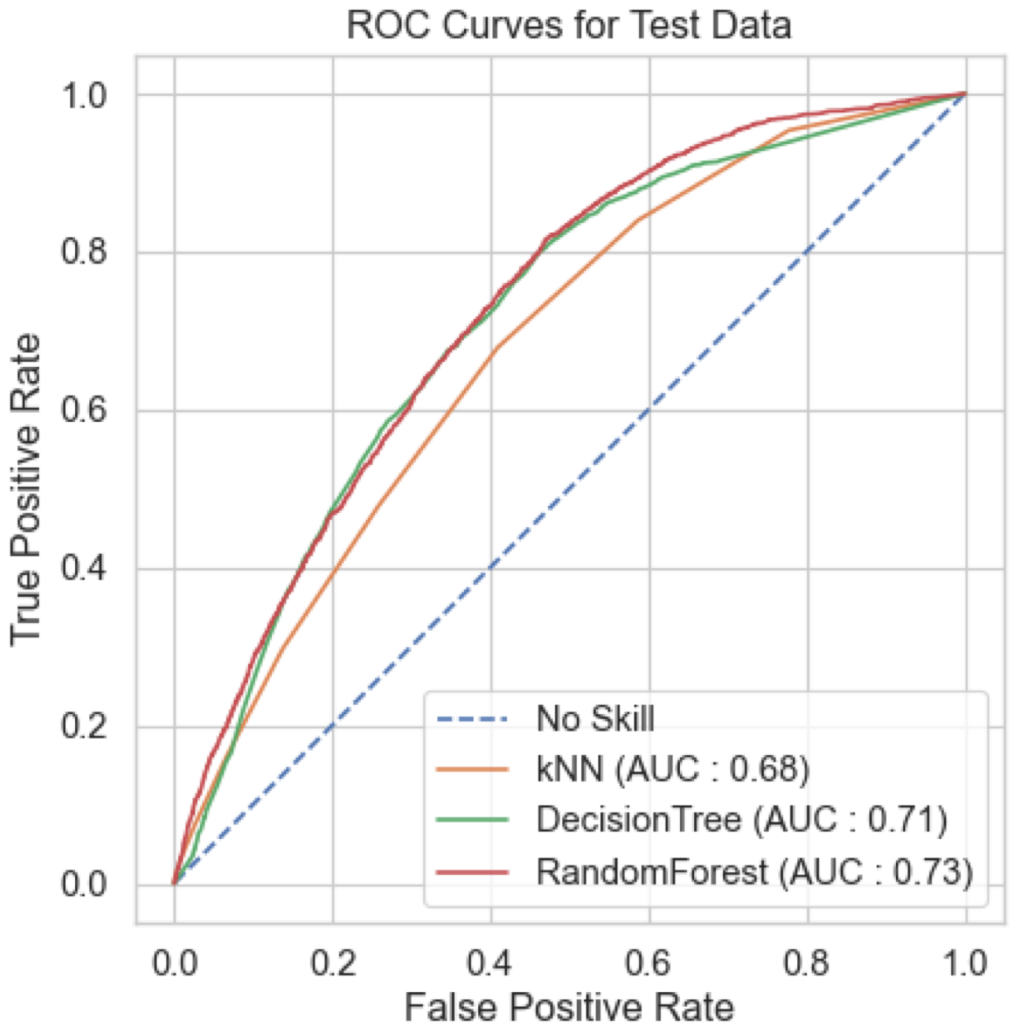
- **Prediction Target:** conflicting (1) or non-conflicting (0)

- ~60,000 entries of data
- 1:3 conflicting to non-conflicting

- Features I chose:
  - Categorical: Name of the gene, variant type, impact of the variant
  - Numeric: Variant frequency, ‘deleteriousness’ score, ‘loss-of-function’ score

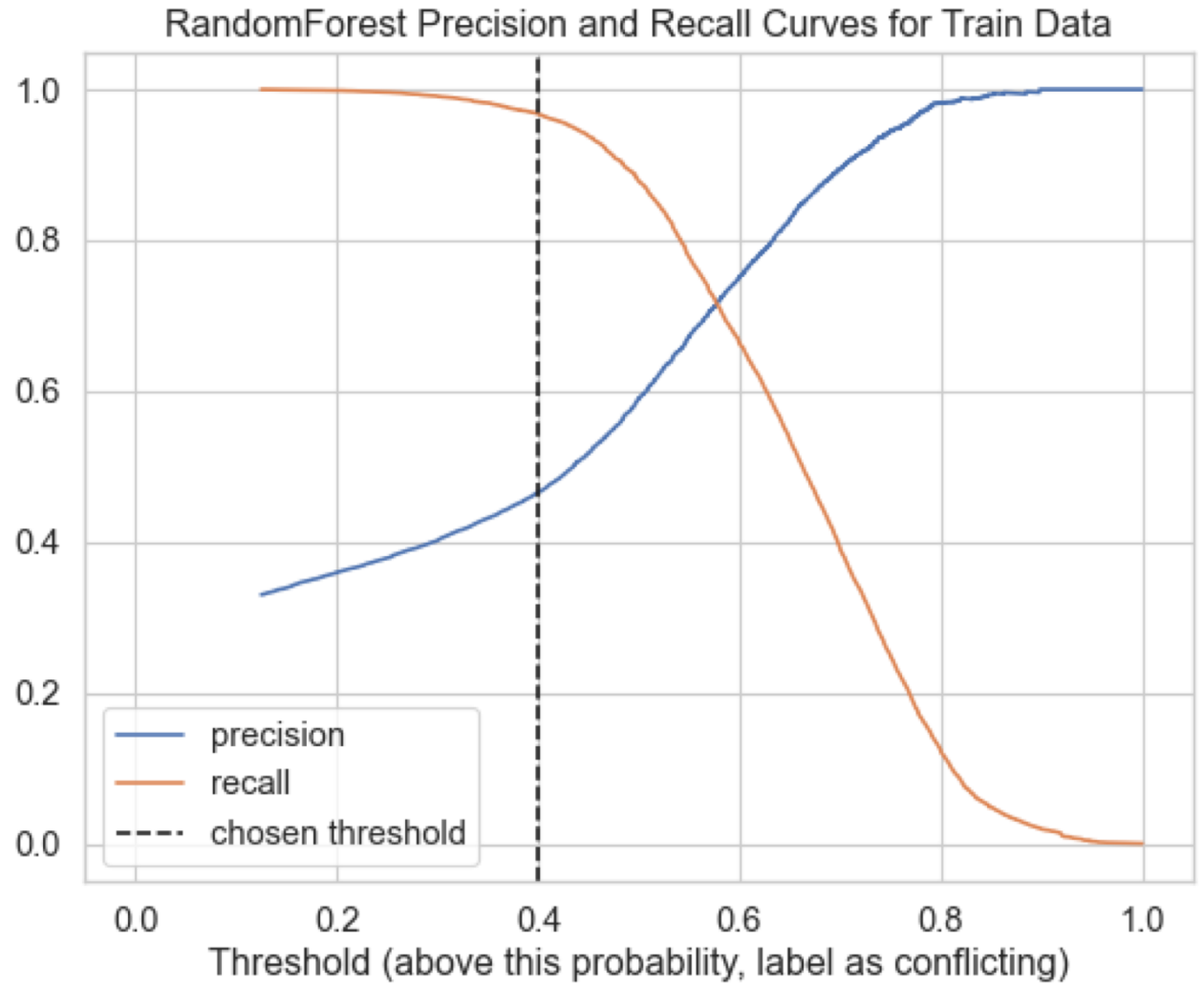
Modeling

- Tried kNN, DecisionTree, and RandomForest
- Used GridSearchCV to tune parameters
- Ultimately chose the RandomForest since it performed the best (and most consistently) on the training and validation data

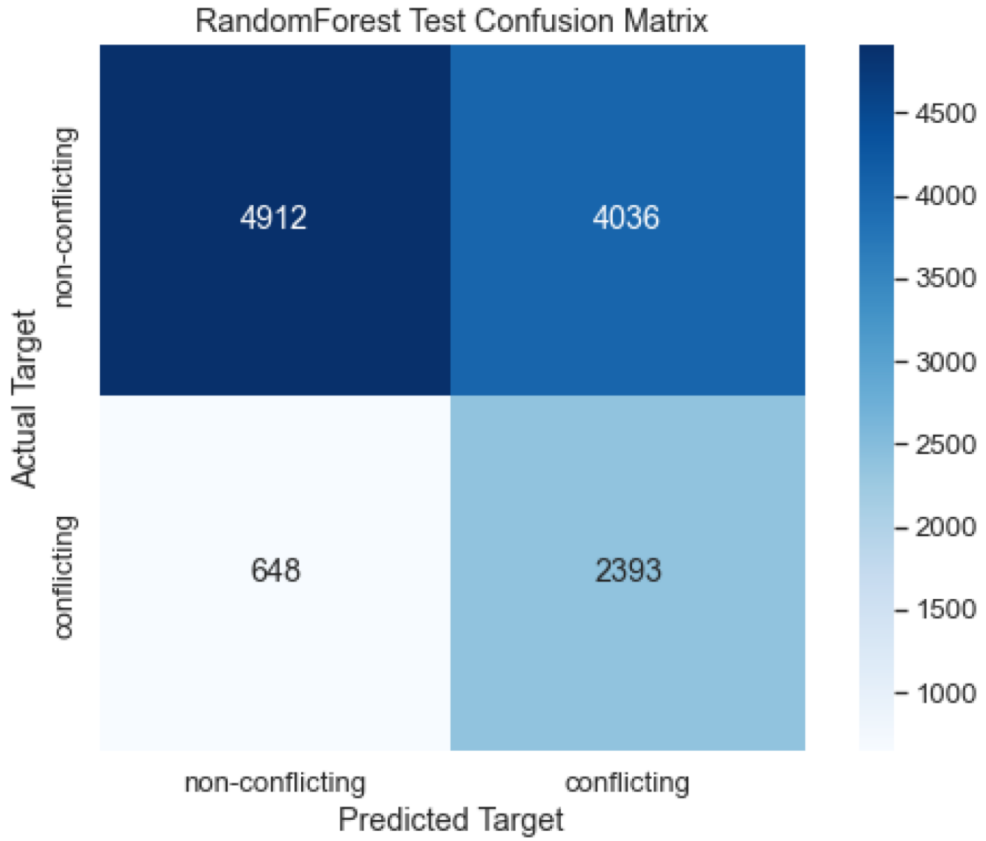


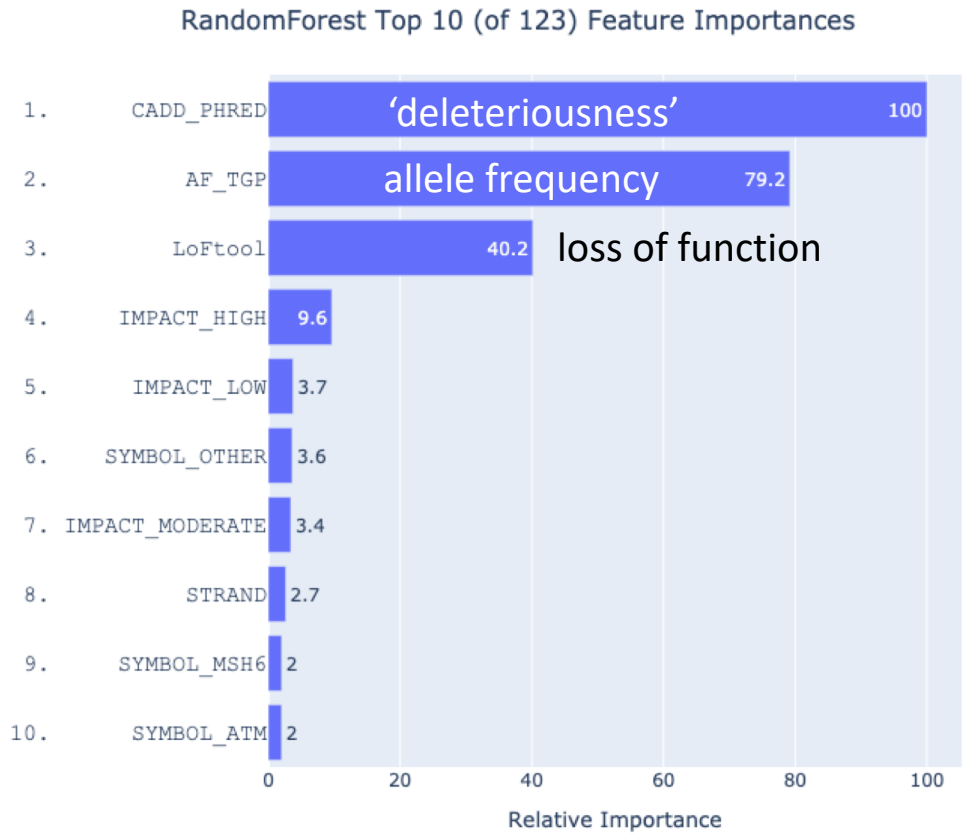
Modeling

- Optimized on recall without entirely sacrificing precision
- Chose threshold of 0.4
- That leaves me with more false positives but it's worth it for the increased number of true positives



Results





‘deleteriousness’

allele frequency

loss of function

Recall: 0.78

Most important feature: CADD_PHRED

Future Work

- There was one or two features which encoded some interesting information which could potentially be used

- The target could be changed from conflicting vs non-conflicting to a multi-label classification problem on the pathogenicity of the variant
  - This would be a more useful application
  - (Indeed there are companies who are working on this stuff)

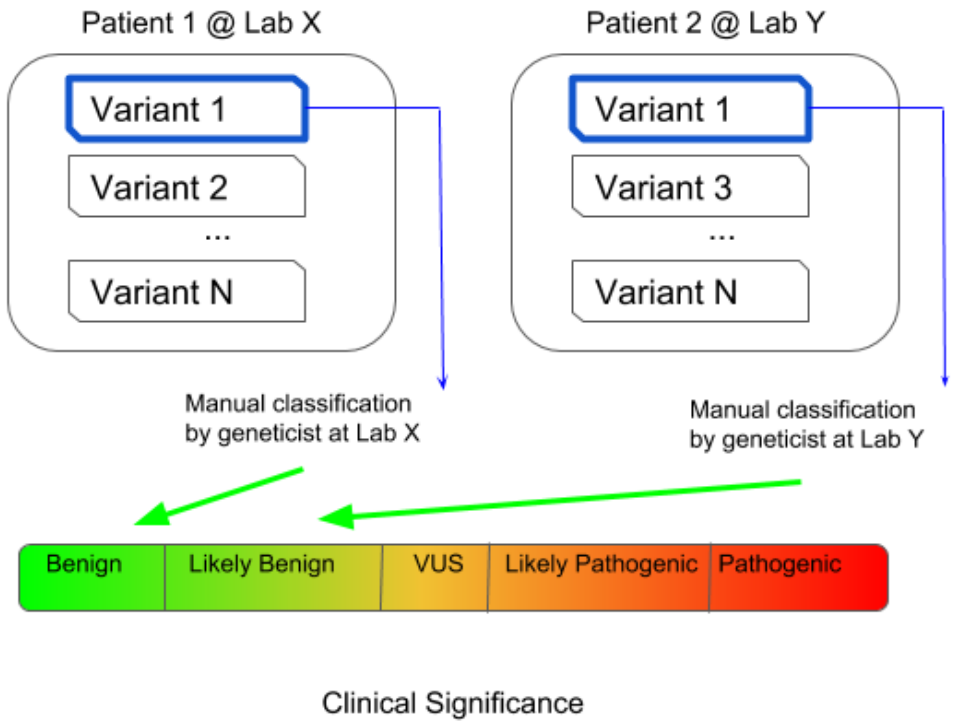
End

"The human genome is only on the order of a gigabyte of data, which is a tiny little database. If you take the entire living biosphere…about one petabyte…that's still very small compared with Google or the Wikipedia. And somehow mother nature manages to create…this incredibly rich environment with this amazingly small amount of data."

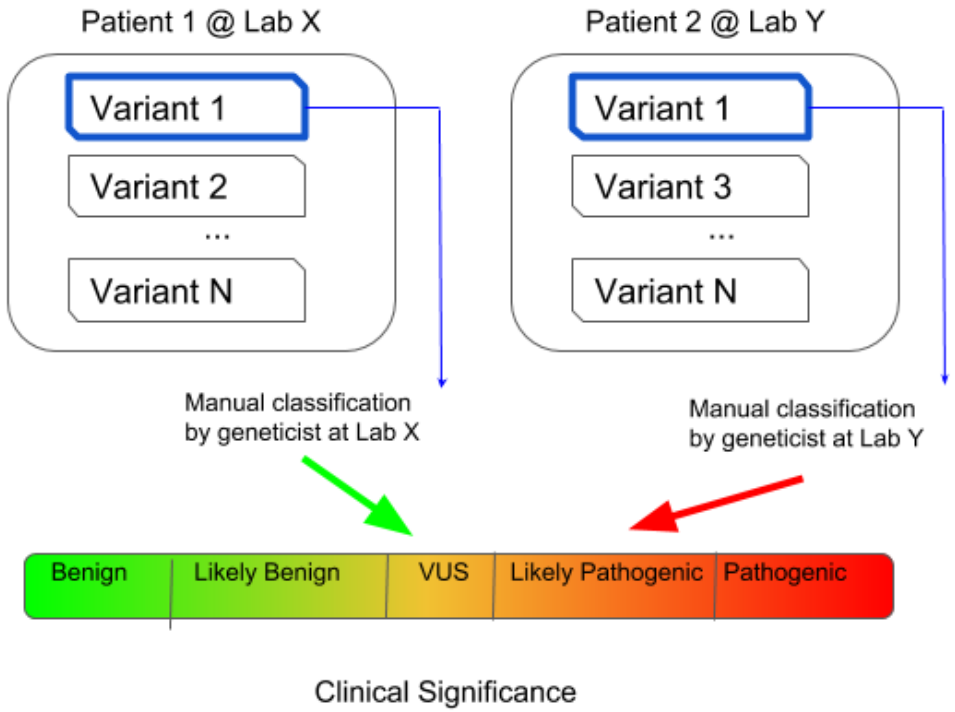
- Freeman Dyson

Appendix

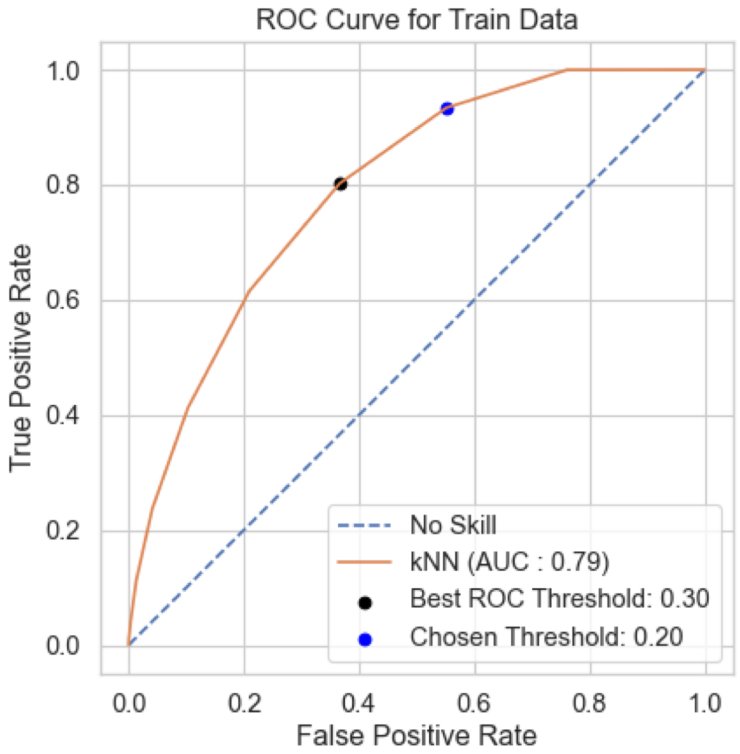
Concordant Variant Classification - **Class: 0**

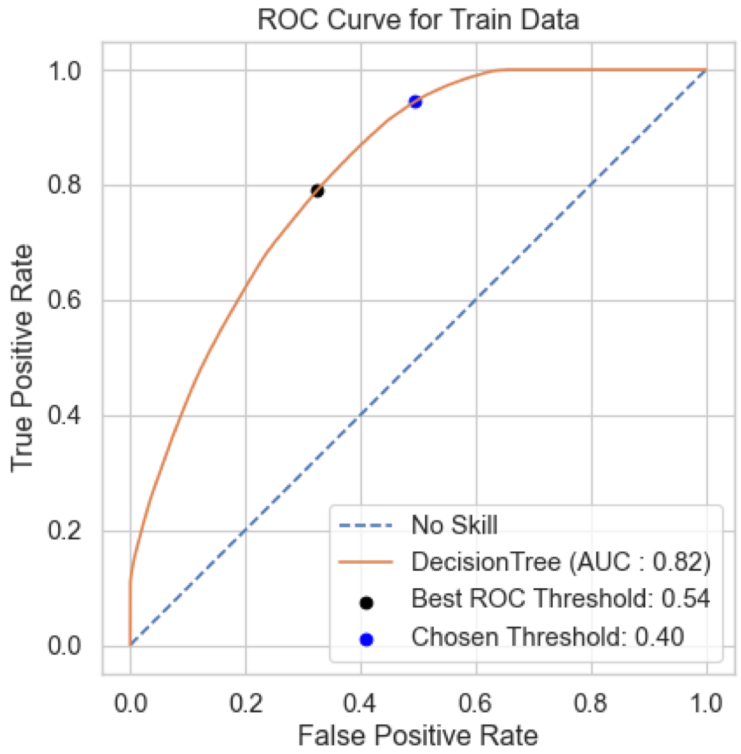


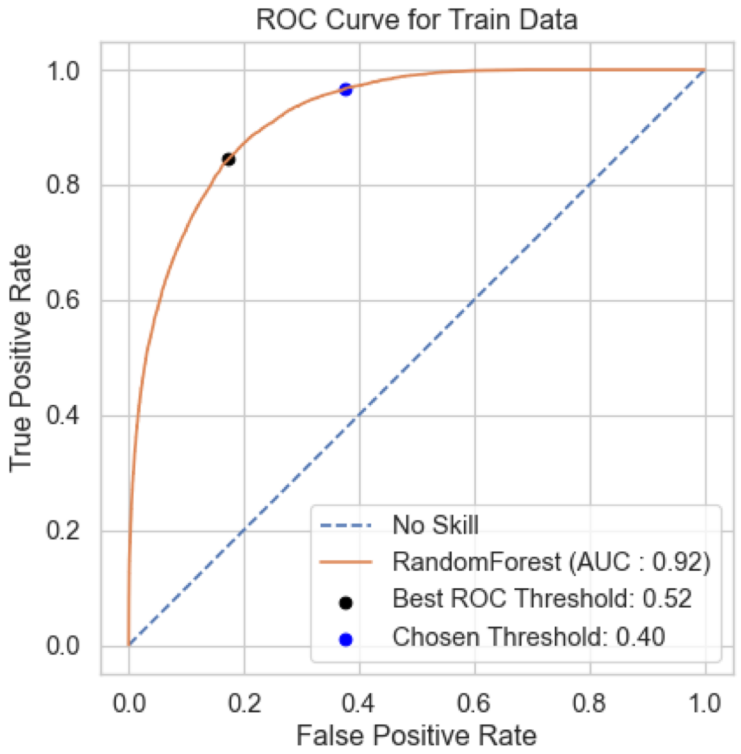
Conflicting Variant Classification - **Class: 1**

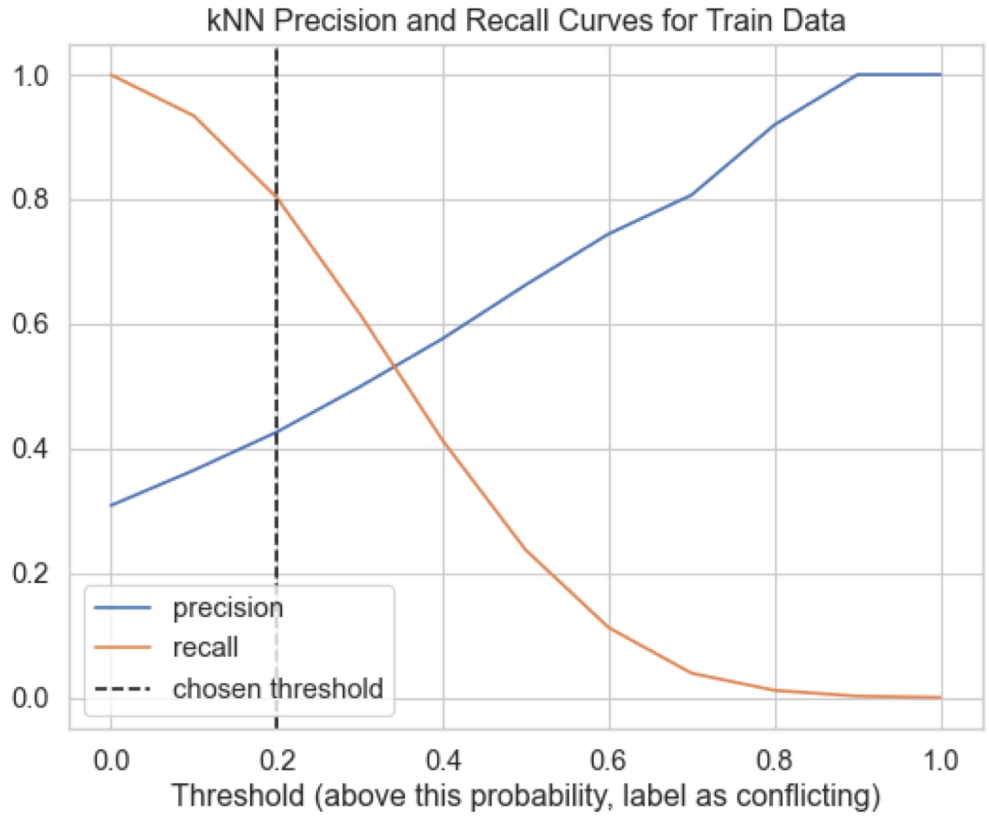


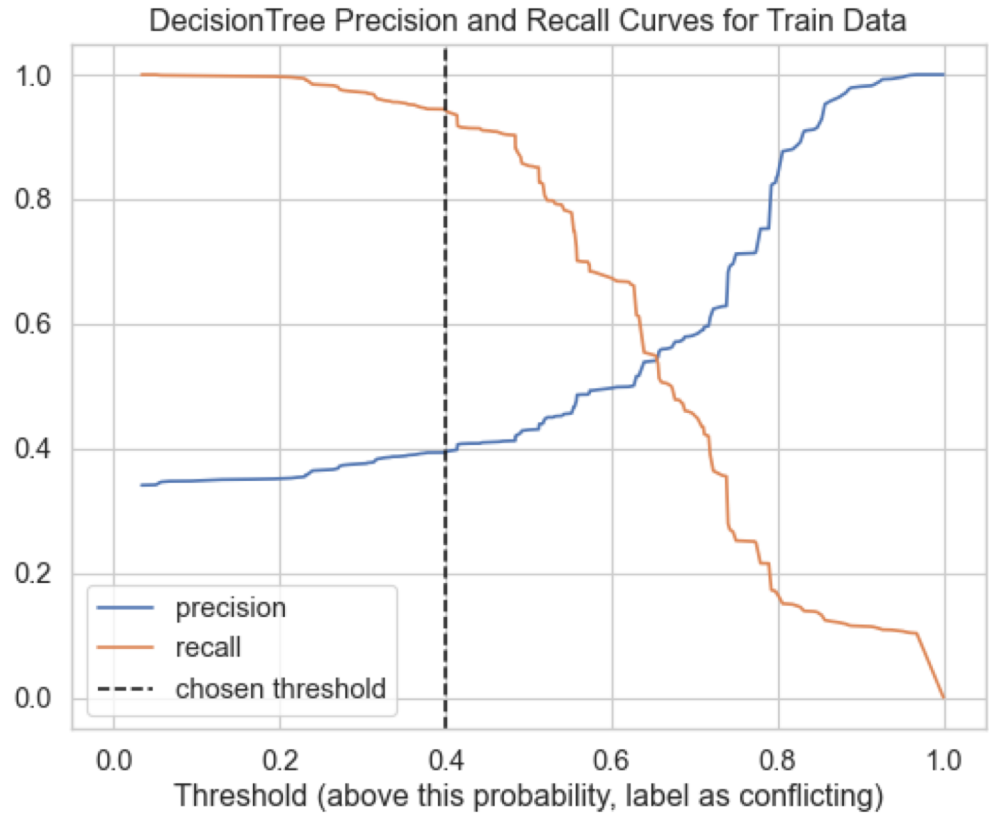
- knn_model = KNeighborsClassifier(n_neighbors=10)
- dt_model = DecisionTreeClassifier(max_depth=25, max_features=20, random_state=121, class_weight={0:1,1:3})
- rf_model = RandomForestClassifier(n_estimators=10, max_depth=25, max_features=20, random_state=121, class_weight={0:1,1:3})

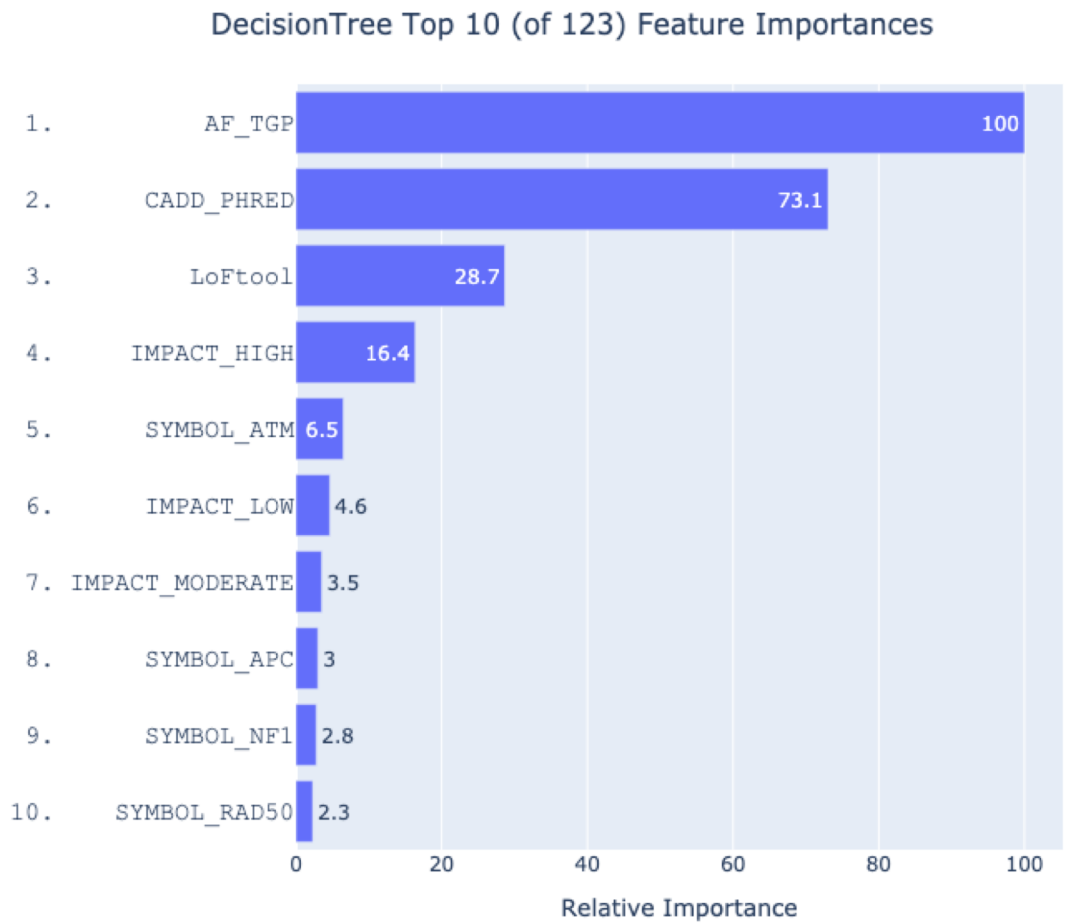

ROC Curve for Train Data
True Positive Rate
False Positive Rate
No Skill
kNN (AUC : 0.79)
Best ROC Threshold: 0.30
Chosen Threshold: 0.20

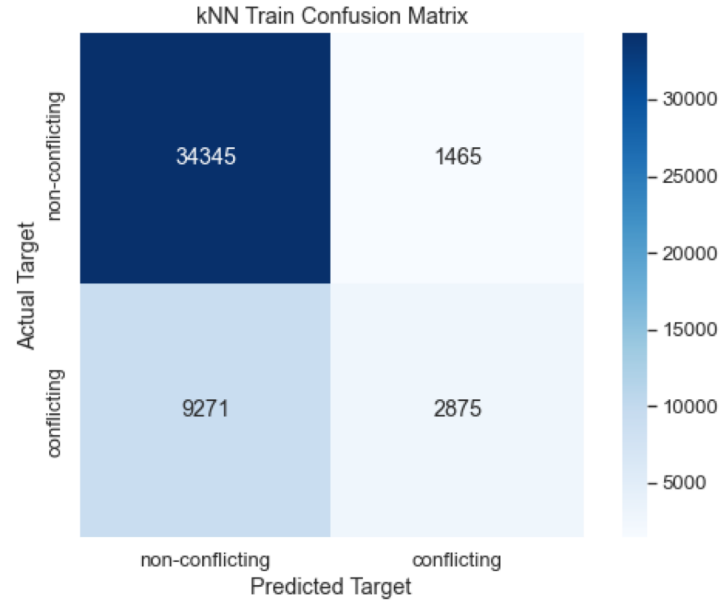

ROC Curve for Train Data
True Positive Rate
False Positive Rate
No Skill
DecisionTree (AUC : 0.82)
Best ROC Threshold: 0.54
Chosen Threshold: 0.40

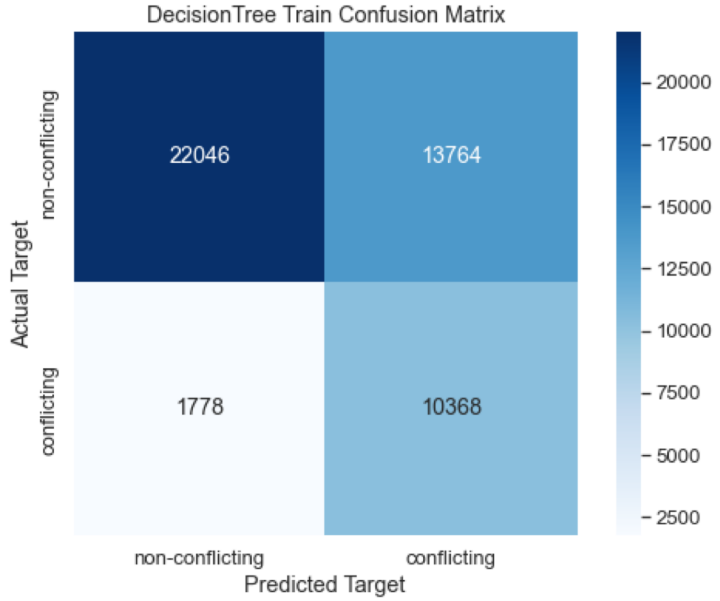

ROC Curve for Train Data
True Positive Rate
False Positive Rate
No Skill
RandomForest (AUC : 0.92)
Best ROC Threshold: 0.52
Chosen Threshold: 0.40

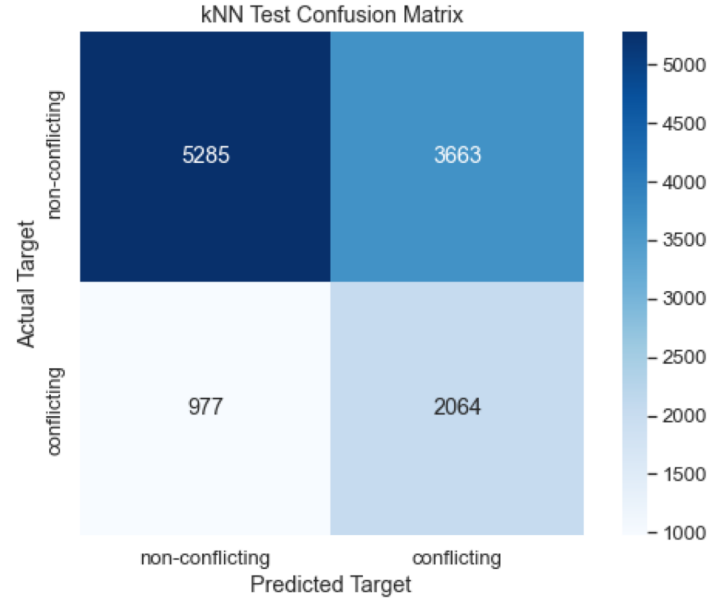

kNN Precision and Recall Curves for Train Data
1.0
0.8
0.6
0.4
0.2
0.0
precision
recall
chosen threshold
0.0
0.2
0.4
0.6
0.8
1.0
Threshold (above this probability, label as conflicting)

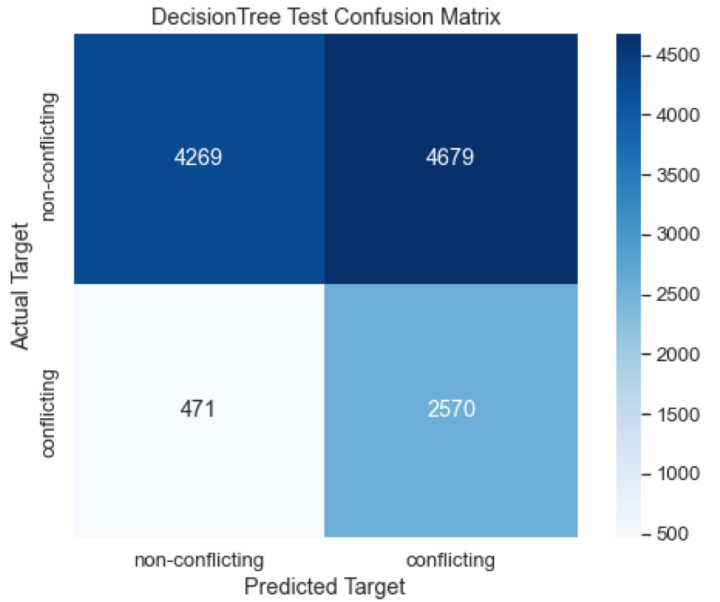

DecisionTree Precision and Recall Curves for Train Data
1.0
0.8
0.6
0.4
0.2
0.0
precision
recall
chosen threshold
0.0
0.2
0.4
0.6
0.8
1.0
Threshold (above this probability, label as conflicting)

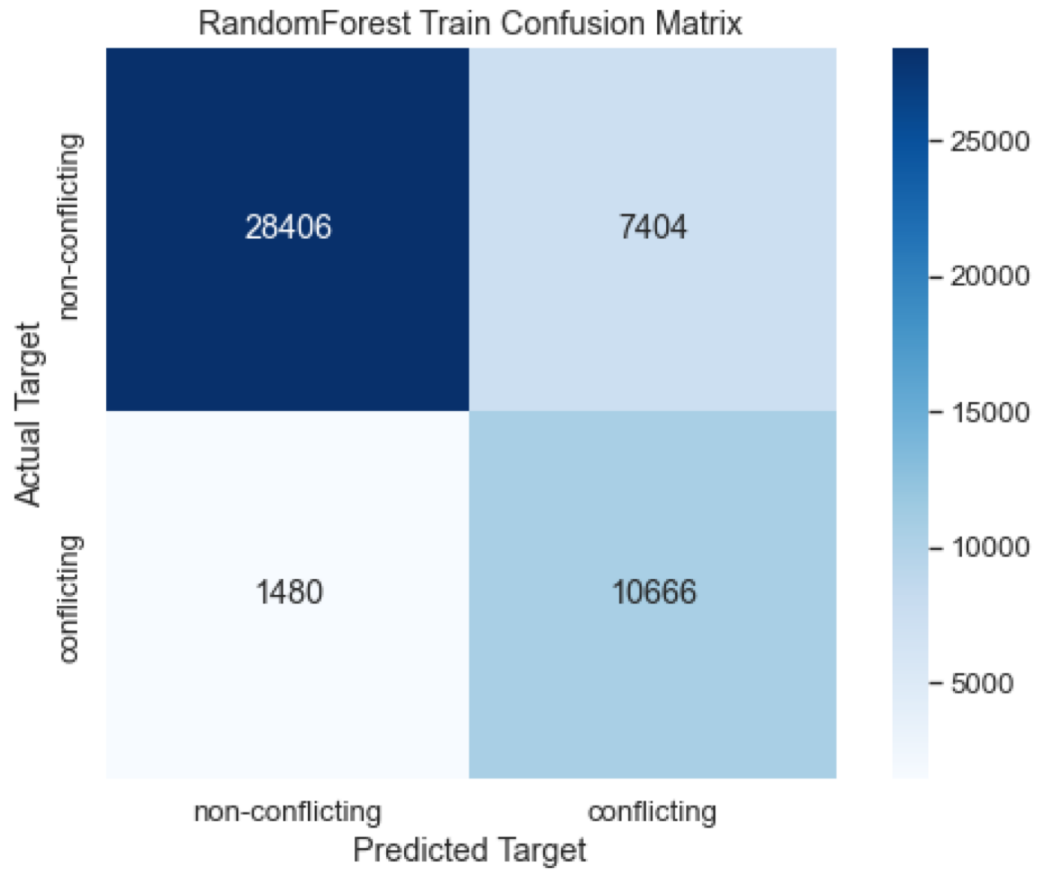

DecisionTree Top 10 (of 123) Feature Importances
1. AF_TGP
2. CADD_PHRED
3. LoFtool
4. IMPACT_HIGH
5. SYMBOL_ATM
6. IMPACT_LOW
7. IMPACT_MODERATE
8. SYMBOL_APC
9. SYMBOL_NF1
10. SYMBOL_RAD50
100
73.1
28.7
16.4
6.5
4.6
3.5
3
2.8
2.3
0
20
40
60
80
100
Relative Importance


kNN Train Confusion Matrix
34345
1465
9271
2875
non-conflicting
conflicting
Actual Target
non-conflicting
conflicting
Predicted Target
30000
25000
20000
15000
10000
5000


DecisionTree Train Confusion Matrix
22046
13764
1778
10368
non-conflicting
conflicting
Actual Target
non-conflicting
conflicting
Predicted Target
20000
17500
15000
12500
10000
7500
5000
2500


kNN Test Confusion Matrix
5285
3663
977
2064
non-conflicting
conflicting
Actual Target
non-conflicting
conflicting
Predicted Target
5000
4500
4000
3500
3000
2500
2000
1500
1000


DecisionTree Test Confusion Matrix
4269
4679
471
2570
non-conflicting
conflicting
Actual Target
non-conflicting
conflicting
Predicted Target
4500
4000
3500
3000
2500
2000
1500
1000
500


RandomForest Train Confusion Matrix
Actual Target
non-conflicting
conflicting
28406
7404
1480
10666
25000
20000
15000
10000
5000
non-conflicting
conflicting
Predicted Target

- Categorical features
  - SYMBOL – Name of the gene
  - CLNVC – Variant type
  - IMPACT – Impact of the variant
- Numeric features
  - AF_TGP – Frequency of allele
  - CADD_PHRED – ‘Deleteriousness’ score
  - LoFtool – Loss of function score
  - Strand – Forward or backward DNA strand