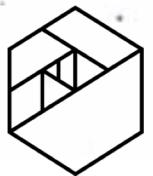


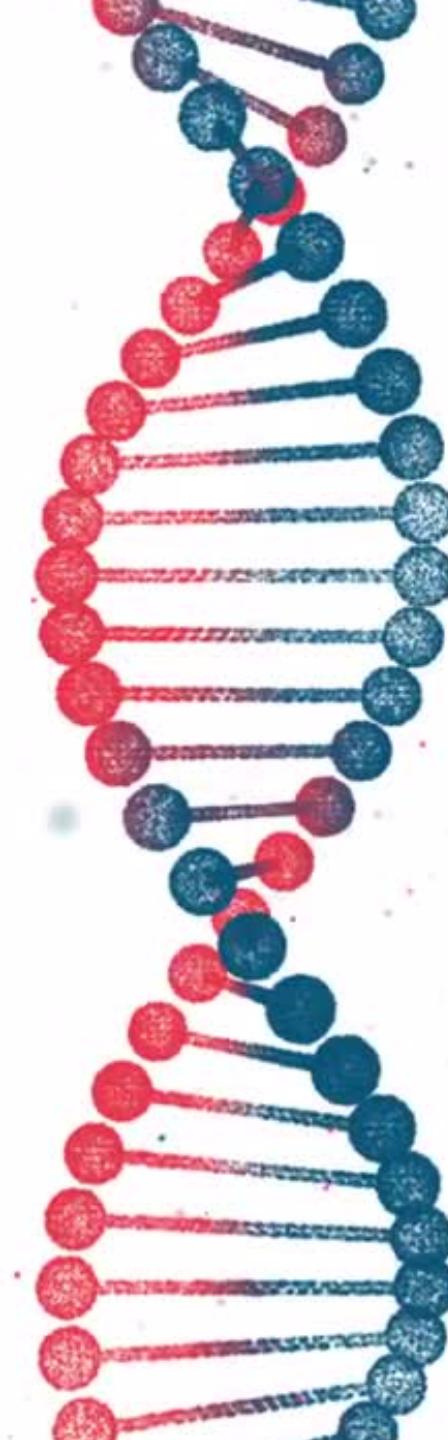
# Genetic Variant Classification (Classification)

Nick Kinnaird

2/10/21



**METIS** Winter 2021 Cohort

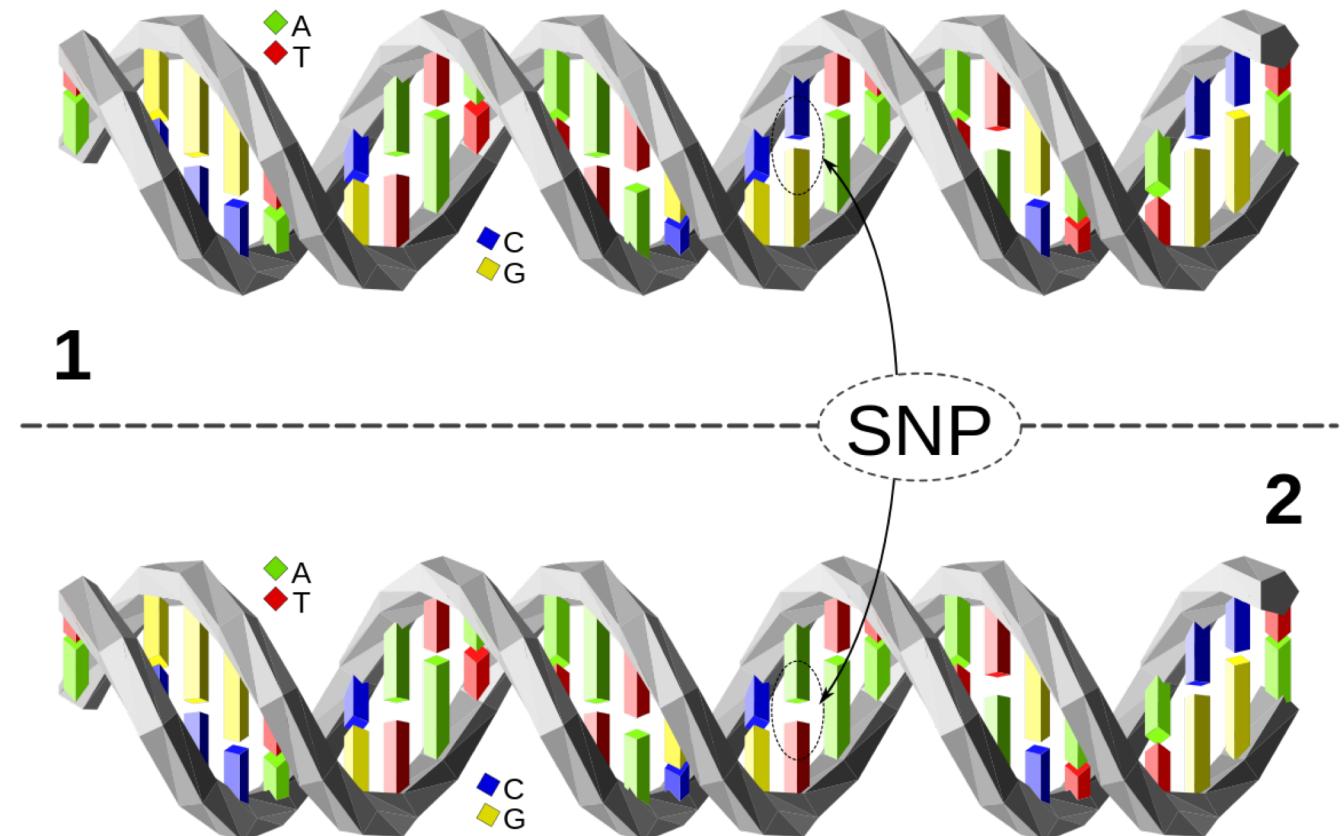




# DNA Primer



- 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)

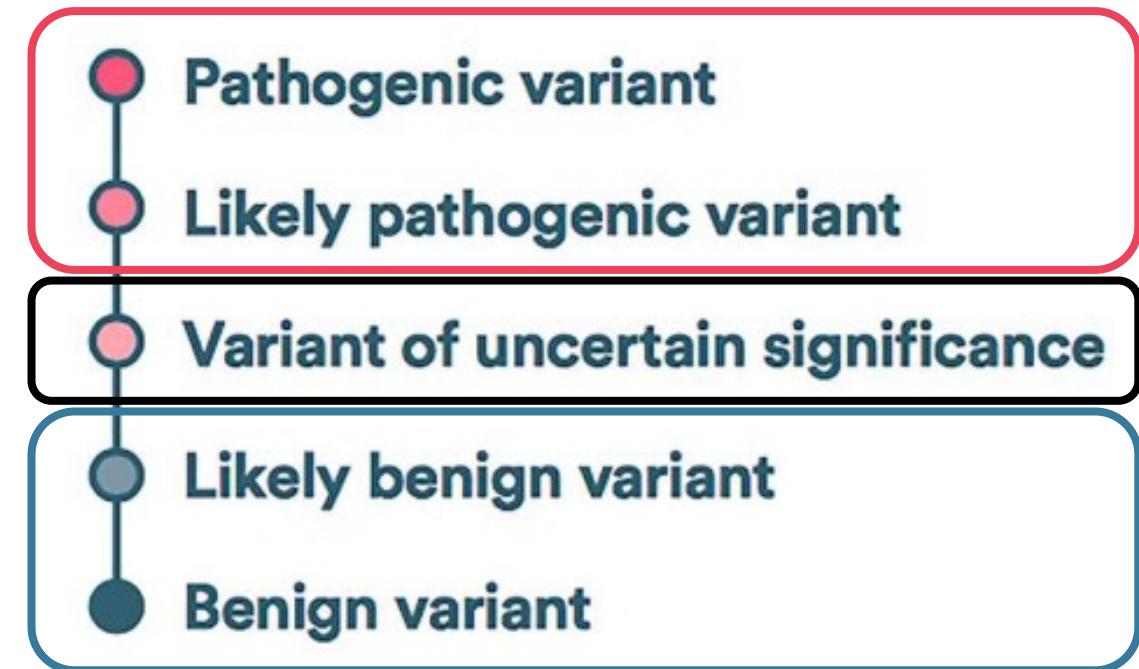




# Gene Classification

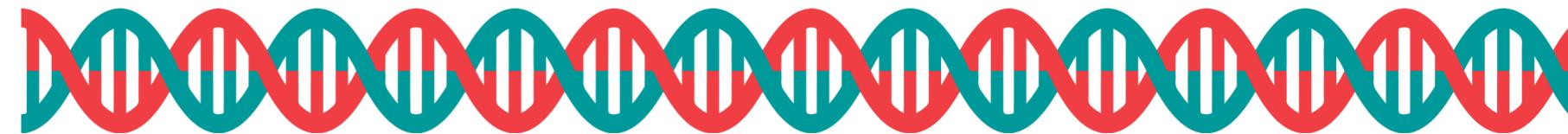


- Genetic variants are classified at labs (usually manually) in one of 5 different ways in terms of clinical significance
- These are split into 3 categories
- When results from separate labs are classified as different categories, they are said to be “conflicting”





# Purpose



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



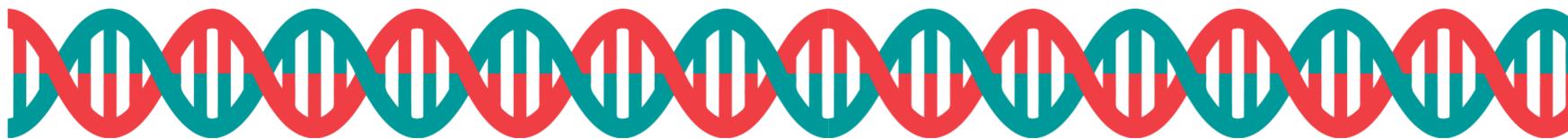
# The Data



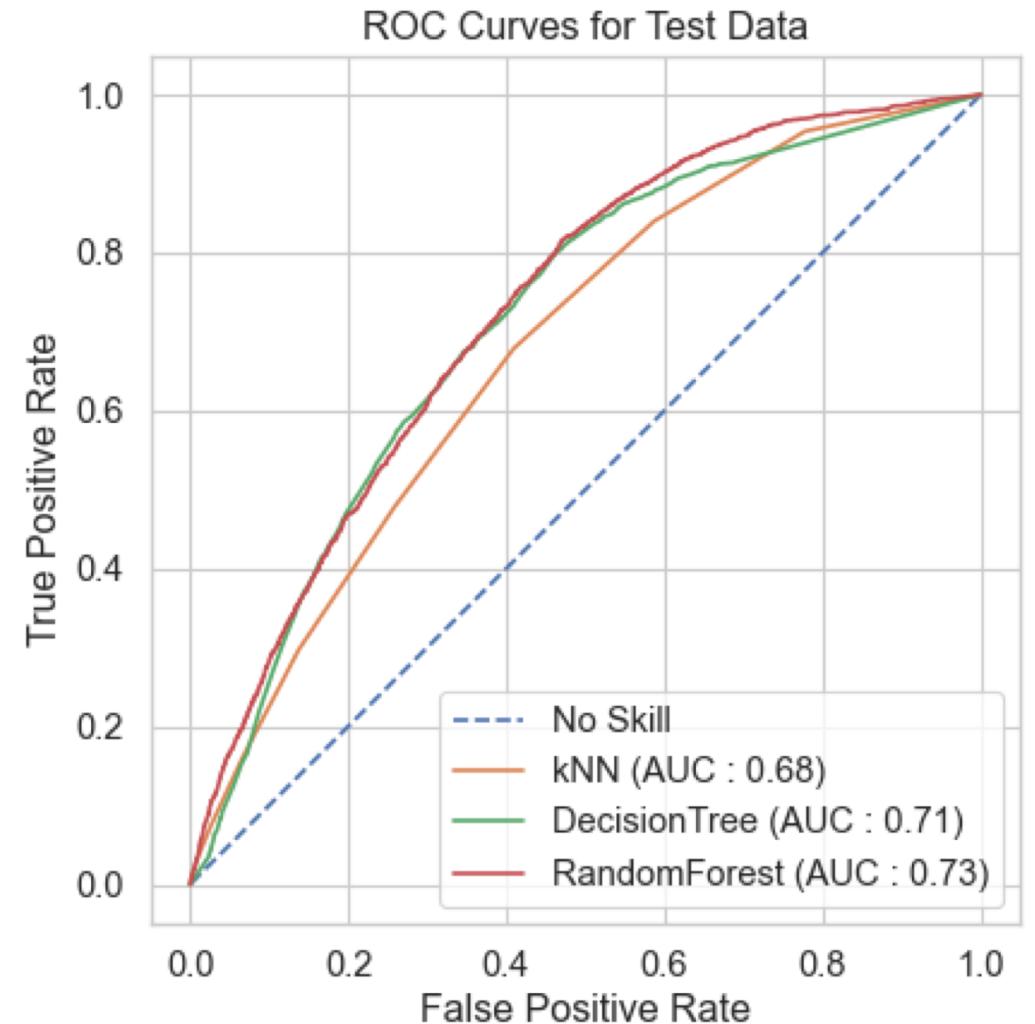
- Kaggle Dataset: <https://www.kaggle.com/kevinarvai/clinvar-conflicting/>
  - ClinVar – public archive of gene reports
- **Prediction Target:** conflicting (1) or non-conflicting (0)
- ~60,000 entries of data
- 1:3 conflicting to non-conflicting
- Features:
  - Categorical: Name of the gene, variant type, impact of the variant
  - Numeric: Variant frequency, ‘deleteriousness’ score, ‘loss-of-function’ score



# Modeling

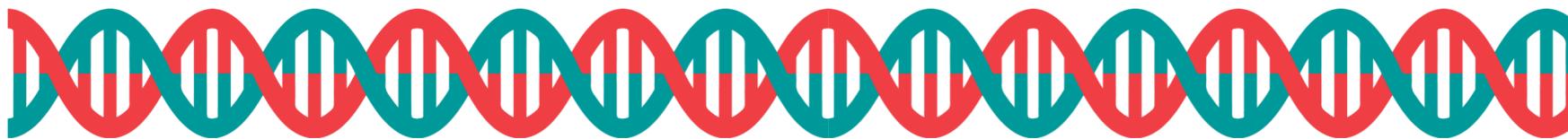


- Tried kNN, DecisionTree, and RandomForest
- Used GridSearchCV to tune parameters
- Ultimately chose the **RandomForest** since it performed the best (and most consistently) on the training and validation data

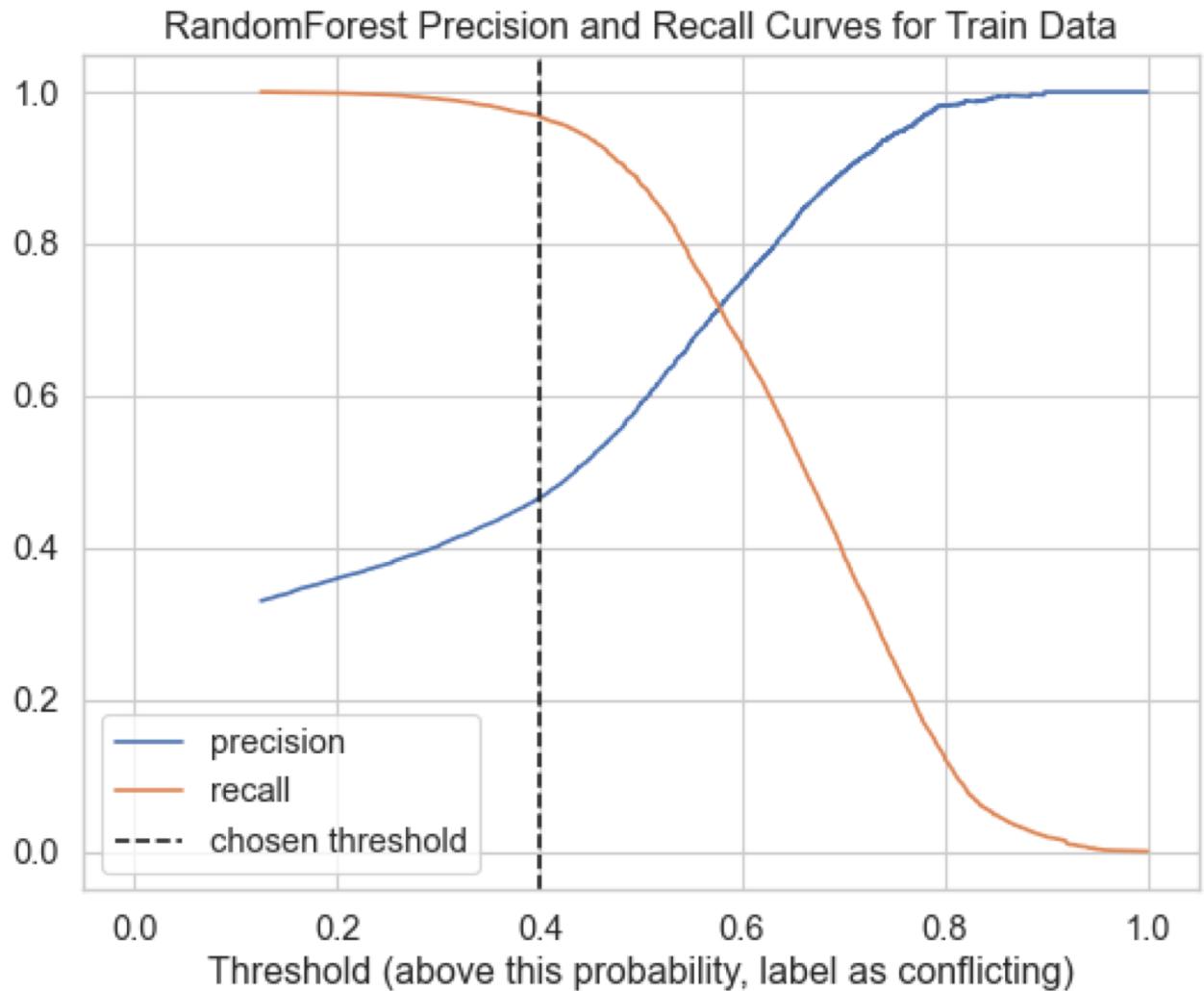




# Modeling

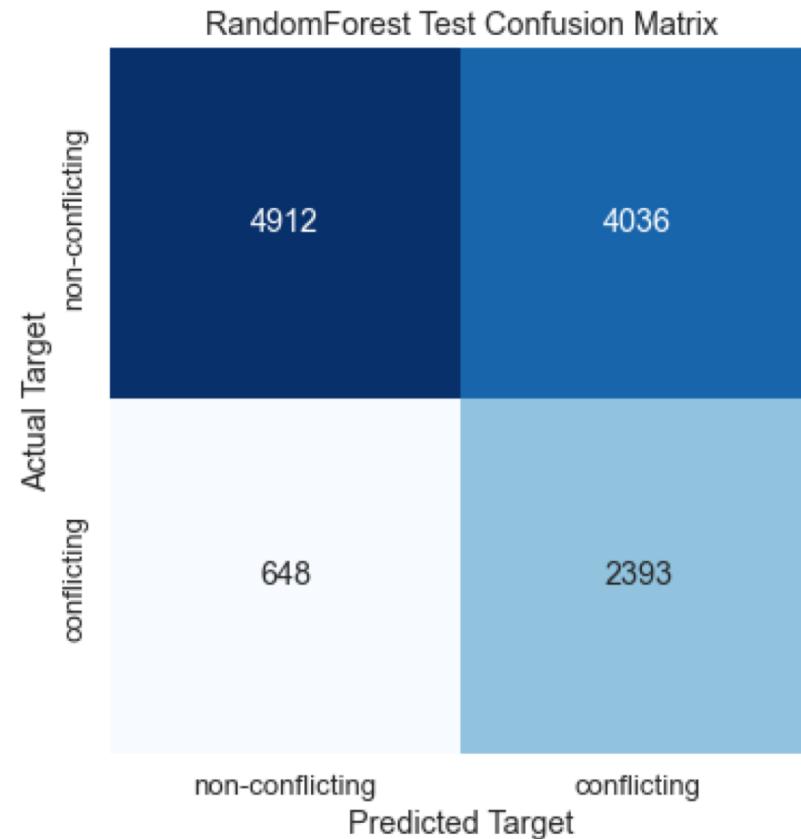


- Optimized on recall without entirely sacrificing precision
- More true positives at the cost of false positives which is acceptable
- Chose threshold of **0.4**

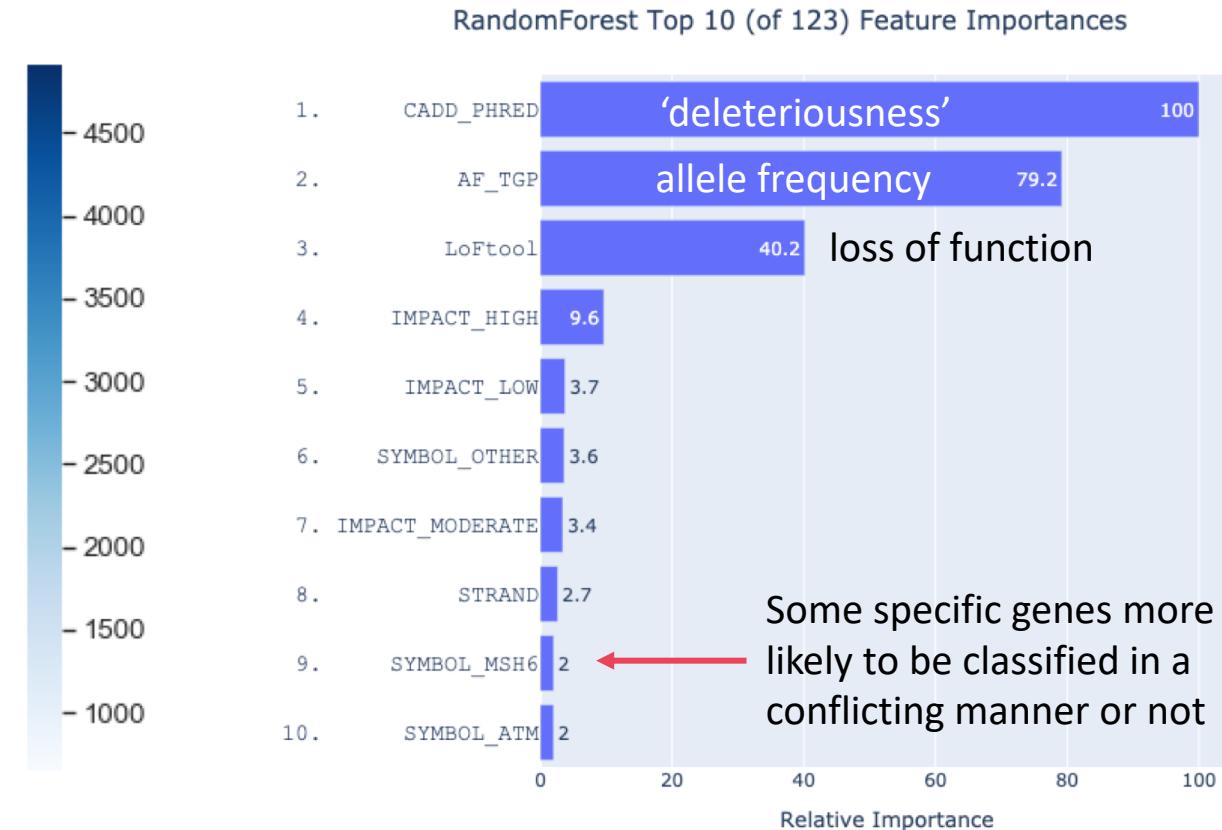




# Results



Recall: 0.78



Most important feature: CADD\_PHRED



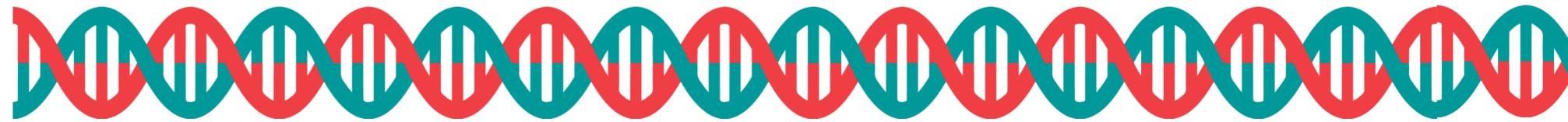
## Future Work



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



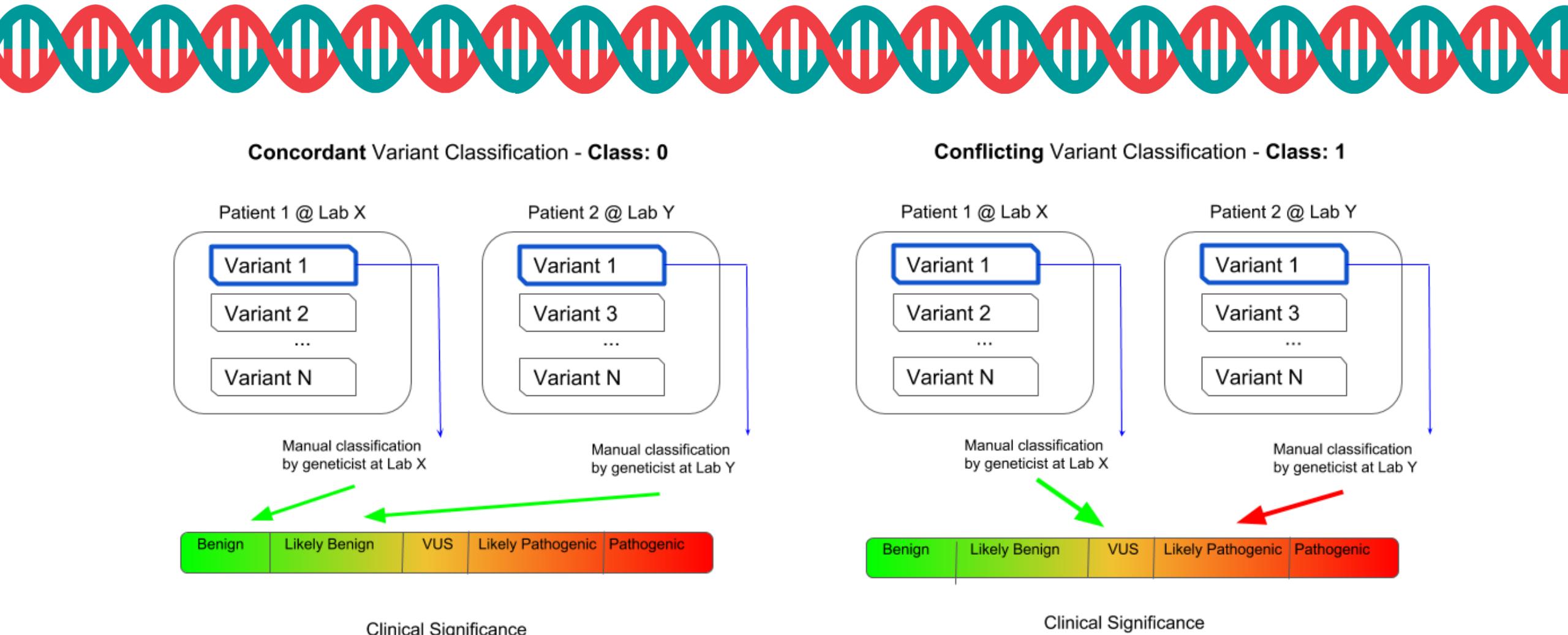
*“The human genome is only on the order of a gigabyte of data, which is a tiny little database. If you take the entire living biosphere...about one petabyte...that’s still very small compared with Google or the Wikipedia. And somehow mother nature manages to create...this incredibly rich environment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

*- Freeman Dyson*

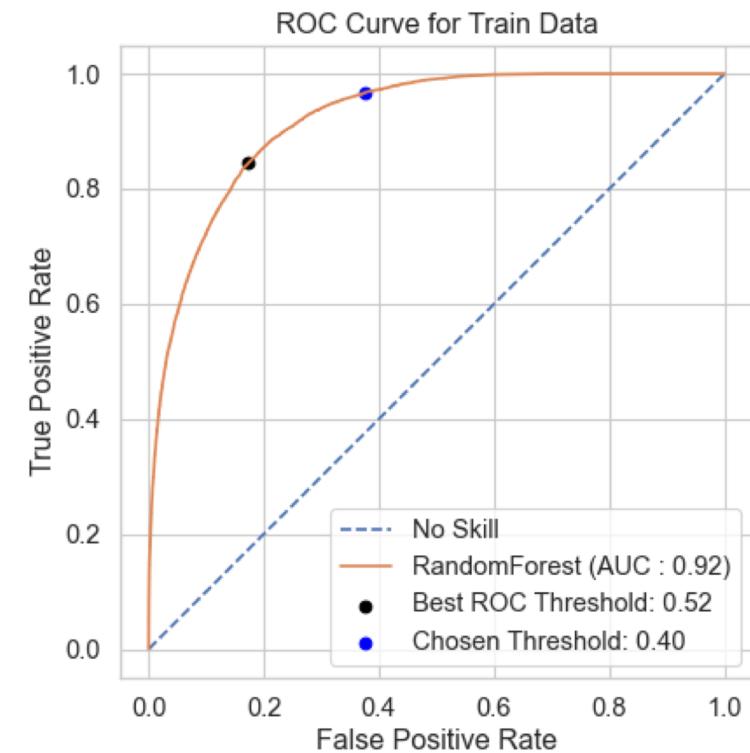
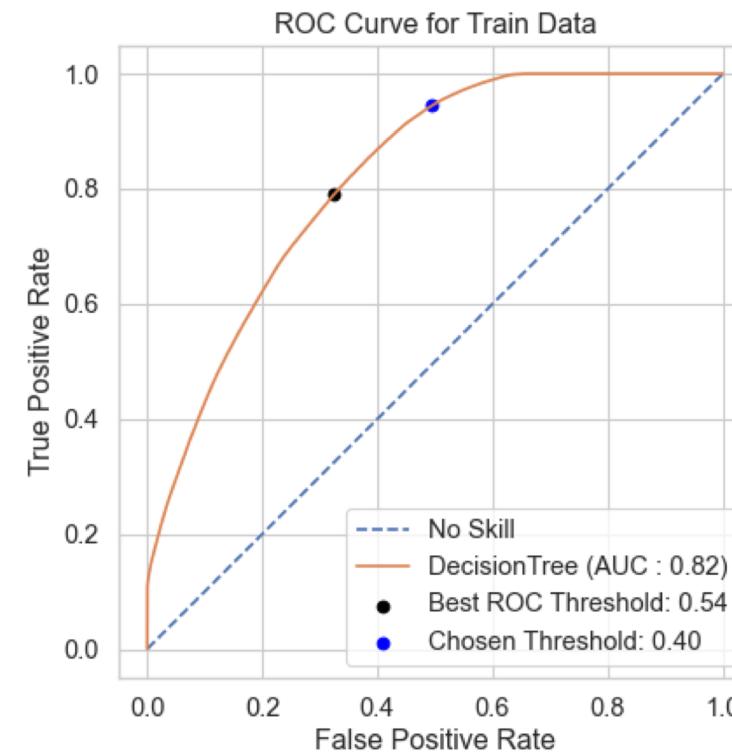
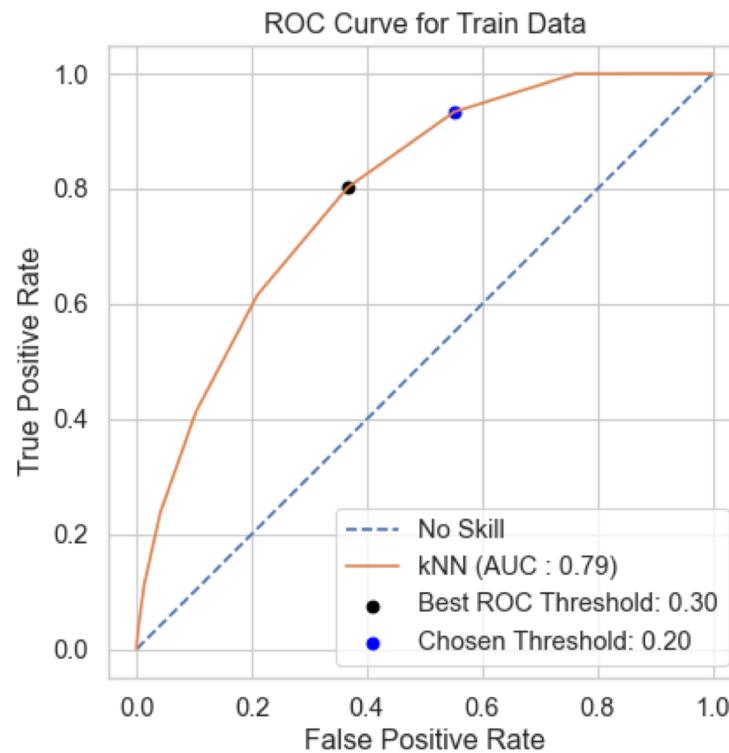


## Appendix



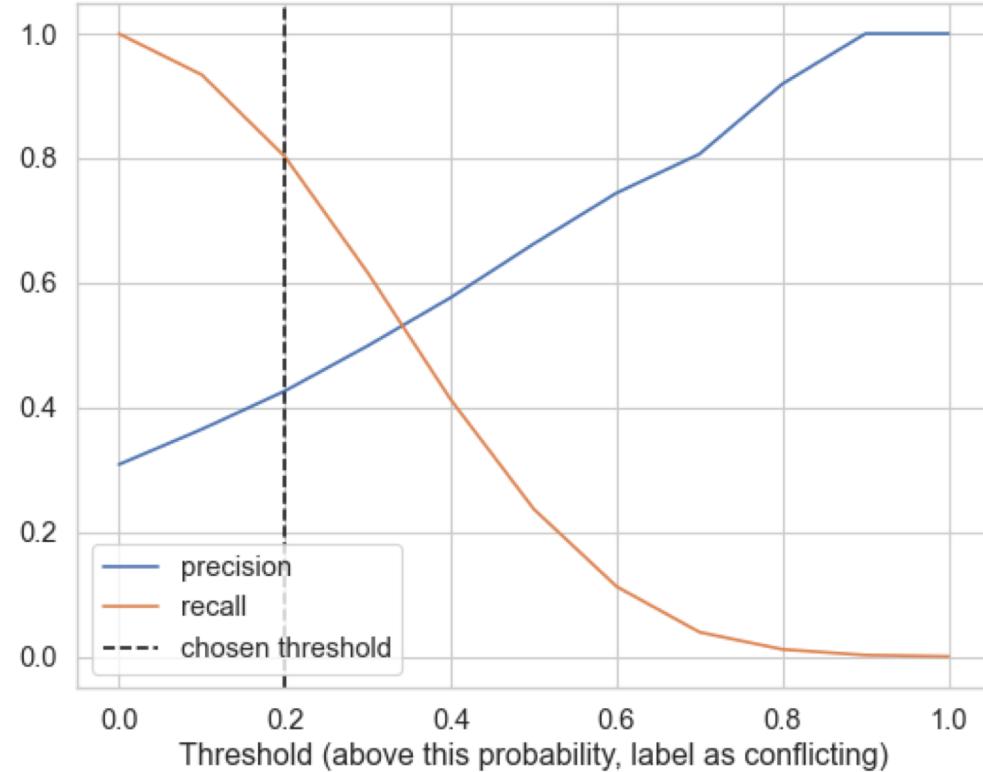


- 
- knn\_model = KNeighborsClassifier(n\_neighbors=10)
  - dt\_model = DecisionTreeClassifier(max\_depth=25, max\_features=20, random\_state=121, class\_weight={0:1,1:3})
  - rf\_model = RandomForestClassifier(n\_estimators=10, max\_depth=25, max\_features=20, random\_state=121, class\_weight={0:1,1:3})

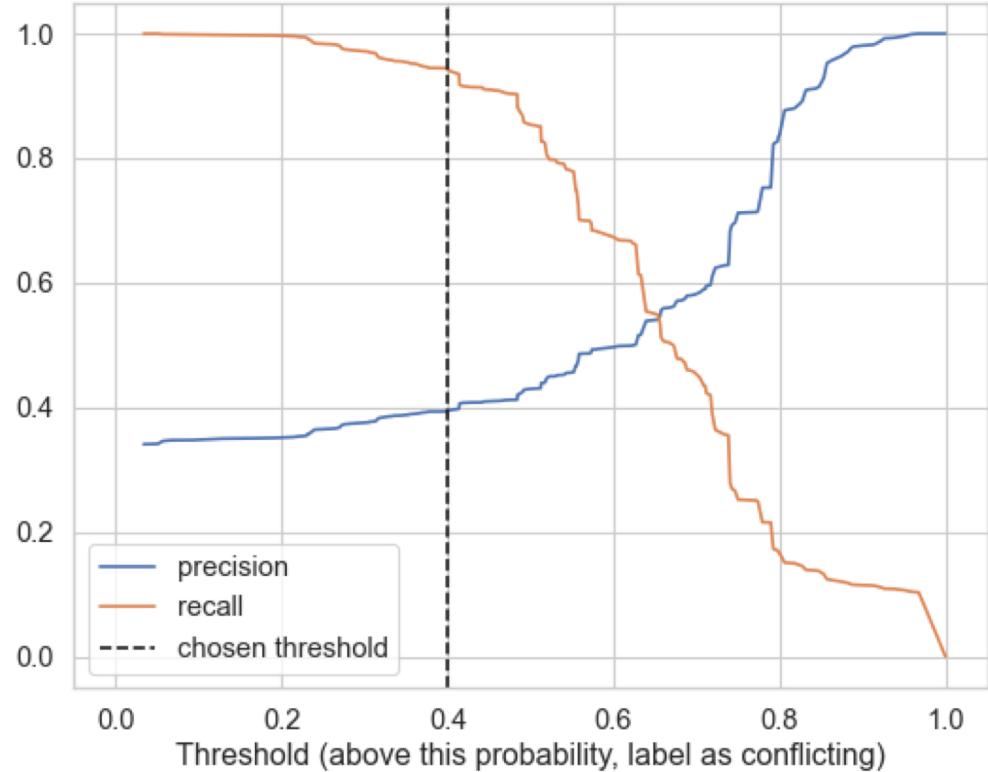




kNN Precision and Recall Curves for Train Data

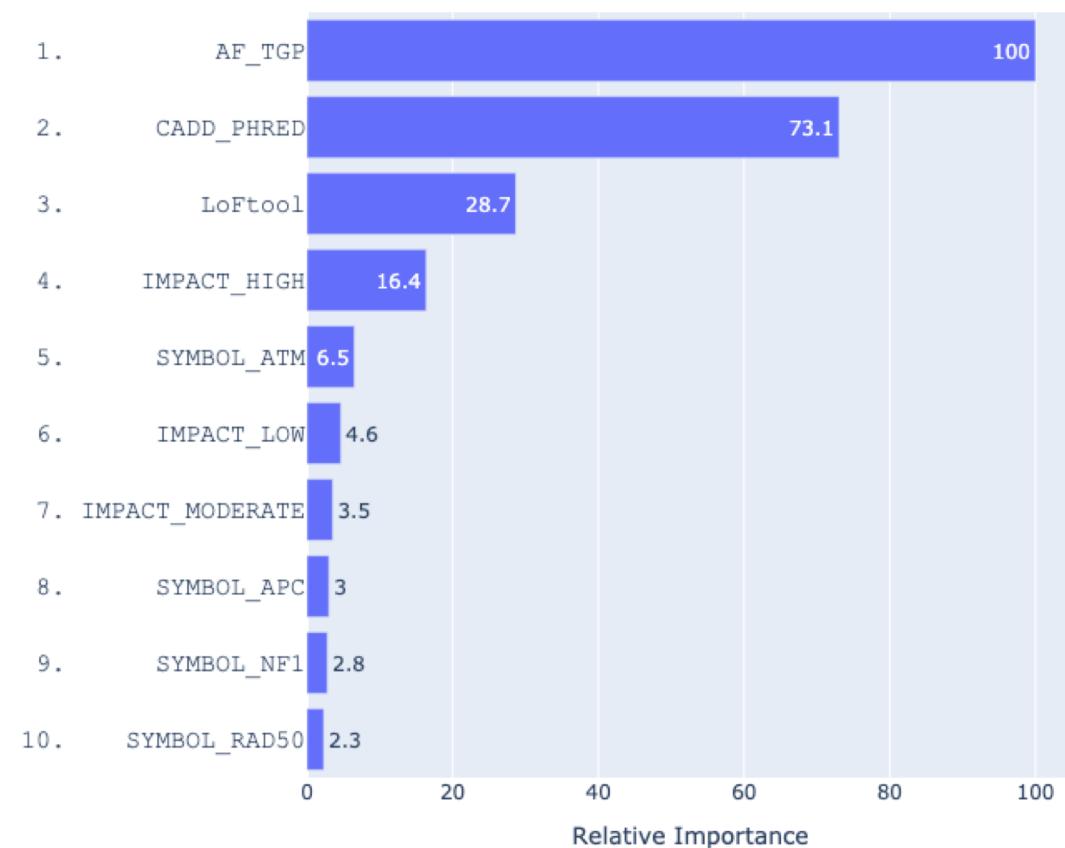


DecisionTree Precision and Recall Curves for Train Data





### DecisionTree Top 10 (of 123) Feature Importances

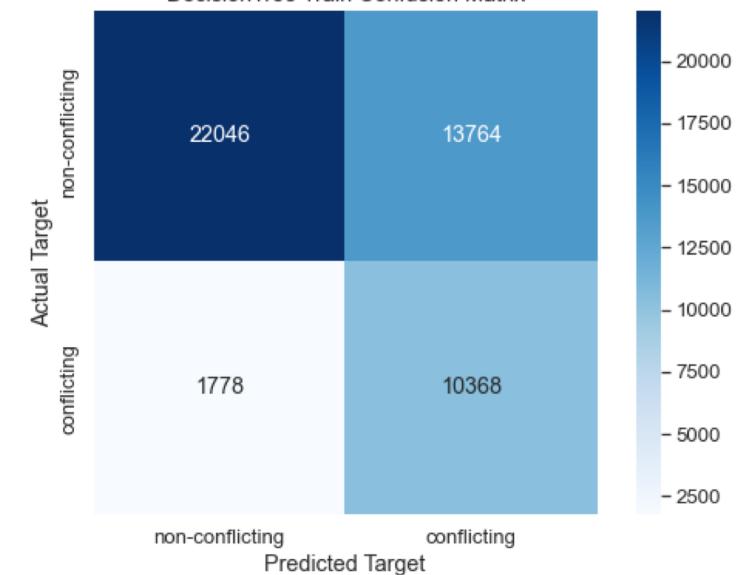




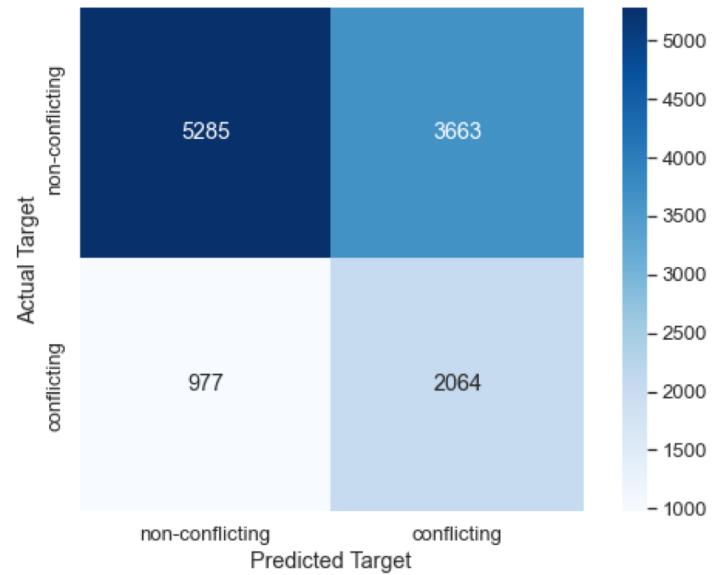
kNN Train Confusion Matrix



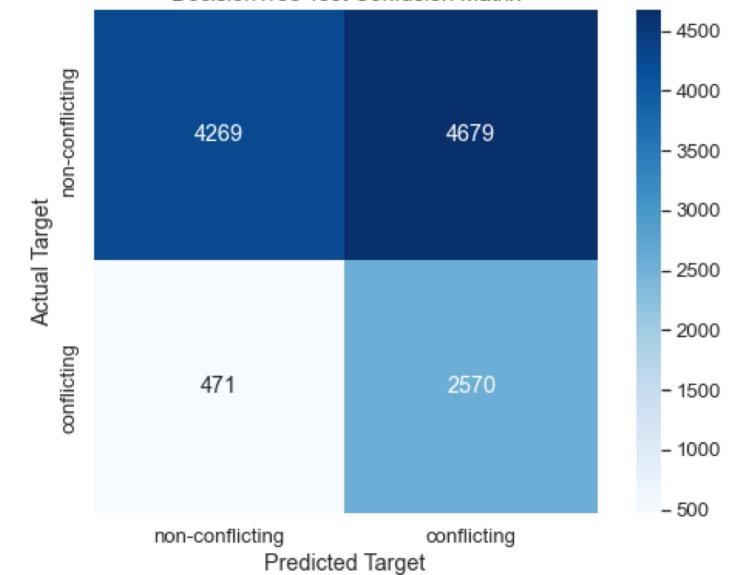
DecisionTree Train Confusion Matrix

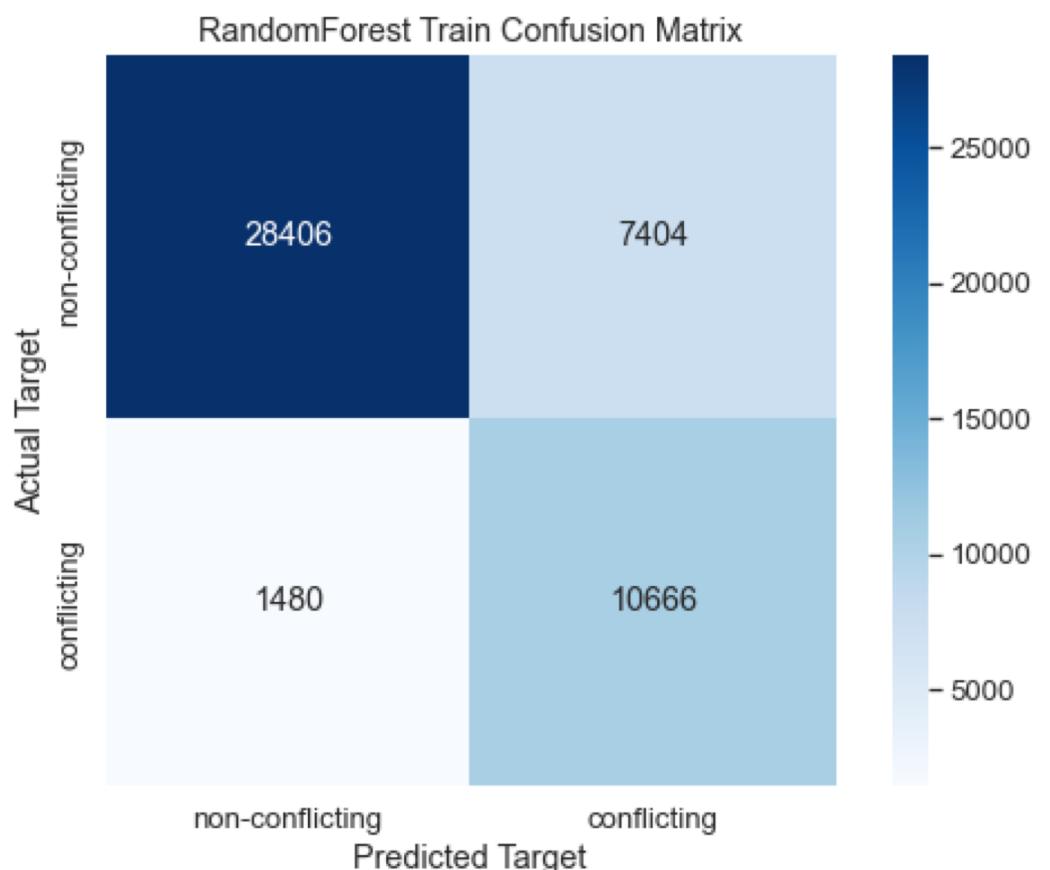


kNN Test Confusion Matrix



DecisionTree Test Confusion Matrix







- 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