Exploring the Oncogenic Impact of the TP53 Gene Variants and Classifying the Pathogenicity

By Alison Vaz

Quick Recap: What Is the TP53 Gene?

A: Part of the DNA responsible for regulating cell division

B: Contains nucleotides A, T, G, C

C: Present on the 17th chromosome

D: None of the above

E: All of the above except D

Quick Recap: What Is the TP53 Gene?

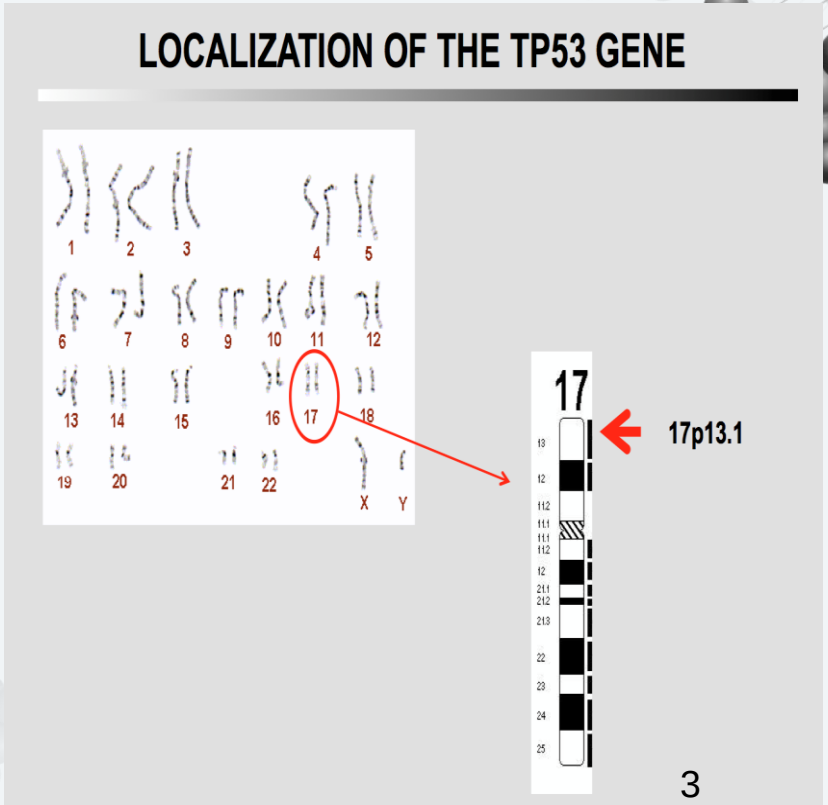
A: Part of the DNA

B: Contains nucleotides A, T, G, C

C: Present on the 17th chromosome

D: None of the above

E: All of the above except D



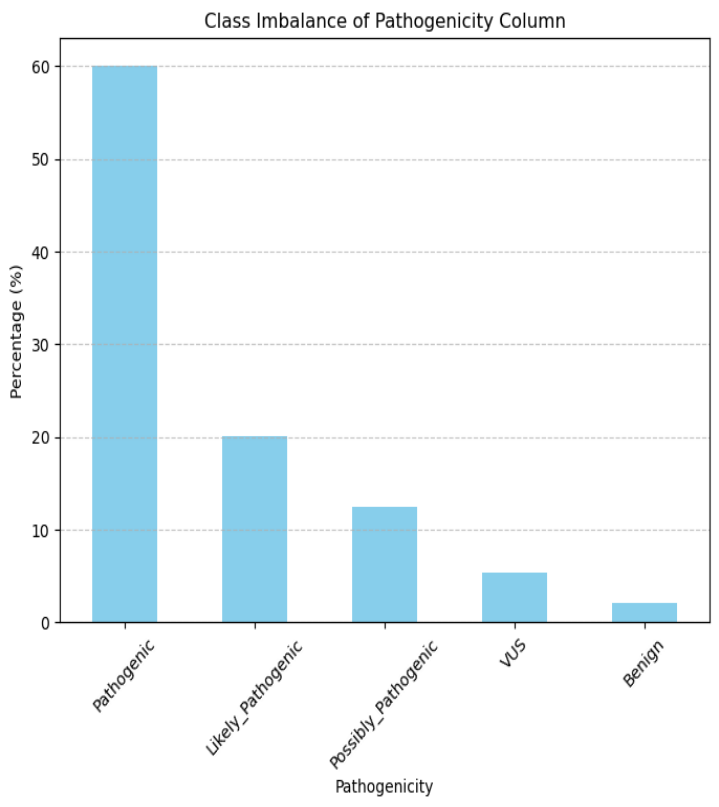
01 Data Dictionary

OG shape: 80406, 133

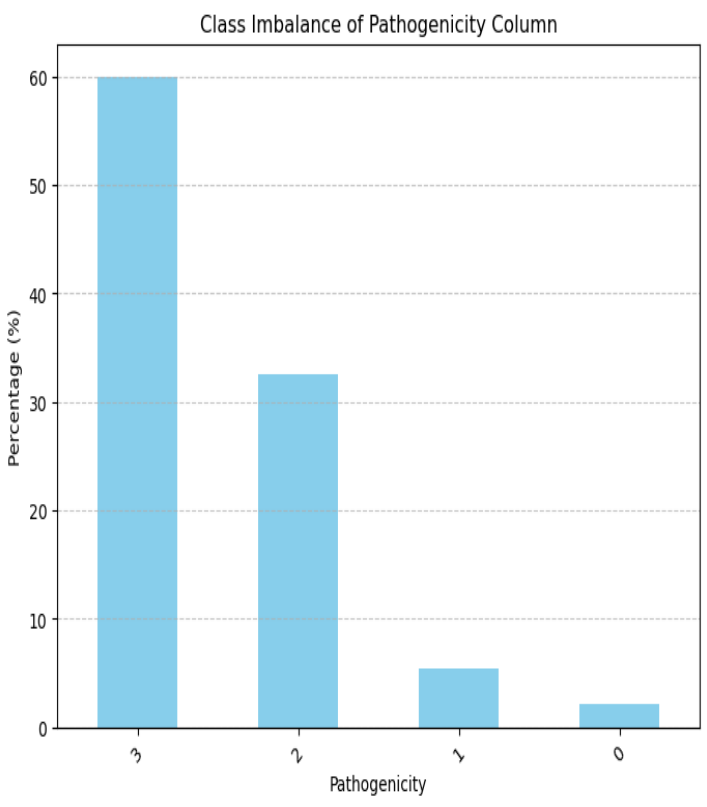
Features	Details
HG38_Start	Mutation start coordinates against the ref genome (int)
Mutant_Codon	**NNN:** Sequence of the mutated codon (69 types)
Disease	Name of the disease (11 including unknown/others)
Variant_Type	SNP, DNP, TNP, ONP, INS, DEL
Somatic_Stat	Frequency of the variant (cDNA_nomenclature) found as a somatic event
Target	**Details**
Pathogenicity	Benign, Unknown, Likely Pathogenic, Pathogenic

AfterEDA: 80337, 78

EDA

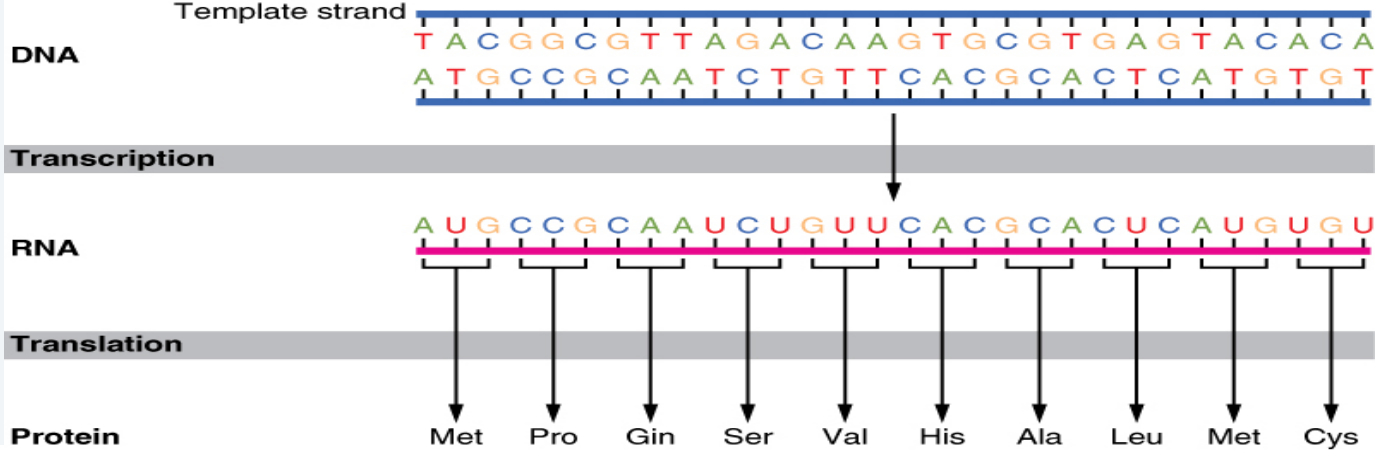


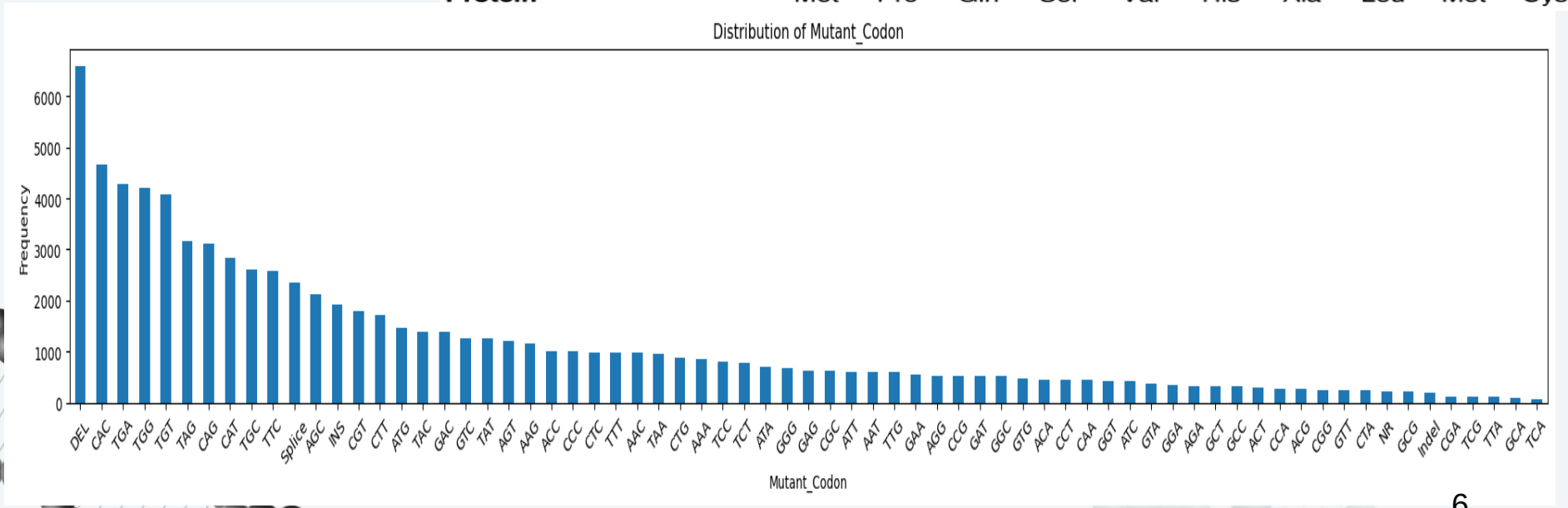
Previously



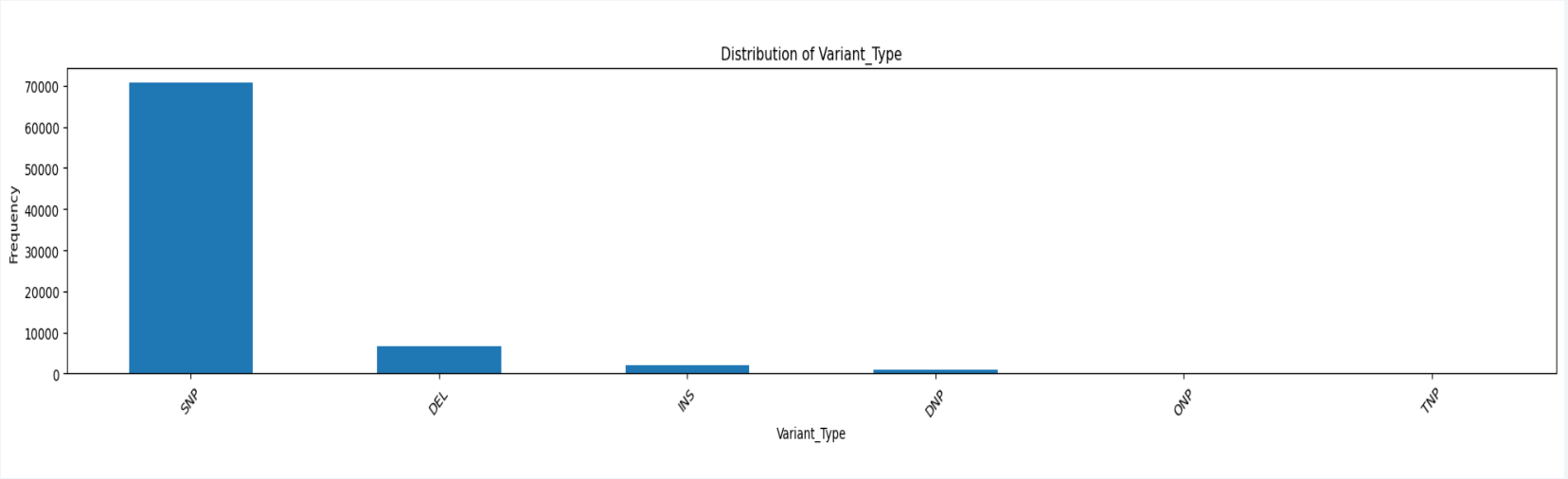
Now

EDA

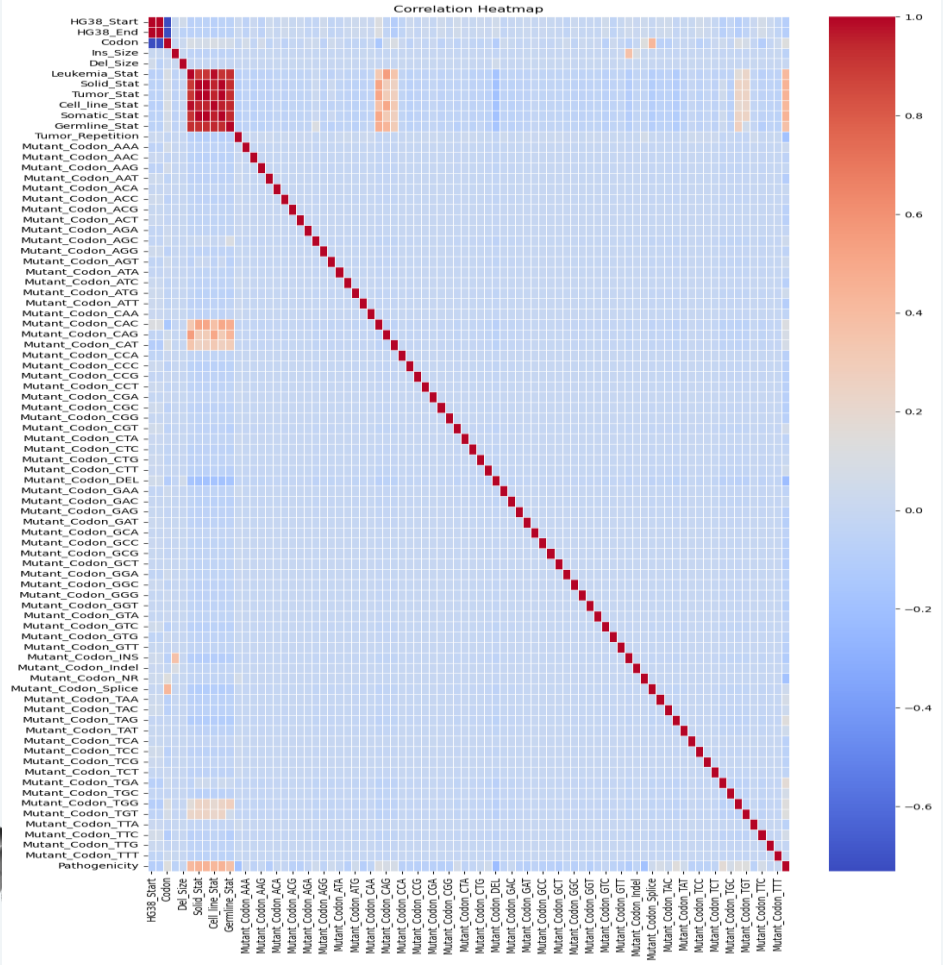

Template strand
DNA
T A C G G C G T T A G A C A A G T G C G T G A G T A C A C A
A T G C C G C A A T C T G T T C A C G C A C T C A T G T G T
Transcription
RNA
A U G C C G C A A U C U G U U C A C G C A C U C A U G U G U
Translation
Protein
Met
Pro
Gin
Ser
Val
His
Ala
Leu
Met
Cys

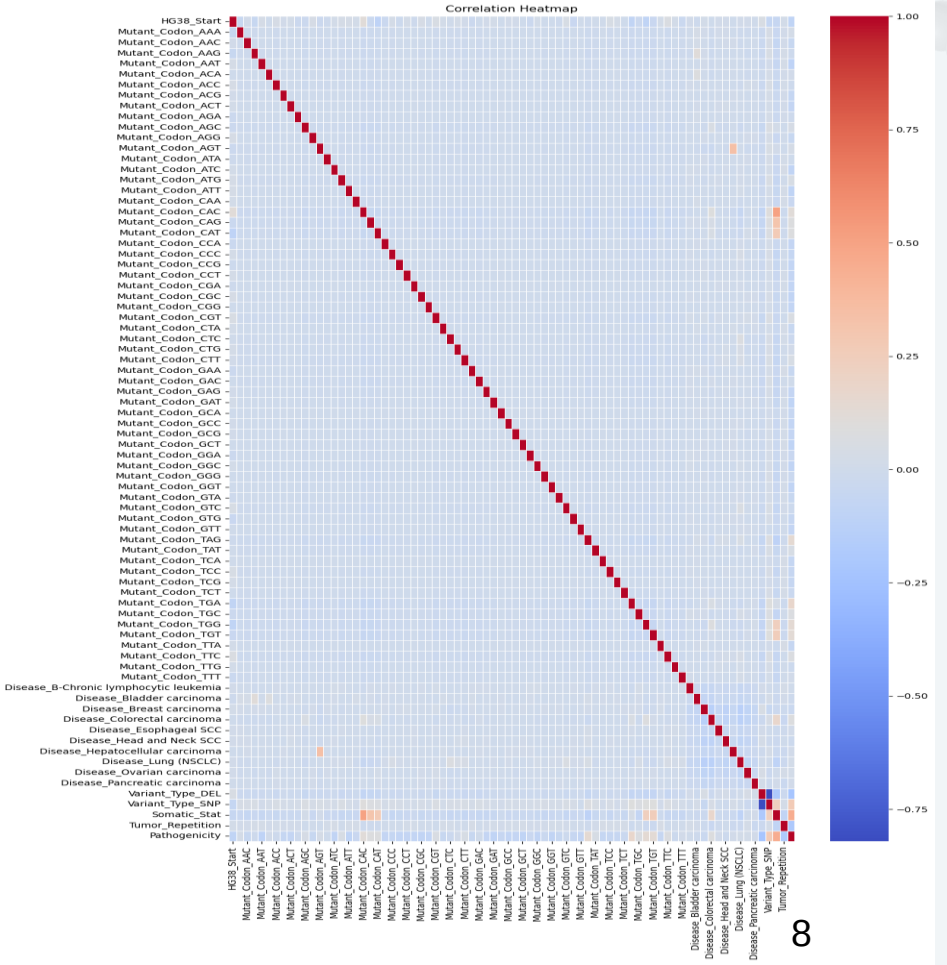

Distribution of Mutant_Codon
Frequency
0
1000
2000
3000
4000
5000
6000
DEL CAC TGA TGG TGT TAG CAG CAT TGC TTC Splice AGC INS CGT CTT ATG TAC GAC GTC TAT AGT AAG ACC CCC CTC TTT AAC TAA CTG AAA TCC TCT ATA GGG GAG CGC AIT AAT TTG GAA AGG CCG GAT GGC GTG ACA CCT CAA GGT ATC GTA GGA AGA GCT GCC ACT CCA ACG CGG GTT CTA NR GCG Indel CGA TCG TTA GCA TCA
Mutant_Codon

EDA

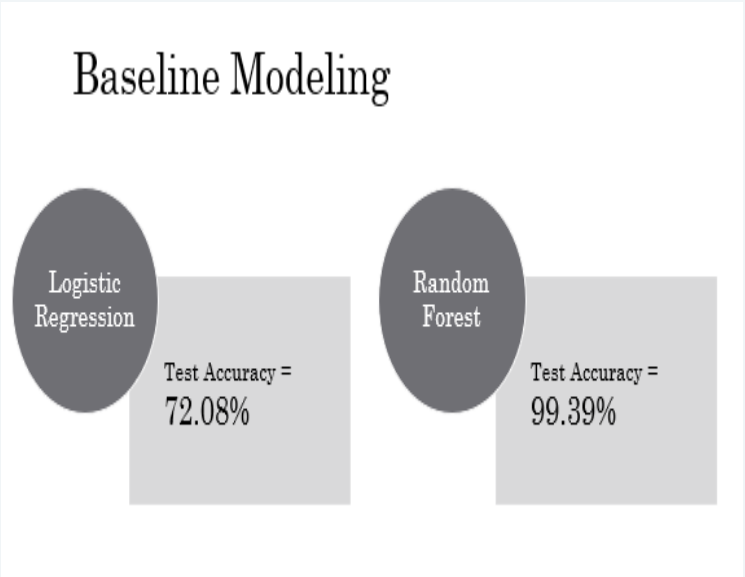

Distribution of Variant_Type
Frequency
70000
60000
50000
40000
30000
20000
10000
0
SNP
DEL
INS
DNP
ONP
TNP
Variant_Type

OHE


Correlation Heatmap
HG38_Start
HG38_End
Codon
Ins_Size
Del_Size
Leukemia_Stat
Solid_Stat
Tumor_Stat
Cell_line_Stat
Somatic_Stat
Germline_Stat
Tumor_Repetition
Mutant_Codon_AAA
Mutant_Codon_AAC
Mutant_Codon_AAG
Mutant_Codon_AAT
Mutant_Codon_ACA
Mutant_Codon_ACC
Mutant_Codon_ACG
Mutant_Codon_ACT
Mutant_Codon_AGA
Mutant_Codon_AGC
Mutant_Codon_AGG
Mutant_Codon_AGT
Mutant_Codon_ATA
Mutant_Codon_ATC
Mutant_Codon_ATG
Mutant_Codon_ATT
Mutant_Codon_CAA
Mutant_Codon_CAC
Mutant_Codon_CAG
Mutant_Codon_CAT
Mutant_Codon_CCA
Mutant_Codon_CCC
Mutant_Codon_CCG
Mutant_Codon_CCT
Mutant_Codon_CGA
Mutant_Codon_CGC
Mutant_Codon_CGG
Mutant_Codon_CGT
Mutant_Codon_CTA
Mutant_Codon_CTC
Mutant_Codon_CTG
Mutant_Codon_CTT
Mutant_Codon_DEL
Mutant_Codon_GAA
Mutant_Codon_GAC
Mutant_Codon_GAG
Mutant_Codon_GAT
Mutant_Codon_GCA
Mutant_Codon_GCC
Mutant_Codon_GCG
Mutant_Codon_GCT
Mutant_Codon_GGA
Mutant_Codon_GGC
Mutant_Codon_GGG
Mutant_Codon_GGT
Mutant_Codon_GTA
Mutant_Codon_GTC
Mutant_Codon_GTG
Mutant_Codon_GTT
Mutant_Codon_INS
Mutant_Codon_Indel
Mutant_Codon_NR
Mutant_Codon_Splice
Mutant_Codon_TAA
Mutant_Codon_TAC
Mutant_Codon_TAG
Mutant_Codon_TAT
Mutant_Codon_TCA
Mutant_Codon_TCC
Mutant_Codon_TCG
Mutant_Codon_TCT
Mutant_Codon_TGA
Mutant_Codon_TGC
Mutant_Codon_TGG
Mutant_Codon_TGT
Mutant_Codon_TTA
Mutant_Codon_TTC
Mutant_Codon_TTG
Mutant_Codon_TTT
Pathogenicity
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6


Correlation Heatmap
HG38_Start
Mutant_Codon_AAA
Mutant_Codon_AAC
Mutant_Codon_AAG
Mutant_Codon_AAT
Mutant_Codon_ACA
Mutant_Codon_ACC
Mutant_Codon_ACG
Mutant_Codon_ACT
Mutant_Codon_AGA
Mutant_Codon_AGC
Mutant_Codon_AGG
Mutant_Codon_AGT
Mutant_Codon_ATA
Mutant_Codon_ATC
Mutant_Codon_ATG
Mutant_Codon_ATT
Mutant_Codon_CAA
Mutant_Codon_CAC
Mutant_Codon_CAG
Mutant_Codon_CAT
Mutant_Codon_CCA
Mutant_Codon_CCC
Mutant_Codon_CCG
Mutant_Codon_CCT
Mutant_Codon_CGA
Mutant_Codon_CGC
Mutant_Codon_CGG
Mutant_Codon_CGT
Mutant_Codon_CTA
Mutant_Codon_CTC
Mutant_Codon_CTG
Mutant_Codon_CTT
Mutant_Codon_GAA
Mutant_Codon_GAC
Mutant_Codon_GAG
Mutant_Codon_GAT
Mutant_Codon_GCA
Mutant_Codon_GCC
Mutant_Codon_GCG
Mutant_Codon_GCT
Mutant_Codon_GGA
Mutant_Codon_GGC
Mutant_Codon_GGG
Mutant_Codon_GGT
Mutant_Codon_GTA
Mutant_Codon_GTC
Mutant_Codon_GTG
Mutant_Codon_GTT
Mutant_Codon_TAG
Mutant_Codon_TAT
Mutant_Codon_TCA
Mutant_Codon_TCC
Mutant_Codon_TCG
Mutant_Codon_TCT
Mutant_Codon_TGA
Mutant_Codon_TGC
Mutant_Codon_TGG
Mutant_Codon_TGT
Mutant_Codon_TTA
Mutant_Codon_TTC
Mutant_Codon_TTG
Mutant_Codon_TTT
Disease_B-Chronic lymphocytic leukemia
Disease_Bladder carcinoma
Disease_Breast carcinoma
Disease_Colorectal carcinoma
Disease_Esophageal SCC
Disease_Head and Neck SCC
Disease_Hepatocellular carcinoma
Disease_Lung (NSCLC)
Disease_Ovarian carcinoma
Disease_Pancreatic carcinoma
Variant_Type_DEL
Variant_Type_SNP
Somatic_Stat
Tumor_Repetition
Pathogenicity
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75

How is my modeling going?



Random Forest Test Accuracy: 0.88

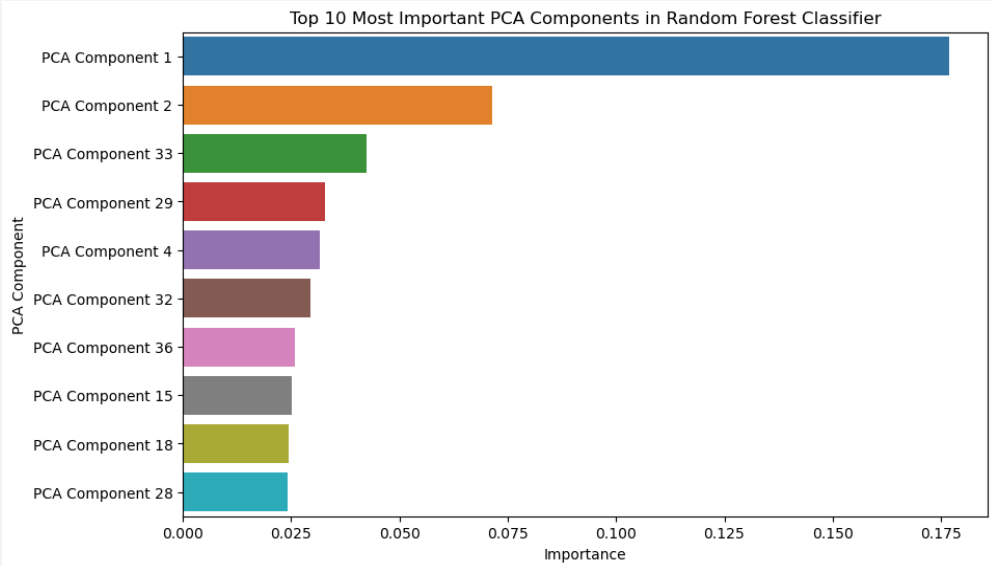
Classification Report for Test Data:

	precision	recall	f1-score	support
0	0.50	0.21	0.29	338
1	0.54	0.36	0.43	860
2	0.81	0.88	0.84	5221
3	0.94	0.94	0.94	9649
accuracy			0.88	16068
macro avg	0.70	0.60	0.63	16068
weighted avg	0.87	0.88	0.87	16068

Previously

Now

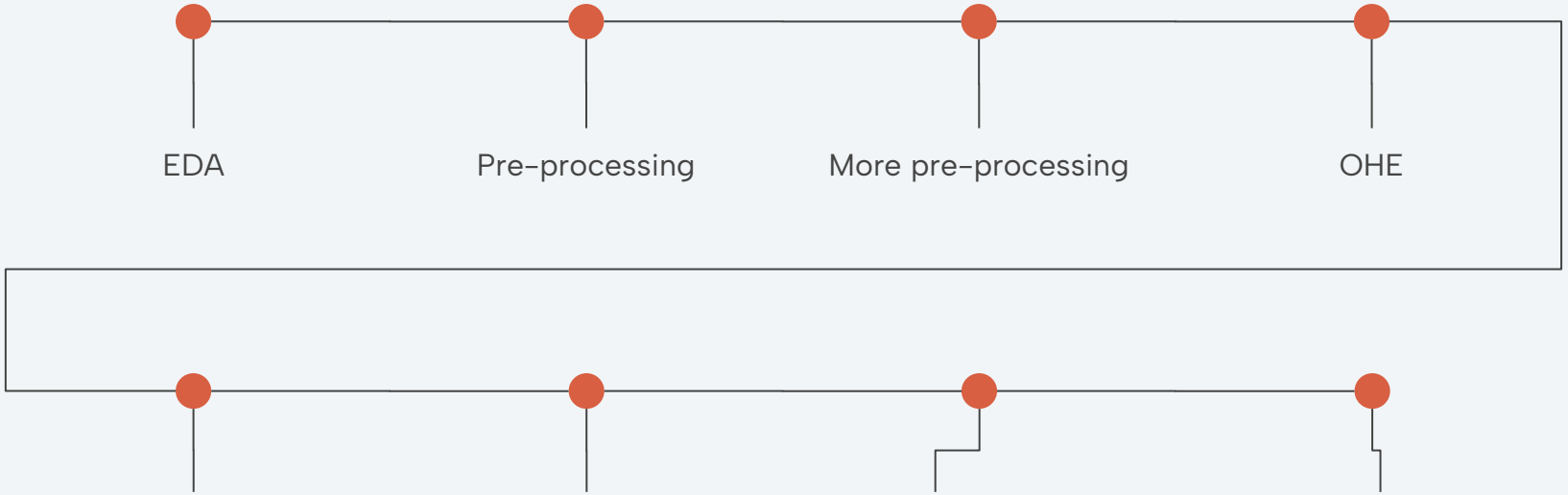
Feature contributions



Top contributing original features:

Variant_Type_DEL	0.052650
Tumor_Repetition	0.038162
Mutant_Codon_TCA	0.034861
Mutant_Codon_CAA	0.032330
Mutant_Codon_CGA	0.031971
HG38_Start	0.031281
Mutant_Codon_CCC	0.029289
Mutant_Codon_AAT	0.029276
Disease_Hepatocellular carcinoma	0.028071
Mutant_Codon_TTA	0.027566

Milestones reached



Thanks!

Do you have any questions?