# Human Genetic Variants

The dataset for this project was collected from kaggle and originates from ClinVar. ClinVar is a public resource containing annotations about human genetic variants. These variants are classified by clinical laboratories on a categorical spectrum ranging from benign, likely benign, uncertain significance, likely pathogenic, and pathogenic. Variants that have conflicting classifications (from laboratory to laboratory) can cause confusion when clinicians or researchers try to interpret whether the variant has an impact on the disease of a given patient.

The objective of the project is to predict whether a ClinVar variant will have conflicting classifications. This is presented here as a binary classification problem, where each record in the dataset is a genetic variant.

Conflicting classifications are when two of any of the following three categories are present for one variant, two submissions of one category are not considered conflicting.

- Likely benign or benign
- VUS
- Likely pathogenic or pathogenic

Conflicting classification has been assigned to the CLASS column. It is a binary representation of whether or not a variant has conflicting classifications, where 0 represents consistent classifications and 1 represents conflicting classifications.

IN the project. Four different classifier models will be deployed ot find the best candidate algorithm that accurately predicts whether a ClinVar variant will have conflicting classifications.

**Data summary:**

1. Uploaded variation - as chromosome_start_alleles
2. Location - in standard coordinate format (chr:start or chr:start-end)
3. Allele - the variant allele used to calculate the consequence
4. Gene - Ensembl stable ID of affected gene
5. Feature - Ensembl stable ID of feature
6. Feature type - type of feature. Currently one of Transcript, RegulatoryFeature, MotifFeature.
7. Consequence - consequence type of this variant
8. Position in cDNA - relative position of base pair in cDNA sequence
9. Position in CDS - relative position of base pair in coding sequence
10. Position in protein - relative position of amino acid in protein
11. Amino acid change - only given if the variant affects the protein-coding sequence
12. Codon change - the alternative codons with the variant base in upper case
13. Co-located variation - known identifier of existing variant
14. Extra - this column contains extra information as key=value pairs separated by ";", see below.

Other output fields:

- REF_ALLELE - the reference allele
- IMPACT - the impact modifier for the consequence type
- VARIANT_CLASS - Sequence Ontology variant class
- SYMBOL - the gene symbol
- SYMBOL_SOURCE - the source of the gene symbol
- STRAND - the DNA strand (1 or -1) on which the transcript/feature lies
- ENSP - the Ensembl protein identifier of the affected transcript
- FLAGS - transcript quality flags:
    - *cds_start_NF:* CDS 5' incomplete
    - *cds_end_NF:* CDS 3' incomplete
- SWISSPROT - Best match UniProtKB/Swiss-Prot accession of protein product
- TREMBL - Best match UniProtKB/TrEMBL accession of protein product
- UNIPARC - Best match UniParc accession of protein product
- HGVSc - the HGVS coding sequence name
- HGVSp - the HGVS protein sequence name
- HGVSg - the HGVS genomic sequence name
- HGVS_OFFSET - Indicates by how many bases the HGVS notations for this variant have been shifted
- NEAREST - Identifier(s) of nearest transcription start site
- SIFT - the SIFT prediction and/or score, with both given as prediction(score)
- PolyPhen - the PolyPhen prediction and/or score
- MOTIF_NAME - the source and identifier of a transcription factor binding profile aligned at this position
- MOTIF_POS - The relative position of the variation in the aligned TFBP
- HIGH_INF_POS - a flag indicating if the variant falls in a high information position of a transcription factor binding profile (TFBP)
- MOTIF_SCORE_CHANGE - The difference in motif score of the reference and variant sequences for the TFBP
- CELL_TYPE - List of cell types and classifications for regulatory feature
- CANONICAL - a flag indicating if the transcript is denoted as the canonical transcript for this gene
- CCDS - the CCDS identifer for this transcript, where applicable
- INTRON - the intron number (out of total number)
- EXON - the exon number (out of total number)
- DOMAINS - the source and identifer of any overlapping protein domains
- DISTANCE - Shortest distance from variant to transcript
- IND - individual name
- ZYG - zygosity of individual genotype at this locus
- SV - IDs of overlapping structural variants
- FREQS - Frequencies of overlapping variants used in filtering
- AF - Frequency of existing variant in 1000 Genomes
- AFR_AF - Frequency of existing variant in 1000 Genomes combined African population
- AMR_AF - Frequency of existing variant in 1000 Genomes combined American population

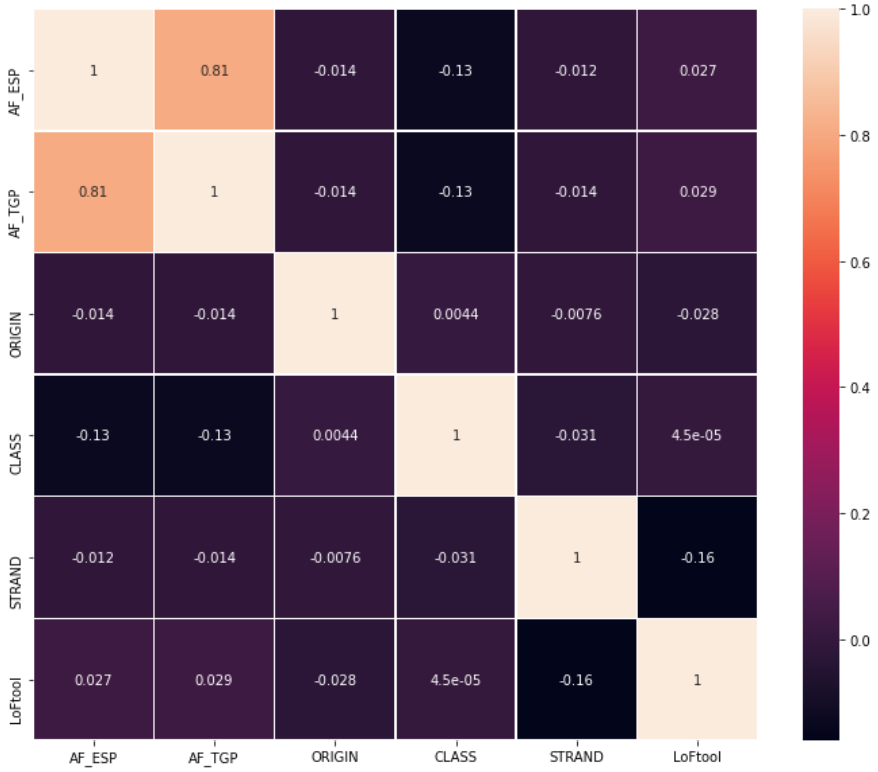
- ASN_AF - Frequency of existing variant in 1000 Genomes combined Asian population
- EUR_AF - Frequency of existing variant in 1000 Genomes combined European population
- EAS_AF - Frequency of existing variant in 1000 Genomes combined East Asian population
- SAS_AF - Frequency of existing variant in 1000 Genomes combined South Asian population
- AA_AF - Frequency of existing variant in NHLBI-ESP African American population
- EA_AF - Frequency of existing variant in NHLBI-ESP European American population
- gnomAD_AF - Frequency of existing variant in gnomAD exomes combined population
- gnomAD_AFR_AF - Frequency of existing variant in gnomAD exomes African/American population
- gnomAD_AMR_AF - Frequency of existing variant in gnomAD exomes American population
- gnomAD_ASJ_AF - Frequency of existing variant in gnomAD exomes Ashkenazi Jewish population
- gnomAD_EAS_AF - Frequency of existing variant in gnomAD exomes East Asian population
- gnomAD_FIN_AF - Frequency of existing variant in gnomAD exomes Finnish population
- gnomAD_NFE_AF - Frequency of existing variant in gnomAD exomes Non-Finnish European population
- gnomAD_OTH_AF - Frequency of existing variant in gnomAD exomes combined other combined populations
- gnomAD_SAS_AF - Frequency of existing variant in gnomAD exomes South Asian population
- MAX_AF - Maximum observed allele frequency in 1000 Genomes, ESP and gnomAD
- MAX_AF_POPS - Populations in which maximum allele frequency was observed
- CLIN_SIG - ClinVar clinical significance of the dbSNP variant
- BIOTYPE - Biotype of transcript or regulatory feature
- APPRIS - Annotates alternatively spliced transcripts as primary or alternate based on a range of computational methods. NB: not available for GRCh37
- TSL - Transcript support level. NB: not available for GRCh37
- PUBMED - Pubmed ID(s) of publications that cite existing variant
- SOMATIC - Somatic status of existing variant(s); multiple values correspond to multiple values in the Existing_variation field
- PHENO - Indicates if existing variant is associated with a phenotype, disease or trait; multiple values correspond to multiple values in the Existing_variation field
- GENE_PHENO - Indicates if overlapped gene is associated with a phenotype, disease or trait
- ALLELE_NUM - Allele number from input; 0 is reference, 1 is first alternate etc
- MINIMISED - Alleles in this variant have been converted to minimal representation before consequence calculation
- PICK - indicates if this block of consequence data was picked by --flag_pick or --flag_pick_allele
- BAM_EDIT - Indicates success or failure of edit using BAM file

- GIVEN_REF - Reference allele from input
- USED_REF - Reference allele as used to get consequences
- REFSEQ_MATCH - the RefSeq transcript match status; contains a number of flags indicating whether this RefSeq transcript matches the underlying reference sequence and/or an Ensembl transcript (more information).
    - *rseq_3p_mismatch:* signifies a mismatch between the RefSeq transcript and the underlying primary genome assembly sequence. Specifically, there is a mismatch in the 3' UTR of the RefSeq model with respect to the primary genome assembly (e.g. GRCh37/GRCh38).
    - *rseq_5p_mismatch:* signifies a mismatch between the RefSeq transcript and the underlying primary genome assembly sequence. Specifically, there is a mismatch in the 5' UTR of the RefSeq model with respect to the primary genome assembly.
    - *rseq_cds_mismatch:* signifies a mismatch between the RefSeq transcript and the underlying primary genome assembly sequence. Specifically, there is a mismatch in the CDS of the RefSeq model with respect to the primary genome assembly.
    - *rseq_ens_match_cds:* signifies that for the RefSeq transcript there is an overlapping Ensembl model that is identical across the CDS region only. A CDS match is defined as follows: the CDS and peptide sequences are identical and the genomic coordinates of every translatable exon match. Useful related attributes are: rseq_ens_match_wt and rseq_ens_no_match.
    - *rseq_ens_match_wt:* signifies that for the RefSeq transcript there is an overlapping Ensembl model that is identical across the whole transcript. A whole transcript match is defined as follows: 1) In the case that both models are coding, the transcript, CDS and peptide sequences are all identical and the genomic coordinates of every exon match. 2) In the case that both transcripts are non-coding the transcript sequences and the genomic coordinates of every exon are identical. No comparison is made between a coding and a non-coding transcript. Useful related attributes are: rseq_ens_match_cds and rseq_ens_no_match.
    - *rseq_ens_no_match:* signifies that for the RefSeq transcript there is no overlapping Ensembl model that is identical across either the whole transcript or the CDS. This is caused by differences between the transcript, CDS or peptide sequences or between the exon genomic coordinates. Useful related attributes are: rseq_ens_match_wt and rseq_ens_match_cds.
    - *rseq_mrna_match:* signifies an exact match between the RefSeq transcript and the underlying primary genome assembly sequence (based on a match between the transcript stable id and an accession in the RefSeq mRNA file). An exact match occurs when the underlying genomic sequence of the model can be perfectly aligned to the mRNA sequence post polyA clipping.
    - *rseq_mrna_nonmatch:* signifies a non-match between the RefSeq transcript and the underlying primary genome assembly sequence. A non-match is deemed to have occurred if the underlying genomic sequence does not have a perfect alignment to the mRNA sequence post polyA clipping. It can also signify that no comparison was possible as the model stable id may not have had a corresponding entry in the RefSeq mRNA file (sometimes happens when accessions are retired or changed). When a non-match occurs one or

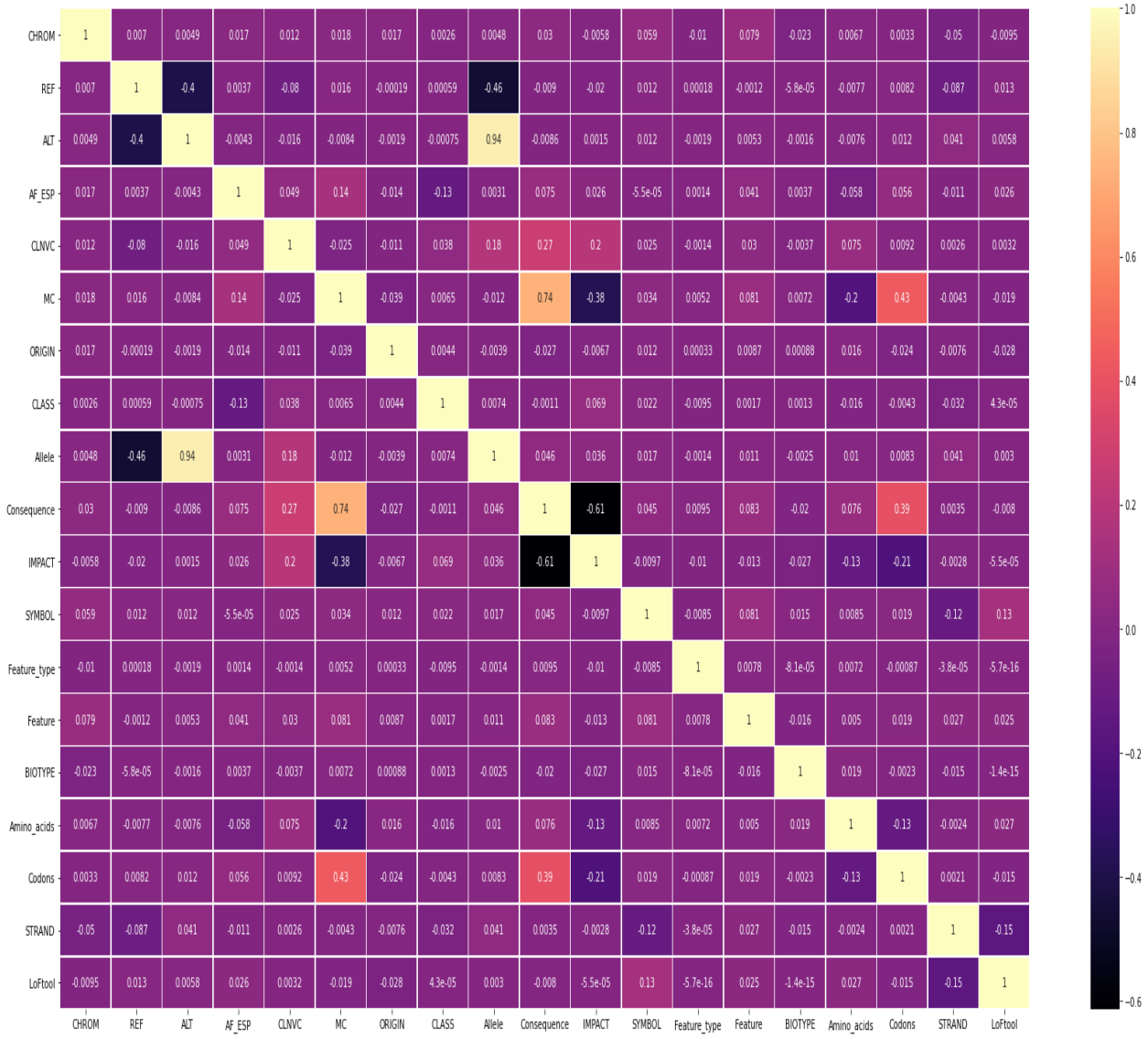
several of the following transcript attributes will also be present to provide more detail on the nature of the non-match: rseq_5p_mismatch, rseq_cds_mismatch, rseq_3p_mismatch, rseq_nctran_mismatch, rseq_no_comparison
    - *rseq_nctran_mismatch:* signifies a mismatch between the RefSeq transcript and the underlying primary genome assembly sequence. This is a comparison between the entire underlying genomic sequence of the RefSeq model to the mRNA in the case of RefSeq models that are non-coding.
    - *rseq_no_comparison:* signifies that no alignment was carried out between the underlying primary genome assembly sequence and a corresponding RefSeq mRNA. The reason for this is generally that no corresponding, unversioned accession was found in the RefSeq mRNA file for the transcript stable id. This sometimes happens when accessions are retired or replaced. A second possibility is that the sequences were too long and problematic to align (though this is rare).
- OverlapBP - Number of base pairs overlapping with the corresponding structural variation feature
- OverlapPC - Percentage of corresponding structural variation feature overlapped by the given input
- CHECK_REF - Reports variants where the input reference does not match the expected reference
- AMBIGUITY - IUPAC allele ambiguity code

**Exploratory Data Analysis:**

- THe data set has 65k+ rows and 46 features.
- The Target feature is a lot more consistent than conflicting classifications.



- Dropped columns that has too many unique values (>3000).
- Dropped columns having missing values more than 20%.
- From the correlation map, AF_ESP with AF_TGP has a relation more than 0.8, hence dropping it.
- Replacing nan in MC, SYMBOL, Feature_type, Feature, BIOTYPE, Amino_acids, Codons, Strand with the most frequent value, also replacing nan in LoFtool with the mean.
- Created the list of ordinal, binary and categorical columns and encoded them.
- Correlation of ALT with Allele and MC with Consequence are both above 0.8, hence dropped.

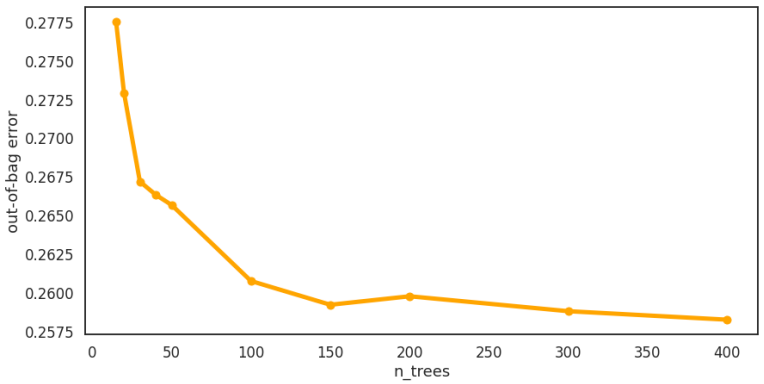



**Feature scaling:**

- Categorical and numerical variables using MinMaxScaler
- Saved the processed file.

**Splitting & Training models & Results:**

- Split using stratified sampling to keep the data set balanced.
- Training models:
    - Logistic regression
    - K-nearest neighbors
    - Decision Tree
    - RandomForest
- Metrixs
    - Precision
    - Recall
    - Accuracy
    - F1_score
    - Roc_auc

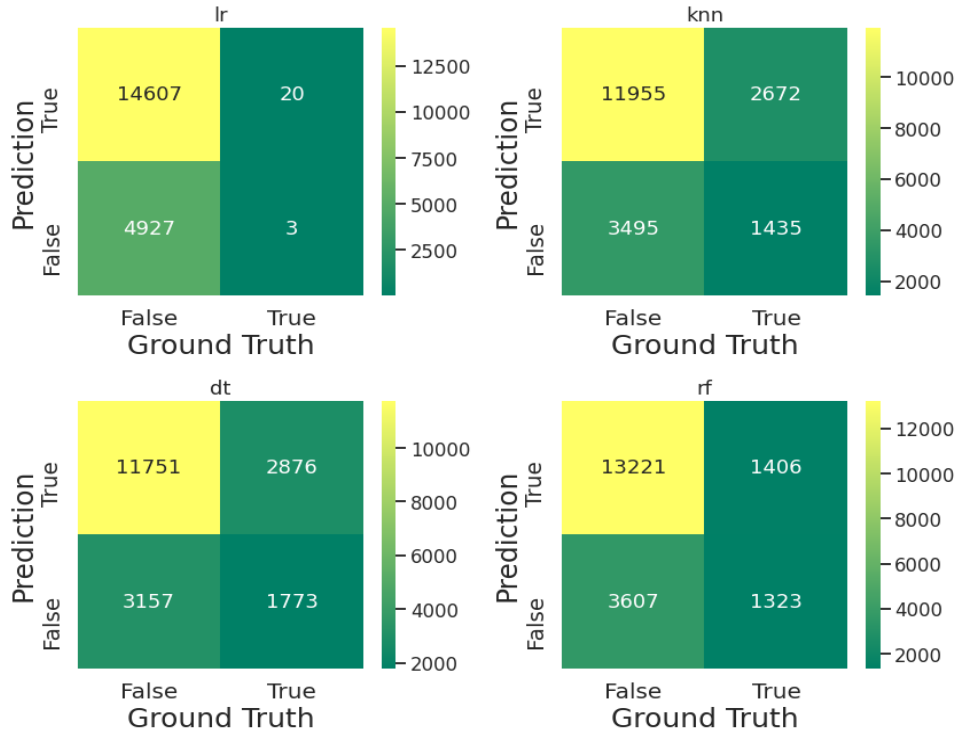



- In RandomForest, it seems like error has plateaued around 150 trees.

**Results:**

| | precision | recall | f1score |
|---|---|---|---|
| Logistic Regression | 0.75 | 0.99 | 0.85 |
| KNN | 0.77 | 0.81 | 0.79 |
| Decision Tree | 0.79 | 0.80 | 0.80 |
| Random Forest | 0.79 | 0.90 | 0.84 |

The classification report of each classifier shows that I am able to predict consistent classification with an F1_score 0f  0.85 in Logistic regression, but it seems to overfit the model, Similarly, RandomForest has the f1_score around 84% which seems better than it is less prone to overfitting than logistic regression.




**Go Forth:**

Further pruning the trees and optimizing with hyperparameters, using GridSearchCV or boosting algorithm can significantly improve the scores.