

Project A4: KAGGLE-CLINVAR

Team members: Ilze Polekauskaite, Roza Rostam Nejad, Liene Gutmane

Link to the repository: <https://github.com/RozaRostamnejad/DS-2024-Project-A4-Group-1>

Kaggle project: <https://www.kaggle.com/datasets/kevinarvai/clinvar-conflicting>

## Business understanding

- Identifying your business goals

- Background

ClinVar is a database for clinical laboratories worldwide to upload their findings on DNA or protein sequencing variants. They can share structural information on the gene products and describe the patient's clinical history. According to the American Clinical Genetics Society (ACGS) guidelines and based on their observations, the user decides on a classification for the variant. These are: likely pathogenic, pathogenic, uncertain significance, likely benign, benign. A user viewing the collective decisions of labs for a single variant can understand how confident the classification is. It is common for clinical practices to utilise this database in making their own clinical decisions for patients: once they have acquired information on phenotype, family history, sequenced the DNA and obtained variants to investigate, the information on ClinVar works as an additional resource, facilitating diagnosis. Researchers working on a disease also benefit from the database, as they can identify variants to study for describing the underlying mechanisms or drug discovery.

It is problematic when variants have conflicting classifications, as this may hinder treatment, when the specific genetic cause, therefore, a target, cannot be established. Therefore, we aim to predict whether a given variant will have conflicting classifications, which could aid in identifying the overarching reasons for this outcome. This could help future improvements on how to study variants and solve issues arising from more complicated cases.

- Business goals

A long-term goal for the project could be to utilise the information gathered on what makes a variant have conflicting classifications, and derive fine-tuning measures to apply to such cases, improving the ACGS guidelines. The updates would mean that the information gathered is not just a collection, but a practical way for achieving better clinical outcomes.

- Business success criteria

1. Quantitative:
  - a. Prediction accuracy of at least 85%
  - b. Chromosome grouping accuracy of 90%
2. Qualitative:

- a. Insight generation: Identify patterns and factors associated with conflicting classifications of genetic variants
- b. Usability of visualizations: Create clear and interpretable visualizations that help understand patterns in conflicting classifications.
- c. Usability of data: Ensure robustness and relevance of findings related to conflicting classifications.
- d. Population Analysis: Analyze and visualize the frequency of variants across different populations.

- **Assessing your situation**

- Inventory of resources

Data: ClinVar dataset

Hardware: Computing resources for data processing and model training

Software:

1. Data analysis and visualization tools (e.g Python, Tableau)
2. Machine learning libraries
3. Data cleaning and preprocessing tools

- Requirements, assumptions, and constraints

Requirements:

1. Access to the ClinVar dataset
2. Clear documentation of all processes and findings

Assumptions:

1. The ClinVar dataset is representative and comprehensive enough for the analysis
2. Necessary computational resources will be available

Constraints:

1. Limited time frame for project completion (the project deadline)
2. Potential data quality issues in the ClinVar dataset
3. No access to domain experts for validation

- Risks and contingencies

1. Data quality issues: Incomplete or inconsistent data.  
Contingency: Implement robust data cleaning procedures and document any data quality issues.
2. Resource limitations: Insufficient computational resources may delay the project.  
Contingency: Plan for resource allocation.
3. Technical challenges: Issues with software tools or modeling techniques.

Contingency: Have alternative tools or methods ready and seek technical support when needed.

- Terminology

1. ClinVar: A public database containing annotations about human genetic variants.
2. SIFT scores: Scores from the Sorting Intolerant From Tolerant algorithm, predicting whether an amino acid substitution affects protein function.
3. Pathogenic: Likely to cause disease.
4. Benign: Not likely to cause disease.
5. Uncertain significance: The effects of the variant is not clear.
6. Conflicting classifications: Different clinical laboratories have provided different classifications for the same genetic variant.

- Costs and benefits

1. Costs:

Time: Significant time investment required for data cleaning, analysis, and model development.

Resources: Computational resources.

Personnel: The group

2. Benefits:

Improved clinical decision-making: Better understanding of conflicting classifications can aid in more accurate diagnoses and treatments.

Research advancements: Insights into genetic variants can contribute to research on disease mechanisms and drug discovery.

- Defining your data-mining goals

- Data-mining goals

Objective: Predict whether a genetic variant will have conflicting clinical classifications.

Specific Goals:

1. Data Exploration and Cleaning:

Perform initial exploration and cleaning of the ClinVar dataset to ensure data quality and reliability.

2. Data Visualization:

Create visualizations to understand patterns and distributions in the data, particularly focusing on conflicting classifications.

3. Frequent Pattern Mining:

Identify frequent patterns and correlations in the genetic variants that are associated with conflicting classifications.

#### 4. Predictive Modeling:

Develop and validate predictive models to classify genetic variants based on the likelihood of having conflicting classifications.

#### 5. Chromosome Grouping:

Group chromosomes based on the complexity of classifying the variants of the genes located on them, and connect this to gene effects (SIFT scores) and associated diseases.

#### 6. Population Analysis:

Analyze and visualize the frequency of variants across different populations using tools like Tableau.

- Data-mining success criteria

##### 1. Quantitative:

- a. Prediction accuracy of at least 85%
- b. Chromosome grouping accuracy of 90%

##### 2. Qualitative:

- a. Insight generation: Identify patterns and factors associated with conflicting classifications of genetic variants
- b. Usability of visualizations: Create clear and interpretable visualizations that help understand patterns in conflicting classifications.
- c. Usability of data: Ensure robustness and relevance of findings related to conflicting classifications.
- d. Population Analysis: Analyze and visualize the frequency of variants across different populations.

## Data understanding

- Gathering data

- Outline data requirements

The project requires genetic variant data like chromosome numbers (CHROM), genomic positions (POS), and alleles (REF, ALT). Allele frequency data (AF\_ESP, AF\_EXAC, AF\_TGP) is needed for population analysis. Clinical annotations (CLNSIG, CLNDISDB) are crucial for understanding conflicting classifications, while functional scores (SIFT, PolyPhen) and molecular annotations (SYMBOL,

BIOTYPE) provide additional insights. The dataset must be comprehensive, in CSV format, and cover current data.

- Verify data availability:

The primary dataset is `clinvar_conflicting.csv` from Kaggle, with 354,471 entries across 46 columns. Key fields like CHROM, POS, REF, ALT, and CLNSIG are present and complete. Columns with over 99% missing data will likely be excluded. External data like ExAC or gnomAD can supplement allele frequencies if necessary. The dataset is accessible, compatible with Python, and sufficient for most project goals.

Our data is available on Kaggle:

<https://www.kaggle.com/datasets/kevinarvai/clinvar-conflicting/data>

The raw ClinVar vcf file can be found:

[ftp://ftp.ncbi.nlm.nih.gov/pub/clinvar/vcf\\_GRCh37/clinvar.vcf.gz](ftp://ftp.ncbi.nlm.nih.gov/pub/clinvar/vcf_GRCh37/clinvar.vcf.gz) (this is now archived)

The raw files updated in 2024 and archives can now be found here:

<https://ftp.ncbi.nlm.nih.gov/pub/clinvar/>

- Define selection criteria

The dataset's key fields (CHROM, POS, REF, ALT, AF\_ESP, CLNSIG) will be used for analysis. We have some additional tasks that we want to achieve like All 354,471 entries will be included as the size is manageable. Semi-structured fields like INFO require parsing to extract metadata, such as allele frequencies and clinical significance. High-quality fields will be prioritized, and sparse columns excluded to streamline the workflow.

## • Describing and exploring data

**Source:** The dataset, `clinvar_conflicting.csv`, originates from the ClinVar database, which aggregates information on genetic variants and their clinical significance contributed by various clinical laboratories.

**Format:** The file is in CSV (comma separated values) - can be loaded to python by pandas

**Structure:**

**Number of Cases (Rows):** The dataset contains 354,471 entries (variants).

**Number of Fields (Columns):** There are 46 columns, each representing a key attribute of the genetic variant:

['CHROM', 'POS', 'REF', 'ALT', 'AF\_ESP', 'AF\_EXAC', 'AF\_TGP', 'CLNDISDB', 'CLNDISDBINCL', 'CLNDN', 'CLNDNINCL', 'CLNHGVS', 'CLNSIGINCL', 'CLNVC', 'CLNVI', 'MC', 'ORIGIN', 'SSR', 'CLASS', 'Allele', 'Consequence',

'IMPACT', 'SYMBOL', 'Feature\_type', 'Feature', 'BIOTYPE', 'EXON', 'INTRON', 'cDNA\_position', 'CDS\_position', 'Protein\_position', 'Amino\_acids', 'Codons', 'DISTANCE', 'STRAND', 'BAM\_EDIT', 'SIFT', 'PolyPhen', 'MOTIF\_NAME', 'MOTIF\_POS', 'HIGH\_INF\_POS', 'MOTIF\_SCORE\_CHANGE', 'LoFtool', 'CADD\_PHRED', 'CADD\_RAW', 'BLOSUM62']

More information on many of the features can be found at these two links:

[https://useast.ensembl.org/info/docs/tools/vep/vep\\_formats.html#output](https://useast.ensembl.org/info/docs/tools/vep/vep_formats.html#output)

[https://useast.ensembl.org/info/genome/variation/prediction/predicted\\_data.html#consequences](https://useast.ensembl.org/info/genome/variation/prediction/predicted_data.html#consequences)

- Verifying data quality

Key fields like CHROM, POS, REF, ALT, and CLNSIG are complete and reliable. We have more than enough reliable features in our dataset that we could use for different tasks. However, some fields, such as CLNDISDBINCL and MOTIF\_NAME, have over 99% missing values. We will be careful with them and potentially will be excluded from analysis. Outliers in allele frequency fields (AF\_ESP, AF\_EXAC) and semi-structured data in the INFO column require preprocessing.

---

Additional Dataset Information:

**Main Dataset:**

We are going to use the main dataset in our project.

<https://www.kaggle.com/datasets/kevinarvai/clinvar-conflicting/data>

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 65188 entries, 0 to 65187
Data columns (total 46 columns):
 #   Column                Non-Null Count  Dtype
---  -
 0   CHROM                  65188 non-null  object
 1   POS                    65188 non-null  int64
 2   REF                    65188 non-null  object
 3   ALT                    65188 non-null  object
 4   AF_ESP                 65188 non-null  float64
 5   AF_EXAC                65188 non-null  float64
 6   AF_TGP                 65188 non-null  float64
 7   CLNDISDB               65188 non-null  object
 8   CLNDISDBINCL          167 non-null    object
```

```

9   CLNDN                65188 non-null object
10  CLNDNINCL            167 non-null object
11  CLNHGVS              65188 non-null object
12  CLNSIGINCL           167 non-null object
13  CLNVC                65188 non-null object
14  CLNVI                27659 non-null object
15  MC                   64342 non-null object
16  ORIGIN               65188 non-null int64
17  SSR                  130 non-null float64
18  CLASS                65188 non-null int64
19  Allele               65188 non-null object
20  Consequence          65188 non-null object
21  IMPACT               65188 non-null object
22  SYMBOL               65172 non-null object
23  Feature_type         65174 non-null object
24  Feature              65174 non-null object
25  BIOTYPE              65172 non-null object
26  EXON                 56295 non-null object
27  INTRON               8803 non-null object
28  cDNA_position        56304 non-null object
29  CDS_position         55233 non-null object
30  Protein_position     55233 non-null object
31  Amino_acids          55184 non-null object
32  Codons               55184 non-null object
33  DISTANCE             108 non-null float64
34  STRAND               65174 non-null float64
35  BAM_EDIT             31969 non-null object
36  SIFT                 24836 non-null object
37  PolyPhen             24796 non-null object
38  MOTIF_NAME           2 non-null object
39  MOTIF_POS            2 non-null float64
40  HIGH_INF_POS         2 non-null object
41  MOTIF_SCORE_CHANGE  2 non-null float64
42  LoFtool              60975 non-null float64
43  CADD_PHRED           64096 non-null float64
44  CADD_RAW             64096 non-null float64
45  BLOSUM62             25593 non-null float64
dtypes: float64(12), int64(3), object(31)
memory usage: 22.9+ MB

```

### Raw Dataset:

Although we do not use the raw data, we wanted to have more information on it. the author has not documented when he has downloaded and used the raw dataset. But from archive here is details of a raw dataset from 2019:

It has 440,147 entries

File name/link: clinvar\_20190108.vcf.gz

head:

```
VCF Data:
```

```

0      1      2 3 4 5 6 \
0 1 949422 475283 G A . .
1 1 949502 542074 C T . .
2 1 949523 183381 C T . .
3 1 949559 542075 C T . .
4 1 949597 475278 C T . .

7
0 AF_ESP=0.00546;AF_EXAC=0.00165;AF_TGP=0.00619;...
1 AF_ESP=0.00015;AF_EXAC=0.00010;ALLELEID=514926...
2 ALLELEID=181485;CLNDISDB=MedGen:C4015293,OMIM:...
3 ALLELEID=514896;CLNDISDB=MedGen:C4015293,OMIM:...
4 AF_ESP=0.00515;AF_EXAC=0.00831;AF_TGP=0.00339;...

```

Data information:

```

--- Dataset Information ---
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 440147 entries, 0 to 440146
Data columns (total 8 columns):
#   Column  Non-Null Count  Dtype
---  -
0   CHROM    440147 non-null  object
1   POS      440147 non-null  int64
2   ID       440147 non-null  int64
3   REF      440147 non-null  object
4   ALT      440147 non-null  object
5   QUAL     440147 non-null  object
6   FILTER   440147 non-null  object
7   INFO     440147 non-null  object
dtypes: int64(2), object(6)
memory usage: 26.9+ MB
None

```

And here is the raw dataset from 2024:

354,471 entries

head:

```

VCF Data:
0      1      2 3 4 5 6 \
0 1 949422 475283 G A . .
1 1 949523 183381 C T . .
2 1 949597 475278 C T . .
3 1 949608 402986 G A . .
4 1 949696 161455 C CG . .

7
0 ALLELEID=446939;CLNDISDB=MedGen:C4015293,OMIM:...
1 ALLELEID=181485;CLNDISDB=MedGen:C4015293,OMIM:...

```



```
2 ALLELEID=446987;CLNDISDB=MedGen:C4015293,OMIM:...
3 ALLELEID=389314;CLNDISDB=MedGen:CN169374;CLNDN...
```

```
4 ALLELEID=171289;CLNDISDB=MedGen:C4015293,OMIM:...
```

```
Dataset information:
--- Dataset Information ---
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 354471 entries, 0 to 354470
Data columns (total 8 columns):
#   Column  Non-Null Count  Dtype
---  -
0   CHROM    354471 non-null    object
1   POS      354471 non-null    int64
2   ID       354471 non-null    int64
3   REF      354471 non-null    object
4   ALT      354471 non-null    object
5   QUAL     354471 non-null    object
6   FILTER   354471 non-null    object
7   INFO     354471 non-null    object
dtypes: int64(2), object(6)
memory usage: 21.6+ MB
```

## Planning your project

- Make a detailed plan of your project with a list of tasks. There should be at least five tasks. Specify how many hours each team member will contribute to each task.

	Task	Team Members	Hours
1	Data cleaning	Ilze, Liene, Roza	10
2	Data exploration - data visualizations to understand patterns	Liene, Roza	12
3	Frequent pattern mining for variant analysis,	Ilze, Liene	12
4	Modelling; predicting whether a ClinVar variant will have conflicting classifications	Ilze, Liene, Roza	14
5	Group chromosomes based on the complexity of classifying the	Ilze, Liene, Roza	14

	Task	Team Members	Hours
	variants of the genes located on them		
6	Connecting findings of task 5 to the genes, gene effects (SIFT) and diseases they are linked to	Ilze, Roza	12
7	Analyze variant frequency across different populations (potentially using tableau or geopandas).	Liene, Roza	12
8	Final conclusions, report	Ilze, Liene, Roza	14

- List the methods and tools that you plan to use. Add any comments about the tasks that you think are important to clarify.

We are going to use pandas and numpy python library. We all work in jupyter notebook. Potentially using tableau or geopandas for some tasks like analysing variant frequency across different populations. matplotlib and seaborn for data visualization. For predictive modeling, we will rely on scikit-learn, with methods like logistic regression, random forests, and hyperparameter tuning using GridSearchCV. If needed other tools will be added.