Unsupervised Learning

IBM Machine Learning - Project
Mohammed Qasim K.
October 2022

Main Objective

- The main objective of this analysis was to classify the Chromosome on which, the particular genetic variation was present using Classification Methods
- The data set was split into Training set (80%) and Test set (20%) using Train-Test-Split.

About the Data

- The data set comprises of information about human genetic variants.
- This data set has 65188 records and 46 variables.

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

# CHROM	A REF	# AF_ESP	A CLNVC	# ORIGIN
Chromosome the variant is located on	Reference Allele	Allele frequencies from GO-ESP	Variant Type	Allele origin. One or more of the following values may be added: 0 - unknown; 1 - germline; 2 - somatic; 4 - inherited; 8 -

A BIOTYPE	A Amino_acids	A Codons	# STRAND	# LoFtool
Biotype of transcript or regulatory feature	only given if the variant affects the protein-coding sequence	the alternative codons with the variant base in upper case	defined as + (forward) or - (reverse).	Loss of Function tolerance score for loss of function variants: https://github.com/konradjk/loftee

A Allele	A Consequence	A IMPACT	A SYMBOL	A Feature
the variant allele used to calculate the consequence	Type of consequence: https://useast.ensembl.org/info/genome/variation/prediction/predicted_data.html#consequences	the impact modifier for the consequence type	Gene Name	Ensembl stable ID of feature

Data Exploration

- After checking for duplicates, the EDA was conducted on the data set.
- The data description is as follows:

```
CHROM                    0
POS                      0
REF                      0
ALT                      0
AF_ESP                   0
AF_EXAC                  0
AF_TGP                   0
CLNDISDB                 0
CLNDISDBINCL         65021
CLNDN                    0
CLNDNINCL            65021
CLNHGVS                  0
CLNSIGINCL           65021
CLNVC                    0
CLNVI                37529
MC                     846
ORIGIN                   0
SSR                  65058
CLASS                    0
Allele                   0
Consequence              0
IMPACT                   0
SYMBOL                  16
Feature_type            14
Feature                 14
BIOTYPE                 16
EXON                  8893
INTRON               56385
cDNA_position         8884
CDS_position          9955
Protein_position      9955
Amino_acids          10004
Codons               10004
DISTANCE             65080
STRAND                  14
BAM_EDIT             33219
SIFT                 40352
PolyPhen             40392
MOTIF_NAME           65186
MOTIF_POS            65186
HIGH_INF_POS         65186
MOTIF_SCORE_CHANGE   65186
LoFtool               4213
CADD_PHRED            1092
CADD_RAW              1092
BLOSUM62             39595
dtype: int64
```

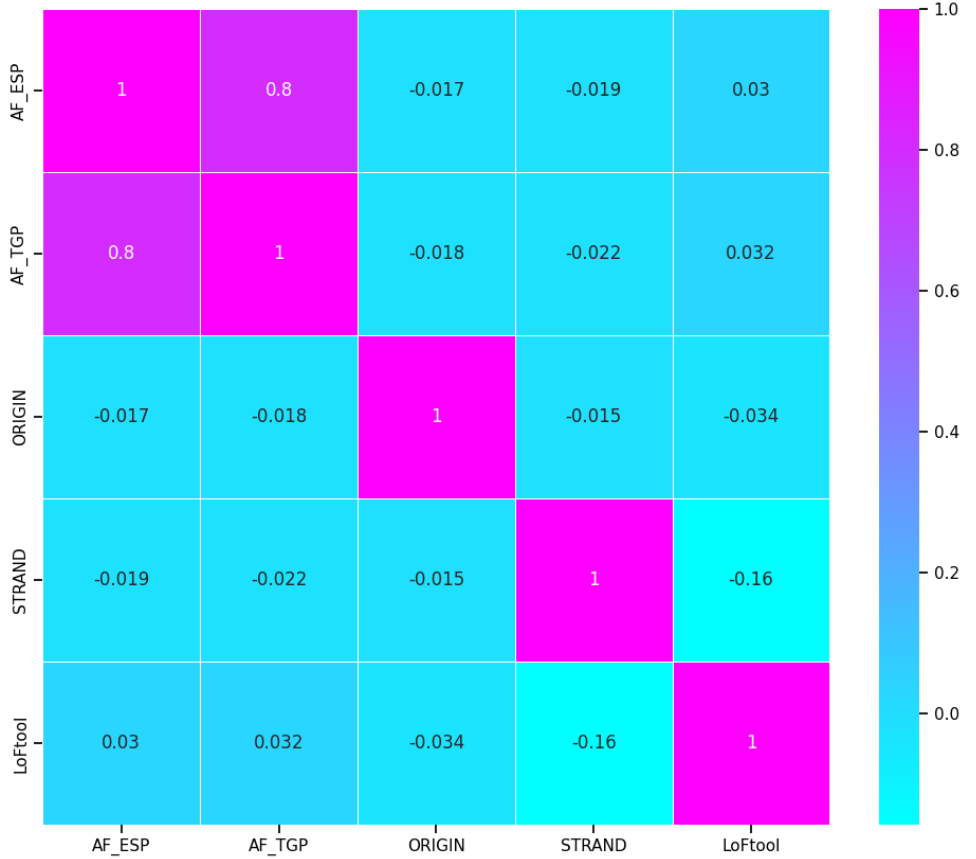
Data Exploration

- Calculated the percentage of missing data for all columns and dropped columns with greater than 20% missing data.
- A total of 20 variables are present in the current dataset.
- The data description after feature engineering is as follows:

```
CHROM              0
REF                0
ALT                0
AF_ESP             0
AF_TGP             0
CLNVC              0
MC               846
ORIGIN             0
CLASS              0
Allele             0
Consequence        0
IMPACT             0
SYMBOL            16
Feature_type      14
Feature           14
BIOTYPE           16
Amino_acids    10004
Codons         10004
STRAND            14
LoFtool         4213
dtype: int64
```

Data Exploration

- Plotted a heat map to check collinearity.
- Dropped *AF_TGP* column due to severe correlation.



Data Exploration

- Column were checked for unique values and columns with greater than 3000 unique values were dropped.
- Rows with CLASS value equal to 1 implying a conflicting classification were dropped along with CLASS column.
- The data set currently has 48754 records and 32 variables.

	0
0	CHROM
1	REF
2	ALT
3	AF_ESP
4	CLNDISDBINCL
5	CLNDNINCL
6	CLNSIGINCL
7	CLNVC
8	MC
9	ORIGIN
10	SSR
11	Allele
12	Consequence
13	IMPACT
14	SYMBOL
15	Feature_type
16	Feature
17	BIOTYPE
18	INTRON
19	Amino_acids
20	Codons
21	DISTANCE
22	STRAND
23	BAM_EDIT
24	SIFT
25	PolyPhen
26	MOTIF_NAME
27	MOTIF_POS
28	HIGH_INF_POS
29	MOTIF_SCORE_CHANGE
30	LoFtool
31	BLOSUM62

Data Exploration

- Column were checked for missing values and columns with greater than 20 % missing values were dropped.
- Other columns with missing values were filled using a mean or mode method depending on the data type.
- Further analysis into the data found the *Feature_type* column to be obsolete and was dropped.
- The data set currently has 48754 records and 17 variables.

	0
0	CHROM
1	REF
2	ALT
3	AF_ESP
4	CLNVC
5	MC
6	ORIGIN
7	Allele
8	Consequence
9	IMPACT
10	SYMBOL
11	Feature
12	BIOTYPE
13	Amino_acids
14	Codons
15	STRAND
16	LoFtool

```
MOTIF_POS              100.00
MOTIF_SCORE_CHANGE     100.00
HIGH_INF_POS           100.00
MOTIF_NAME             100.00
DISTANCE                99.85
CLNDISDBINCL            99.80
CLNDNINCL               99.80
CLNSIGINCL              99.80
SSR                     99.79
INTRON                  86.70
PolyPhen                62.50
SIFT                    62.47
BLOSUM62                61.33
BAM_EDIT                50.71
Amino_acids             14.86
Codons                  14.86
LoFtool                  6.33
MC                       1.21
STRAND                   0.01
Feature                  0.01
BIOTYPE                  0.01
Feature_type             0.01
SYMBOL                   0.01
REF                      0.00
IMPACT                   0.00
...
CLNVC                    0.00
AF_ESP                   0.00
ALT                      0.00
CHROM                    0.00
Name: Percentage of Missing, dtype: float64
```

Data Exploration

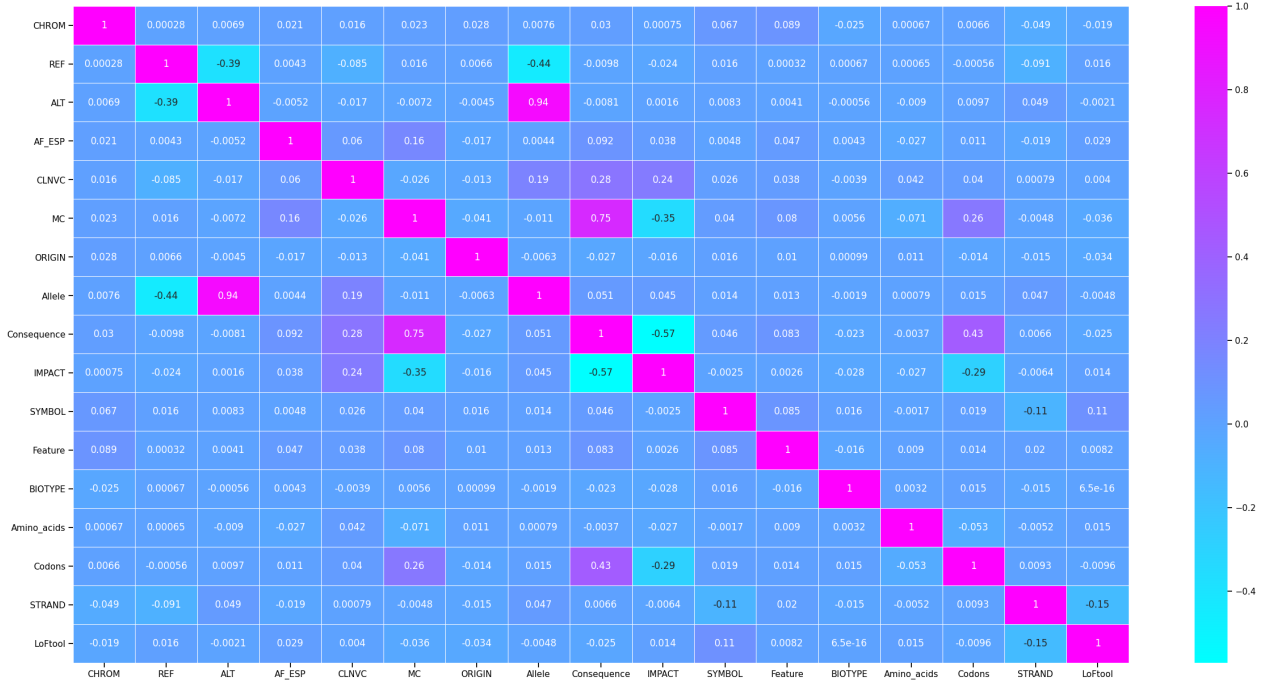
- Binary and Label Encoding was performed on the data set appropriately.
- The data set currently has 48754 records and 17 variables.

	CHROM	REF	ALT	AF_ESP	CLNVC	MC	ORIGIN	Allele	Consequence	IMPACT	SYMBOL	Feature	BIOTYPE	Amino_acids	Codons	STRAND	LoFtool
0	0	351	92	0.08	6	17	1	77	14	2	185	1568	1	227	1497	1.00	0.35
1	0	351	0	0.00	6	17	1	1	14	2	1975	426	1	662	1039	-1.00	0.35
3	0	351	0	0.00	6	17	33	1	14	2	1757	1744	1	334	469	1.00	0.35
4	0	351	281	0.00	6	17	33	246	14	2	1757	1744	1	322	471	1.00	0.35
5	0	351	92	0.00	6	17	33	77	14	2	1757	1744	1	333	470	1.00	0.35
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
65182	23	180	281	0.02	6	78	1	246	43	1	664	59	1	655	1199	-1.00	0.00
65183	23	515	178	0.08	6	78	1	135	43	1	664	59	1	844	1880	-1.00	0.00
65185	23	180	281	0.01	6	78	1	246	43	1	664	59	1	67	1586	-1.00	0.00
65186	23	515	92	0.00	6	78	1	77	43	1	1589	1657	1	929	807	-1.00	0.35
65187	23	351	92	0.00	6	17	1	77	14	2	385	2183	1	657	298	-1.00	0.14

48754 rows × 17 columns

Data Exploration

- Plotted a heat map to check collinearity.
- Dropped *ALT & MC* column due to severe correlation.



Data Exploration

- The target label is the *CHROM* column where:
  - 0 : Chromosome 8
  - 1: Chromosome 2
- The data set currently has 7979 records and 15 variables.
- Min Max Scaling was performed on the dataset.

	CHROM	REF	AF_ESP	CLNVC	ORIGIN	Allele	Consequence	IMPACT	SYMBOL	Feature	BIOTYPE	Amino_acids	Codons	STRAND	LoFtool
4629	0	180	0.00	6	1	246	14	2	1724	828	1	937	679	1.00	0.28
4630	0	351	0.00	6	1	1	14	2	155	445	1	333	465	1.00	0.14
4631	0	0	0.00	6	1	77	14	2	155	445	1	226	1375	1.00	0.14
4632	0	351	0.00	6	1	1	14	2	155	445	1	101	448	1.00	0.14
4633	0	180	0.00	6	1	1	14	2	155	445	1	87	1396	1.00	0.14
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
16503	1	180	0.00	6	1	246	28	1	22	2136	1	511	315	1.00	0.29
16506	1	515	0.06	6	1	77	28	1	22	2136	1	511	315	1.00	0.29
16508	1	180	0.00	6	1	246	14	2	22	2136	1	806	322	1.00	0.29
16509	1	351	0.00	6	1	1	14	2	22	2136	1	334	469	1.00	0.29
16510	1	180	0.05	6	1	246	13	3	22	2136	1	511	315	1.00	0.29

7979 rows × 15 columns

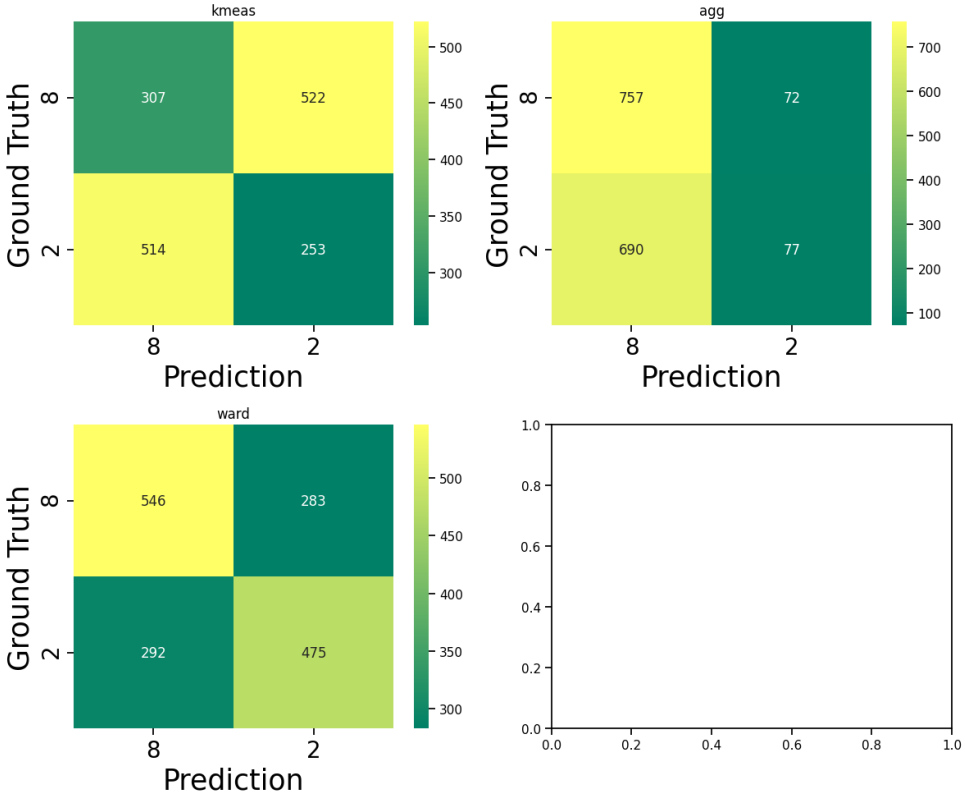
Predictions on the Test Set

The three models were then fit on the unseen Validation Set and the appropriate evaluation metrics were calculated for each model.

- K means with *n_clusters* = 2
- Agglomerative Clustering with *complete* linkage and *n_clusters* = 2
- Agglomerative Clustering with *ward* linkage and *n_clusters* = 2

Comparing the Metrics

- The confusion matrix shows that Agglomerative Clustering with *ward* linkage performed decently with less misclassification
- Agglomerative Clustering with *complete* linkage performed the worst on the validation set classifying most of the data belonging to CHROM = 8.
- The best classification model is *ward* linkage (True Positives= 0.62 & True Negatives = 0.66)



Comparing the Metrics

- The metrics shows that Agglomerative Clustering with *ward* linkage performed decently.
- Agglomerative Clustering with *complete* linkage performed the worst classifying most of the data as CHROM = 8.
- The best classification model is *ward* linkage (Avg F1-score = 0.64)

K Means

	0	1
precision	0.37	0.33
recall	0.37	0.33
f1-score	0.37	0.33

Complete Linkage

	0	1
precision	0.52	0.52
recall	0.91	0.10
f1-score	0.67	0.17

Ward Linkage

	0	1
precision	0.65	0.63
recall	0.66	0.62
f1-score	0.66	0.62

Conclusion

The analysis shows that Feature Engineering has a large effect on the model performance. The best classification model for this application is Agglomerative Clustering with *ward* linkage. Although performance by the model wasn't great improvements can be made by choosing a dataset with a closer ratio between the two classification classes, trying other classification methods not employed in this analysis or using more features which were dropped for the sake of simplicity and computational resources.

Jupyter Notebook can be found here: https://github.com/moqa19/IBM-Machine-Learning/blob/main/ML%204.ipynb