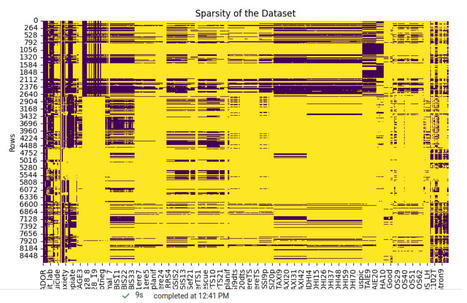
Unsupervised analysis of the relationship between various genetic polymorphisms and impulse-aggresive personality traits

Jorge Parreño Hernández, Filippo Nardi, Riccardo Conforto Galli

Data Science, Universidad Carlos III de Madrid

Preprocessing

The original dataset was very sparse, a heatmap is displayed below.



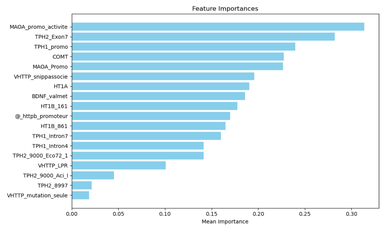
Feature selection

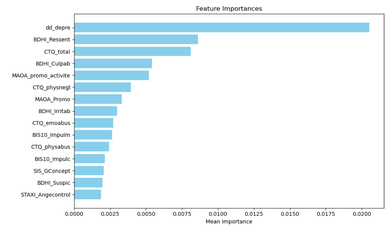
Original dataset contained 604 features. Many of them were redundant, some were questions to standardized personality tests, at the end we chose the following

- **Type, suicide and gender**: suicidal personality type, attempt date and gender
- **Standardized test information**: aggregated scores for BDHI and BIS tests
- **Polymorphisms**: data regarding various genetic polymorphisms closely related to suicidal ideation

Random Forest Feature Importances

We utilized K-Nearest Neighbor for data imputation, followed by applying Random Forest using two distinct subsets: one primarily focused on Polymorphism and the other emphasizing BDHIS tests. Here are the results showcasing Feature Importances, sorted by Mean Decrease Index:





Dimensionality reduction

Even with feature selection, our data still presented very high dimensionality, thus we resorted to various dimensionality reduction techniques, the most optimal being UMAP (Uniform Manifold Approximation Projection), for its topology preserving properties.

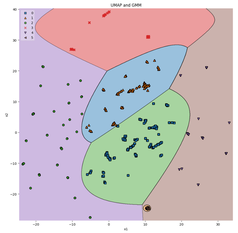
As information regarding polymorphisms takes string values, we employed the BERT model, along with PCA in order to avoid having to impute.

GMM clustering

Gaussian Mixture Models (GMMs) are probabilistic models used for clustering data. Their probability density function is given by:

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

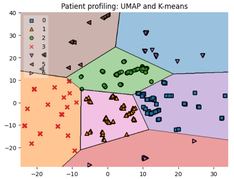
where K is the number of components, which we optimized using Bayesian information criterion, π_k is the weight of the k-th component, and $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ is the Gaussian distribution. The result is as follows.



We also study the correlations of the initial variables with the UMAP components to know which variables play more of a role in each component so that we can effectively group.

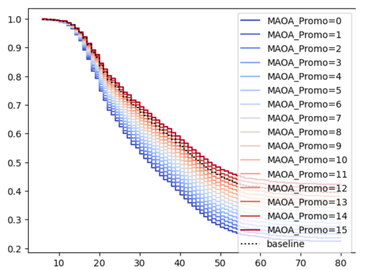
K-means clustering

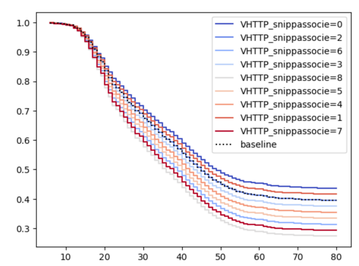
We also employed perhaps the most known clustering algorithm, on two dimensional data, using the Silhouette score as the criteria for selecting the number of components. The results are similar as K-means is nothing but a mixture model where the covariance matrices of the clusters are isotropic and the weights are uniform.



Survival Analysis

We conducted Survival Analysis, employing K-NN algorithms for data imputation and utilizing the Cox model for survival analysis. These results validate the hypothesis posited by the Random Forest method.





Genomic phenotyping using fp-tree

We employed the fp-tree algorithm to find the support for different values of sets of 2 and 3 polymorphisms for different personality types. Thus given some polymorphisms we can, to the extent of our data, know which is the most likely profile that patient belongs to, working backwards, it also aids in clinical analysis.

Contributions

1. Through dimensionality reduction and clustering, we perform patient profiling and enhance our understanding of underlying patterns.
2. Through frequent itemset computation we can better understand the patient profiles and the polymorphism distribution for the different types.
3. Through feature importance analysis using random forest, we find the most important features for determining personality types and improve interpretability.
4. Through survival analysis we improve the understanding and prediction of suicidal attempt for different patient profiles depending solely on genetic polymorphisms.

Source code

[1] Complete code.
`https://github.com/mrjorgeparr/ML-in-healthcare/tree/main/Final%20project`.
Accessed: December 15, 2023.