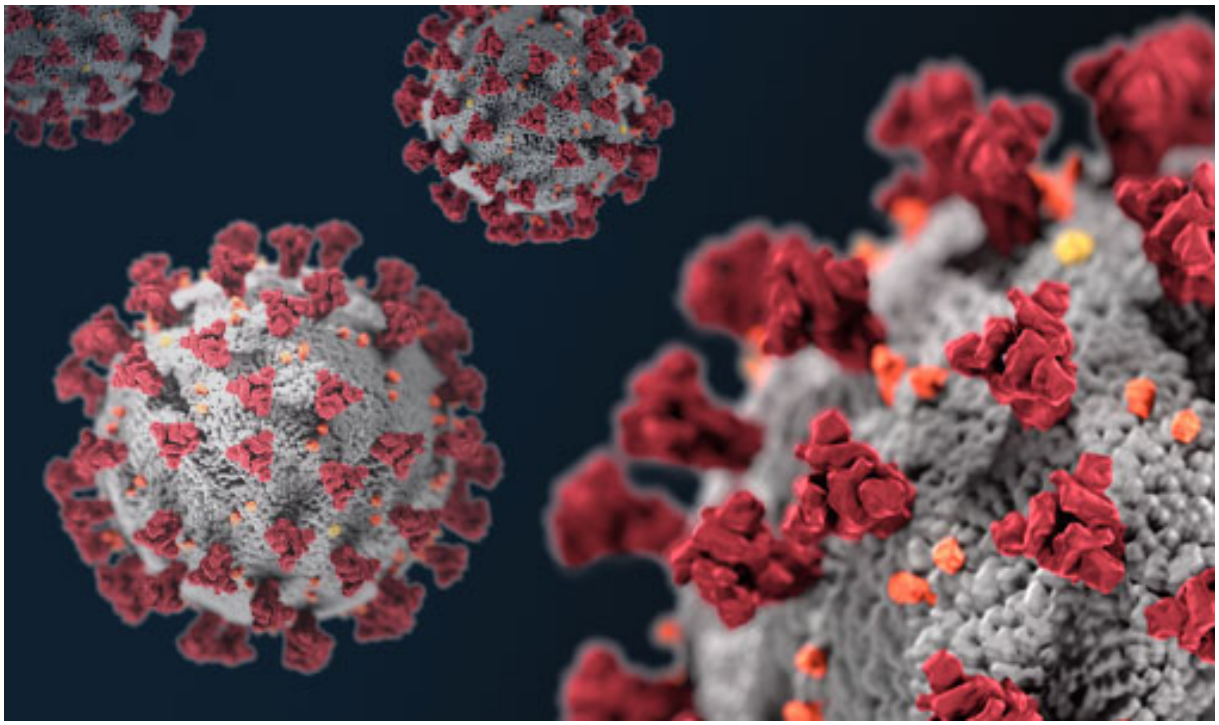
EPFL

Data Science In Practice

MGT - 415

Project

COVID-19 : Risk-factor Prediction



Students
Rayan Chaouche
Yann Martinson
Christopher Padovani
Jules Triomphe

Professor
Dr. Christopher Bruffaerts

Doctoral Assistant
Omar Ballester

May 11, 2020

Contents

Introduction

Coronavirus disease 2019 (COVID-19 or **SARS-CoV-2**) is an infectious disease that first appeared in December 2019 in Wuhan, China [6]. Since then, the virus has been spreading worldwide causing almost 4 million confirmed cases and over 265,000 deaths [3]. In a few months, the disease reached almost every country and exposed the limits of national healthcare systems and international cooperation, while also expanding the influence of disinformation and the impact of poor leadership.

For the moment, the only measure to prevent overcrowded hospitals has been for the authorities to order lock-downs around the world, to stop the propagation of the virus and protect the most vulnerable parts of the population. As of April 1st, over a third of the world population was asked to stay home [12]. This phenomenon was magnified as a result of poor international cooperation. Indeed, countries mostly adopted self-preservation policies first and only later started talking with their neighbours and the global community to preserve the trade of vital merchandise and food. Consequently, large pans of the economy were shut down from one day to the next, causing a global economic slowdown [11]. This decision, though apparently cautious in the short term, is nefarious in the long term. Indeed, it has been shown that life expectancy is largely correlated with economic growth [8]. Hence, a global recession will lead to many deaths in the medium to long term which is why the economy must be restarted. This means, unfortunately, that people will keep on being infected by the coronavirus disease. To limit the number of deaths, it is important to identify the greatest risks related to infections and subsequent deaths so that they may be mitigated.

Fortunately, the knowledge of the virus grows daily. Thousands of studies are conducted around the world in hope of developing treatments and vaccines. Among these studies, some focus on how the disease affects people based on several factors such as patient age and comorbidities. When diagnosing a patient however, other than external symptoms such as cough and fever, tests are done. These tests aim to identify if certain targets are expressed. Hence, it is important to know which targets to look for to determine the level of risk of a patient to an infection of COVID-19 and whether or not to contain them or have them isolate themselves. Moreover, the fewer targets are needed to identify the risk, the more affordable the tests. Given the magnitude of testing campaigns which will be done and the cost of lock-downs, the price of conducting risk-analyses of patients to COVID-19 is an important economic issue with high stakes.

Chapter 1

Problem description

1.1 Understanding the need

COVID-19 testing campaigns will be expensive, but they are necessary to allow a restart of the economy. In the US for example, losses are estimated to be anywhere between $300 billion to $400 billion per month [13]. Hence a testing campaign costing $100 billion but enabling workers to return to work and thus relaunching the economy would be a "modest investment" according to Rockefeller Foundation president Rajiv Shah [13].

Today, costs for genome tests to identify COVID-19 are over $100 [10]. In Switzerland, there have been a few more than 35 tests per 1'000 people [2] as of May 10th. With a population of 8'544'527[14], this represents almost $30 million worth of tests. At a global scale, the represents over $27 billion [7]. Hence, a small price reduction due to an increased efficiency of tests or the identification of people at risk with less costly tests will have a huge impact. This impact will be at least two-fold : the investments required to conduct testing campaigns will decrease and the number of people who may be tested will increase. With countries like France thinking about launching a mass screening campaign [1] and the number of tests per 1'000 people ever increasing [2], the need for increased testing efficiency and risk-identification is ever growing.

1.2 Challenges

To facilitate risk-prediction, the main challenge outside of identifying otherwise obvious traits like age and obesity is to identify the main drug targets associated to disease risk-factors for COVID-19, so that they can be targeted in a test.

Hence, the main challenges of this study are to :

- Create a model to identify which targets are most likely to be risk-factors for COVID-19 based on the association of diseases to COVID-19 and their association score to these targets.
- Design a map of the major targets drugs for COVID-19 with the use of networks.

1.3 Strategy

This project aims to identify and make links between risk-factors (both diseases and drug targets) and COVID-19 based on preexisting studies and databases.

By finding similarities between diseases and COVID-19, it is possible to identify which diseases are closest to COVID-19. This is done with a network analysis based on identified genetic relations between diseases. The relations between diseases are extracted from a previous study [9].

Next, it is possible to extract the association scores of diseases related to COVID-19 and specific drug targets by leveraging the Open Targets Platform API. Association scores between diseases and targets, and between diseases and COVID-19 can thus be put in relation. The use of a Random Forest Regressor creates a model from which the most important drug targets may be identified.

1.4 Expectations

Two outputs are expected from this challenging project :

- The first expected output is an accurate network of the relation between COVID-19 and other diseases through the expression of a selection of drug targets.
- The second expected output is an accurate prediction of the most important target risk-factors for COVID-19.

The network correspondence between the disease, target genes and COVID-19 can help scientist figure out gene families. The expected communities could prove useful by providing a focus on the most represented genes. Moreover, by identifying the closest diseases from the coronavirus, scientists could conceivably try affecting COVID-19 with drugs affecting genes close to it.

The database from the second output could help doctors know in which case they need to pay special attention to patients with identified targets in order to avoid additional infections and/or deaths due to COVID-19. It could also help searchers develop cheaper and more efficient tests to determine the degree of risk a person is exposed to with regards to the virus.

Chapter 2

Data

In response to the COVID-19 pandemic, Kaggle launched a competition : the COVID-19 Open Research Dataset Challenge (CORD-19) [4]. Kaggle is a web platform organizing data science competitions. On this platform, companies propose data science problems and offer a prize to the data scientists with the best performance [16].

2.1 Previous work (Kaggle)

The White House and a coalition of leading research groups have prepared the COVID-19 Open Research Data-set (CORD-19) [4] which is a resource of over 59,000 scholarly articles, including over 47,000 with full text, about COVID-19, SARS-CoV-2, and related coronaviruses. This freely available data-set is provided to the global research community to apply recent advances in natural language processing and other AI techniques to generate new insights in support of the ongoing fight against this infectious disease.

In addition to this data-set, two other sources were used to reach the resultant data-sets from which this study is based upon : the Open Targets Platform database and the Mondo Disease Ontology database. The Open Targets Platform database is composed of aggregates and merges genetic associations curated from both literature and newly-derived loci[1] from UK Biobank, which contains functional genomics data and quantitative trait loci. The Mondo Disease Ontology [5] is a semi-automatically constructed ontology that merges in multiple disease resources to yield a coherent merged ontology.

Two datasets from Christophe Guéret's Kaggle project in response to the COVID-19 Open Research Dataset Challenge (CORD-19) [9] are used. The first is the "COVID_KG_sample.csv" data-set, later referred to as ISD1 for Initial Data-Set 1. It contains the relation between disease codes, drug targets and CORD-19 papers based on predicates.

The predicates are as follows :

- `isAboutGene` connects a CORD-19 paper to a target gene.

[1] A quantitative trait locus (QTL) is a locus (section of DNA) that correlates with variation of a quantitative trait in the phenotype of a population of organisms. QTLs are mapped by identifying which molecular markers (such as SNPs or AFLPs) correlate with an observed trait. This is often an early step in identifying and sequencing the actual genes that cause the trait variation. [15]

- `isAboutDisease` connects a CORD-19 paper to a disease (i.e. a characteristic, condition, or behaviour - following Open Targets terminology).
- `isAssociatedTo` connects a target gene to a Disease (i.e. a characteristic, condition, or behaviour - following Open Targets terminology).
- `belongsToTherapeuticArea` associates a Disease to its therapeutic area, as provided by Open Targets.
- `isASpecific` associates a disease to a higher-up element in the Open Targets hierarchy (ontological path connecting higher classes of diseases to more specific instances).
- `hasGeneticClue` connects a disease to a characteristic, condition, or behaviour if there are genetic evidences that such connection exists.

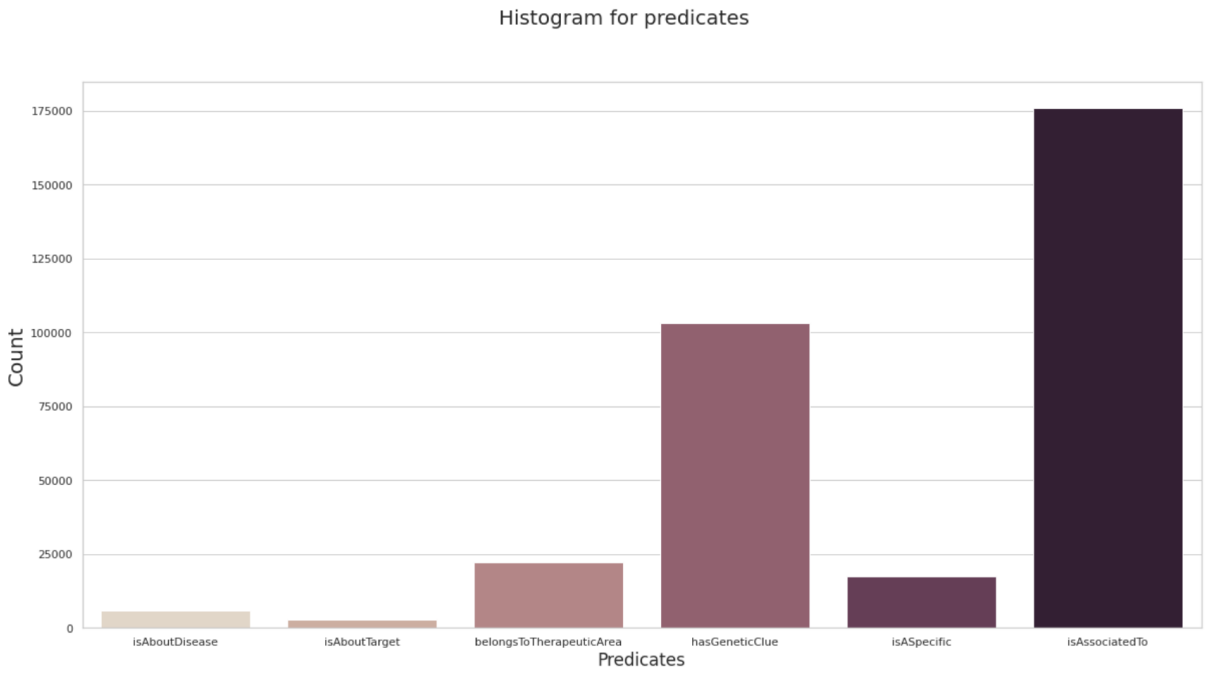


Figure 2.1: Histogram for predicates from the Kaggle project output

Figure 2.1 illustrates the predicates' occurrences depending on their type. A high occurrence of the `isAssociatedTo` predicate is observed. The predicates used in this study are the `hasGeneticClue` and `isASpecific`, however. The former is present in more than 10'000 rows, i.e there are over 10'000 genetic relations between targets and diseases.

The second data-set used is named "predicted_covid19_risk_factors.csv" and is built from a ComplEx model from the previously mentioned data-set including all the predicates except for all but 100 of the `hasGeneticClue` triples. It contains an association score of 7'217 diseases with COVID-19 and is later referred to as IDS2.

Figure 2.2 represents these association scores. The scores are mainly situated between 0.00 and 0.10. This means that the majority of diseases from the databases have little associations with COVID-19.

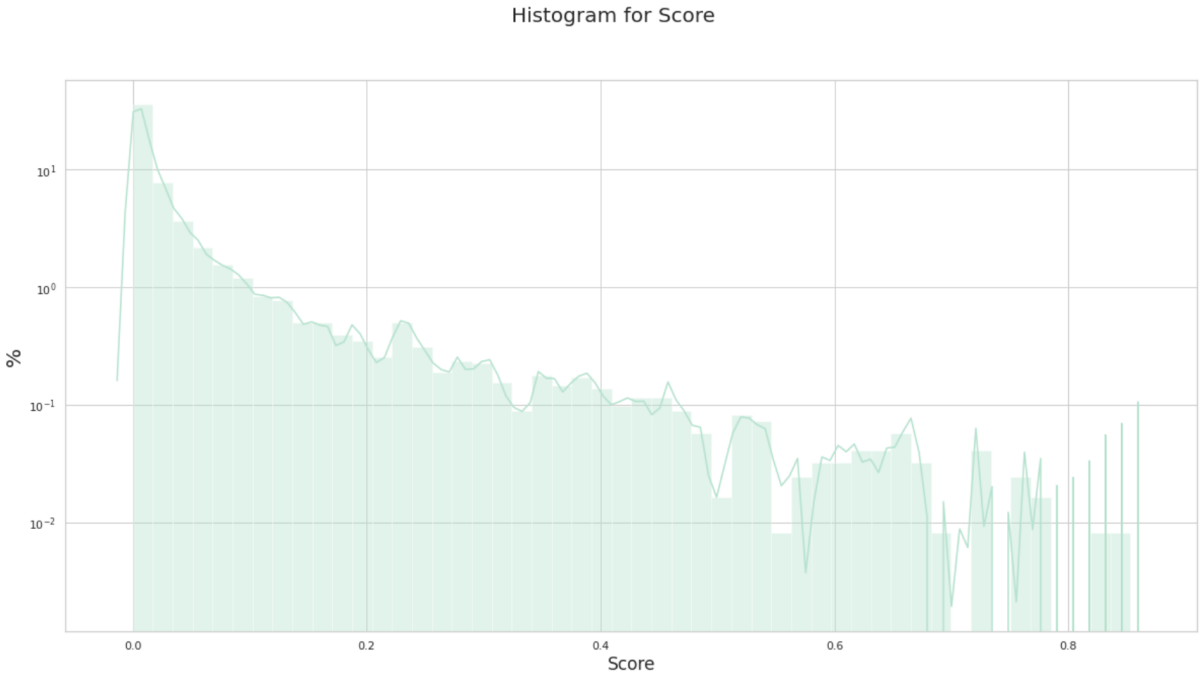


Figure 2.2: Score histogram from the Kaggle project output

2.2 Data-set to predict the most important target risk-factors

IDS2 is a list of full names of diseases and their computed association score to COVID-19. The aim is to create a data-set with these diseases, their association score to targets in the Open Targets Genetics database and their association score to COVID-19. Hence, the association score of diseases to targets needs to be retrieved. This is done using the `opentargets` python library which levers the Open Targets API.

Due to the high number (60'564) of targets and the low computational resources available, the data-set used in this study is created by chunks of 50 diseases (selected at random among the 7'217 diseases) and 100 targets. These 100 targets are selected at random among the 60'564 available ones as there is no knowledge as to whether some of them have particular relations to COVID-19. The final, full data-set is obtained with the merge of 144 of these 50 by 101 (last column for COVID-19) data-sets. It contains the 7'217 diseases and their association score to 100 targets as well as their association score to COVID-19. It is later referred to as DS1.

Chapter 3

Models

Two types of analyses are conducted. The first is a network analysis and the second is a random forest regression.

3.1 Network analysis

The network has been created using a combination of DS1 and IDS1.

In DS1, each disease has been assigned a weighted link with every target. As the model is studying 100 targets and COVID-19, at most 101 edges for are created for each possible disease. For simplicity and computational purposes, all zero weighted associations are avoided and do not appear in the network.

IDS1 considers relationships between diseases. The goal is to extend this data-set to establish a relationship between families/communities/clusters of diseases and COVID-19 while also establishing links with specific targets.

3.2 Random forest regression

For both grid search and random search algorithms, DS1 is used. It does not require particular pre-processing as it is made for these algorithms.

3.2.1 Grid search

To avoid excessive computational times, the focus is put on two parameters, namely the maximum tree depth and the number of estimators. The maximum tree depth is ranged from 5 to 15, and the number of estimators is selected from 10, 20, 50, 100, 200, 500 and 1'000. The minimum number of samples per leaf is also varied between one and two as this has proven to improve results by experimentation. Other parameters are kept to their default values.

The model is therefore run with 5 folds of cross validation for each of the 77 candidates, totalling 385 fits.

Once the model has run with the mentioned variations, features are ordered by importance and the model is run again with the best parameters and the fewest number of parameters required to reach 95% importance.

3.2.2 Random search

As the random search has the benefit of selecting a given number of parameters at random a given number of times, more parameters can be varied to obtain a final model. Hence, the number of estimators is varied from 10 to 1'000 with increments of 50, the maximum tree depth is varied from 5 to 20, the minimum number of samples required to split a node is ranged from 2 to 10 with increments of 2, the minimum number of samples per leaf is ranged from 1 to 5 and the method of computing the number of features considered for the best split is varied between the square root of the number of features, the logarithm and plainly the number of features considered.

The model is run with 5 folds of cross validation with 5 iterations, for a total of 25 fits. The chosen number of iterations emanates from the particularly large run time of these simulations. Provided with more time and more computational resources, it is recommended to increase the number of iterations made.

Once the model has run with the mentioned variations, features are ordered by importance and the model is run again with the best parameters and the fewest number of parameters required to reach 95% importance.

Chapter 4

Results

4.1 Analytical results of the models

4.1.1 Network analysis

The aim in this part is to observe the relationships between the expression of drug targets in a list of 4'000 diseases and the relationship between these diseases and COVID-19.

There are two types of provided data. On one side, a table links the diseases to the drug targets (DS1) via weights. The weights are normalized (between 0 and 1). This weighted edges matrix can be interpreted as an adjacency matrix.

On the other side, a second table establishes links based on section 2.1 (from IDS1). This table allows the collection of more information on the previous adjacency matrix. The goal is to link both to obtain a clearer view of the data-set and the influencing factors.

Knowing that, an association between certain drug targets and COVID-19 is obtained. The degree of proximity between them can thus be defined by means of various centrality analyses. Indeed, knowing the link between the expression of a target in a given disease and COVID-19 is not enough to confirm some assumptions. The use of networks makes it possible to give more weight to observations without arriving to causal relations.

Many zero values are observed when creating the network from the data-set. Most of the time, only one gene is correlated with the selected disease. These correlations with target values, when they are not zero, represent the edges of the graph. A centrality analysis permits the investigation of the disease characteristics.

Centrality measurement yields information on how important nodes are. It makes perfect sense in the scope of this study as the aim is to identify which diseases share close relationships with COVID-19.

Before studying the centrality of the produced graph, it is interesting to observe the distribution among nodes when taking into account all nodes, disease only and genes only. This is illustrated in **Figures 4.1, 4.2 and 4.3**.

It can be observed that the distribution is spread out among the edges. This comes from

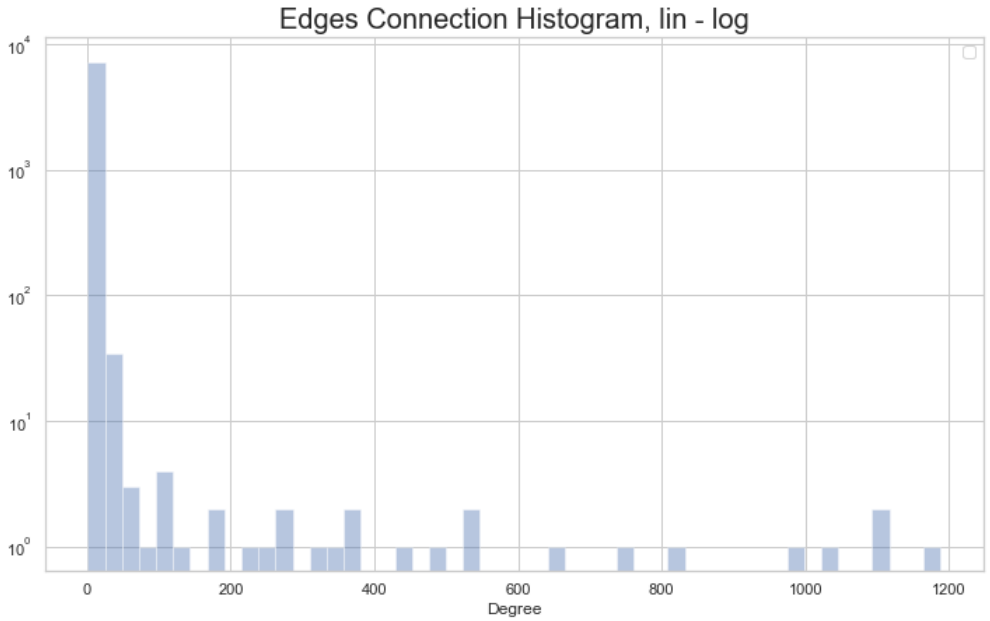


Figure 4.1: Edge histogram

the fact that many diseases do not have associations with targets in DS1. Indeed, most diseases only rely on one relationship other than COVID-19.

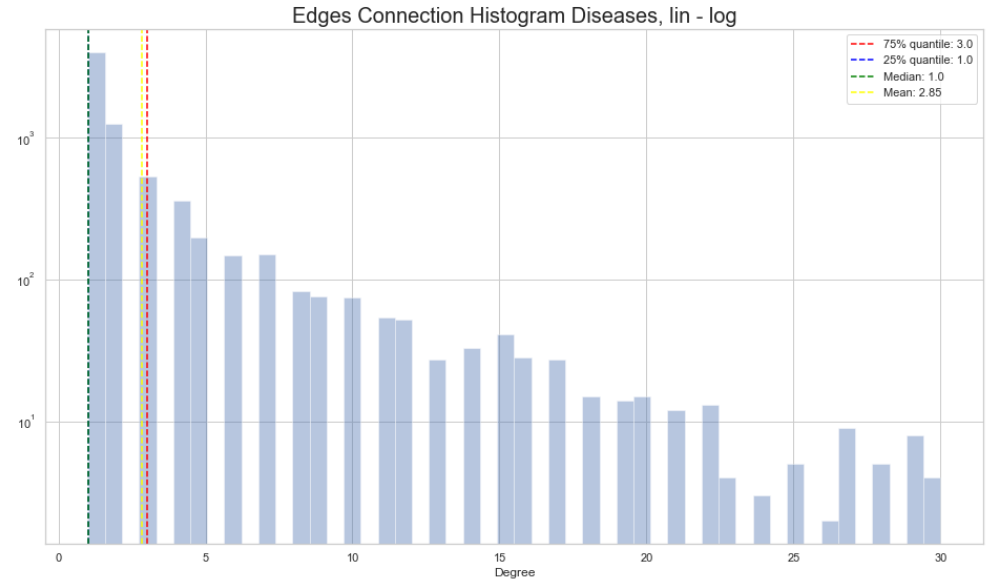


Figure 4.2: Edge histogram - diseases

Focusing on the results of the centrality analysis, rankings for disease edges and gene edges are listed in **Tables 4.1 and 4.2**. Several centrality criterions that express different characteristics among the edges were selected. Noticeably, it is possible to find

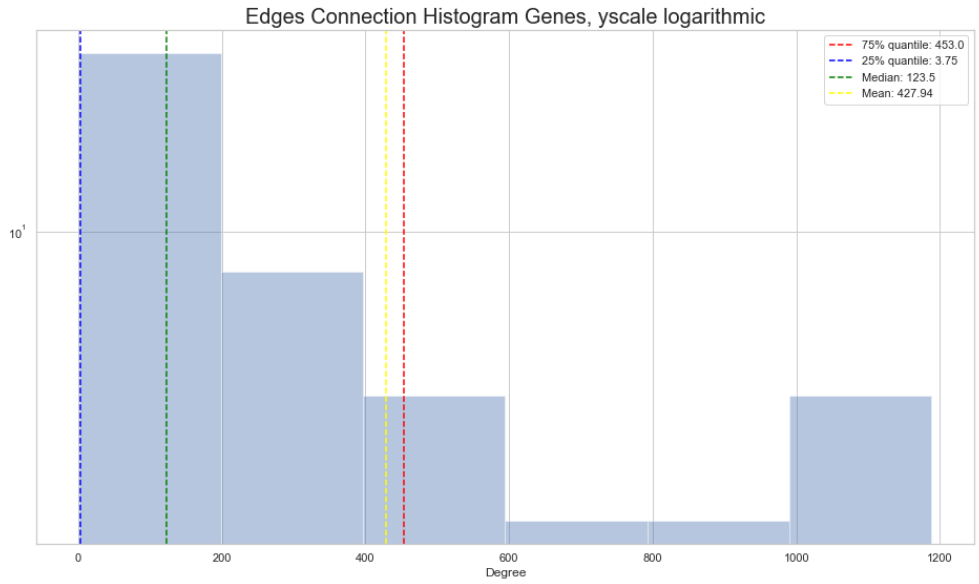


Figure 4.3: Edge histogram - genes

diseases or genes that appear in the first two rankings for different criteria. This means that they represents a special interest for this study.

	Degree Centrality	Closeness Centrality	Betweeness Centrality	Pagerank Centrality
1	genetic disorder	cell proliferation disorder	Paralysis	genetic disorder
2	genetic, familial or congenital disease	neoplasm	Short stature due to GHSR deficiency	genetic, familial or congenital disease
3	cell proliferation disorder	neoplastic disease or syndrome	Combined T and B cell immunodeficiency	cell proliferation disorder
4	neoplastic disease or syndrome	nervous system disease	immunodeficiency disease	neoplastic disease or syndrome
5	neoplasm	bone disease	vision disorder	neoplasm
6	cancer	brain disease	abnormality of brain morphology	cancer
7	respiratory or thoracic disease	central nervous system disease	visceral Leishmaniasis	respiratory or thoracic disease
8	thoracic disease	connective tissue disease	cytomegalovirus infection	thoracic disease
9	epithelial neoplasm	hematological measurement	Lymphangioleiomyomatosis	nervous system disease

Table 4.1: Centrality disease results

	Degree Centrality	Closeness Centrality	Betweeness Centrality	Pagerank Centrality
1	IL21R	LCAT	FDFT1	SCN9A
2	TCEA1	SCN9A	PDGFD	IL21R
3	SLC18A2	CALD1	PKIG	LCAT
4	LCAT	SLC18A2	FKBP10	SLC18A2
5	PDGFD	IL21R	LCAT	TCEA1
6	CALD1	PDGFD	SLC18A2	CALD1
7	DNAJC3	FDFT1	CDC20	PDGFD
8	TRIM13	DHRS11	CALD1	DNAJC3
9	CDC20	CDC20	RRAGD	TRIM13

Table 4.2: Centrality gene results

Finally, the general ranking for diseases and genes can be observed in **Tables 4.1 and 4.4**.

	bone disease	psychiatric disorder	biological process	protein measurement	diabetes mellitus
Overall Ranking	1	2	3	4	5
Number of connections	29	28	25	27	17
Degree Centrality	40	62	79	137	114
Closeness Centrality	6	18	30	25	118
Betweeness Centrality	49	17	38	20	28
Pagerank Centrality	37	60	74	108	107

Table 4.3: Overall disease ranking

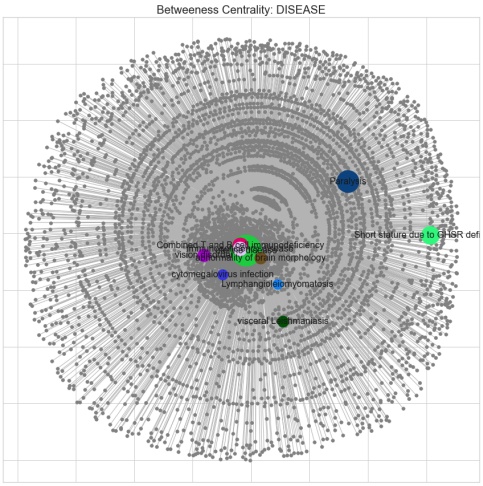
	LCAT	SCN9A	SLC18A2	PDGFD	COVID	CALD1	FDFT1	TCEA1	CDC20	FKBP10
Overall Ranking	1	2	3	4	5	6	7	8	9	10
Number of connections	1188	1103	1041	818	7217	1097	753	527	529	360
Degree Centrality	5	1	4	6	22	7	13	3	10	14
Closeness Centrality	2	3	5	7	1	4	8	11	10	15
Betweeness Centrality	6	12	7	3	1	9	2	19	8	5
Pagerank Centrality	4	2	5	8	1	7	13	6	12	14

Table 4.4: Overall gene ranking

Graph Representation

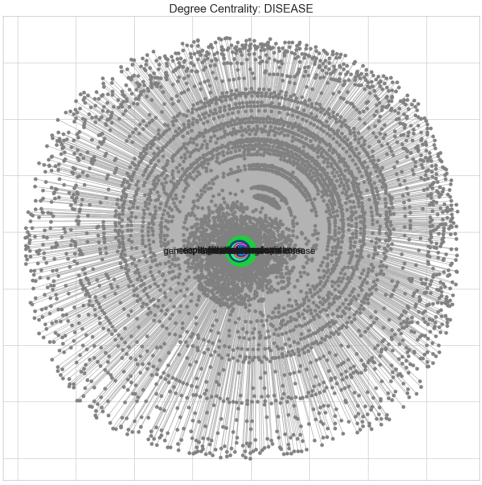
Based on the previous centrality results, a representation of the network can help visualize the previous statistics.

So far the network has not been modeled as a directed graph, although it could have been, but the specificity of the latter is that diseases or genes do not yet have links amongst themselves. This means that the representation is very "disease centered", which complicates the visualization of central diseases. This is illustrated in **Figrue 4.4**.

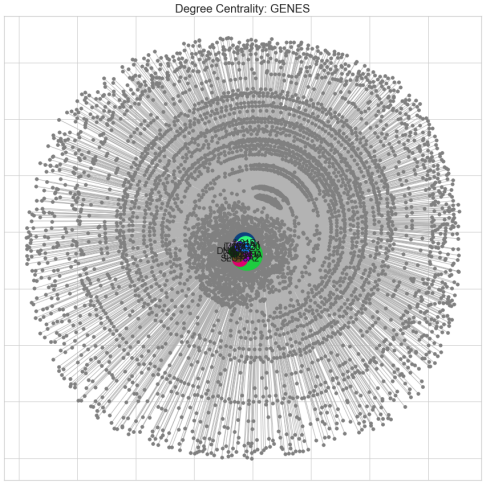


(a) Betweeness Centrality Disease

(b) Betweeness Centrality Genes



(c) Degree Centrality Disease



(d) Degree Centrality Genes

Figure 4.4: Network Centrality Representation

The next section will dive in the heart of the network to take a closer look at the central nodes.

Disease to Disease Network

As previously mentioned, IDS1 contains the relationship information between nodes. Two useful kinds of relationships for the already-built network are :

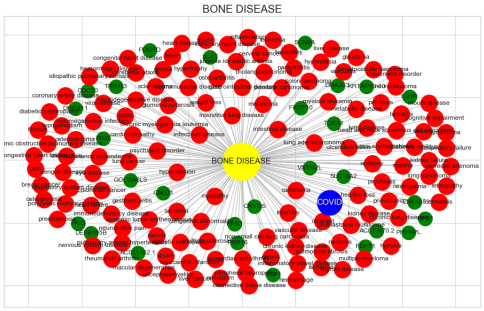
- **hasGeneticClue**: defines a link between two diseases due to genetic or behavioral similarities

- **isASpecic**: defines a link between a disease and its more general kind of disease, that could be interpreted as a cluster or disease category

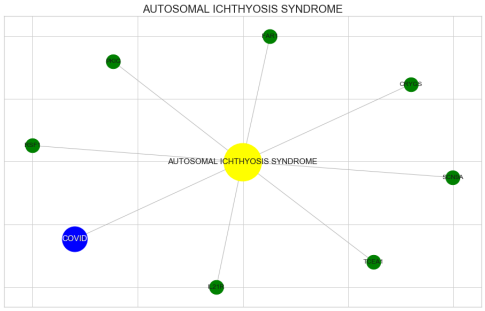
The first relationship enables adding complementary edges on the network since there was no link between diseases previously. Note that now, edges are no longer weighted.

By recalling the previous ranking system used above, a more precise look is taken at the central diseases in **Figure 4.5**.

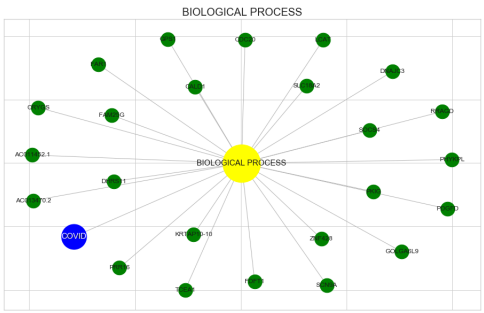
A comparison is made between, on one side, the diseases that scored best on the overall centrality ranking, and on the other side, the best scoring diseases obtained from the data set.



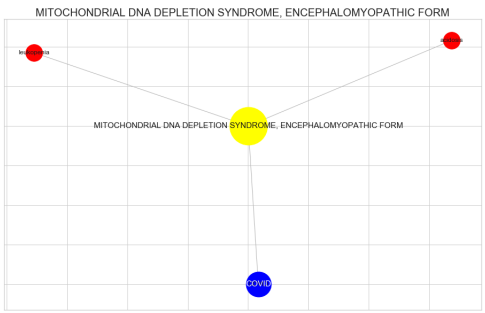
(a) Overall: 1.



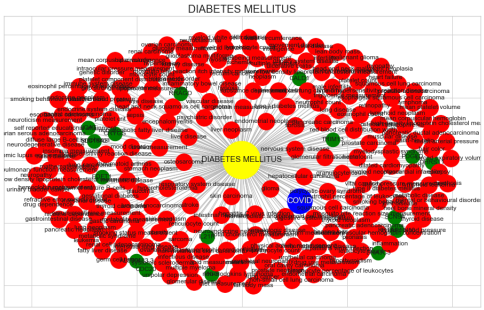
(b) Covid: 1.



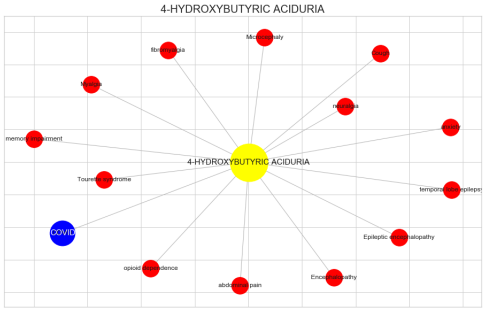
(c) Overall: 2.



(d) Covid: 2.



(e) Overall: 3.



(f) Covid: 3.

Figure 4.5: Ranking vs Covid score

Overall, the diseases scoring high on the overall scores on the left hand side have many more nodes than the diseases that score high on the COVID-19 score in the DS1 data-set. This shows an imbalance in the data-set as well as how complicated it is to find patterns since the most highly ranked COVID nodes offer little information of any kind.

Communities

Most of the diseases that have been seen so far are quite specific and do not usually ring a bell for common people. This is why the second data-set (IDS1) is used once again to help take out a clearer view of the network and the nodes influence. Mentioned above in

the second bullet point, **isASpecic** links a disease to an upper level disease, which can contain many diseases. It will now be referred to as *Community*.

By grouping the nodes in their respective community, the data will be clustered. In each community, the COVID-19 score of the belonging nodes is summed, giving an overall COVID-19 score to the community. The size of the community is also taken into account. Finally, the COVID-19 score is normalized by the size of the community to avoid a size bias.

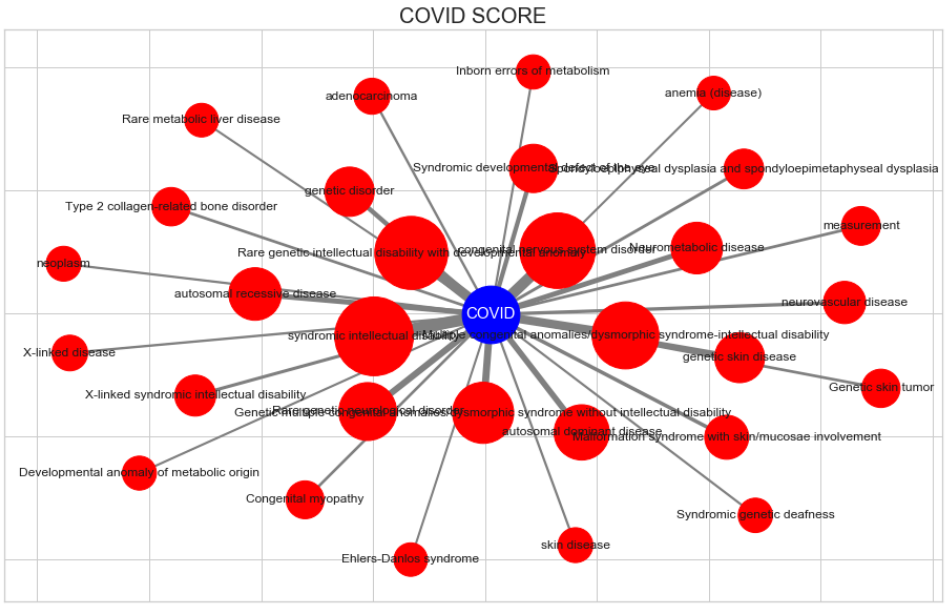


Figure 4.6: Largest Covid Score

Figures 4.6, 4.7 and 4.8 show the communities by size and their associated score to COVID-19 by the size of the link. It is still hard to find a recurring pattern in the 3 graphs or recurring community.

Overall, giving a certain importance to certain parameters compared to others will easily change the perceived influence of either the genes or the diseases on COVID-19. It is therefore complicated to find an accurate model of COVID-19 scoring parameters.

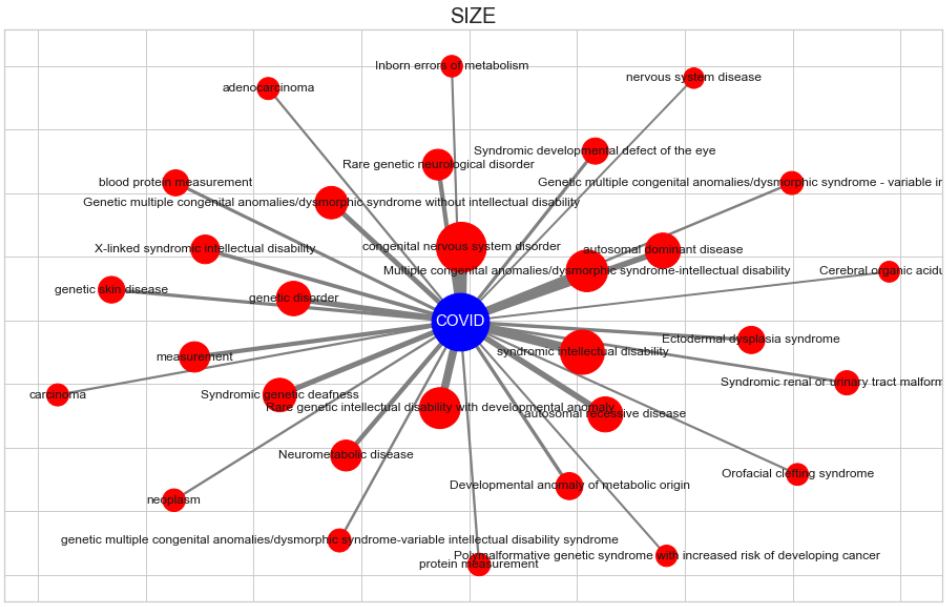


Figure 4.7: Largest Size

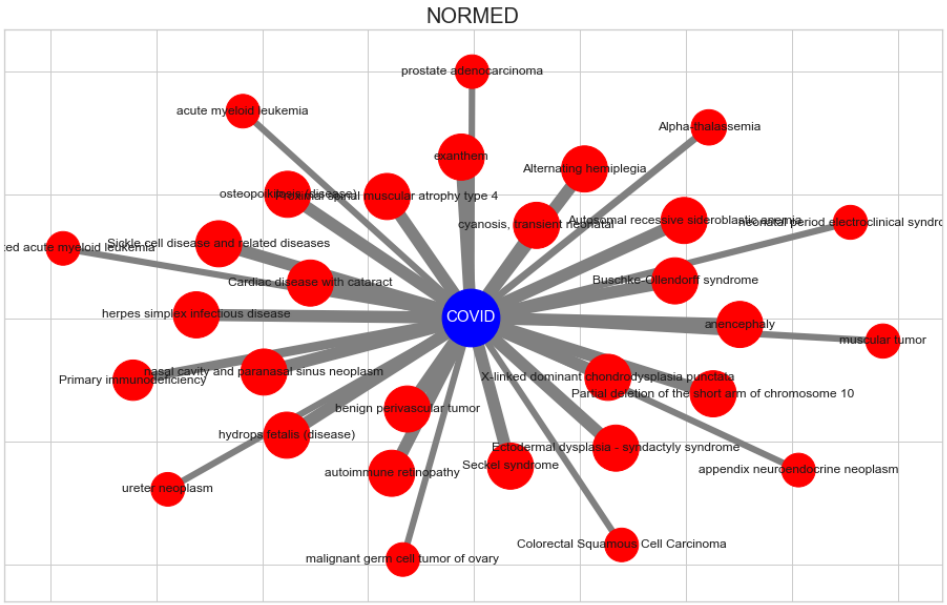


Figure 4.8: Largest Covid Score normed

4.1.2 Random forest regression

Grid search

A feature importance analysis from the random forest regression model shows consistently that target `ENSG00000213398` (lecithin-cholesterol acyltransferase) has an importance level of about 12%. The next 18 most important features range from 2 to 8%, as illustrated in **Figure 4.9**.

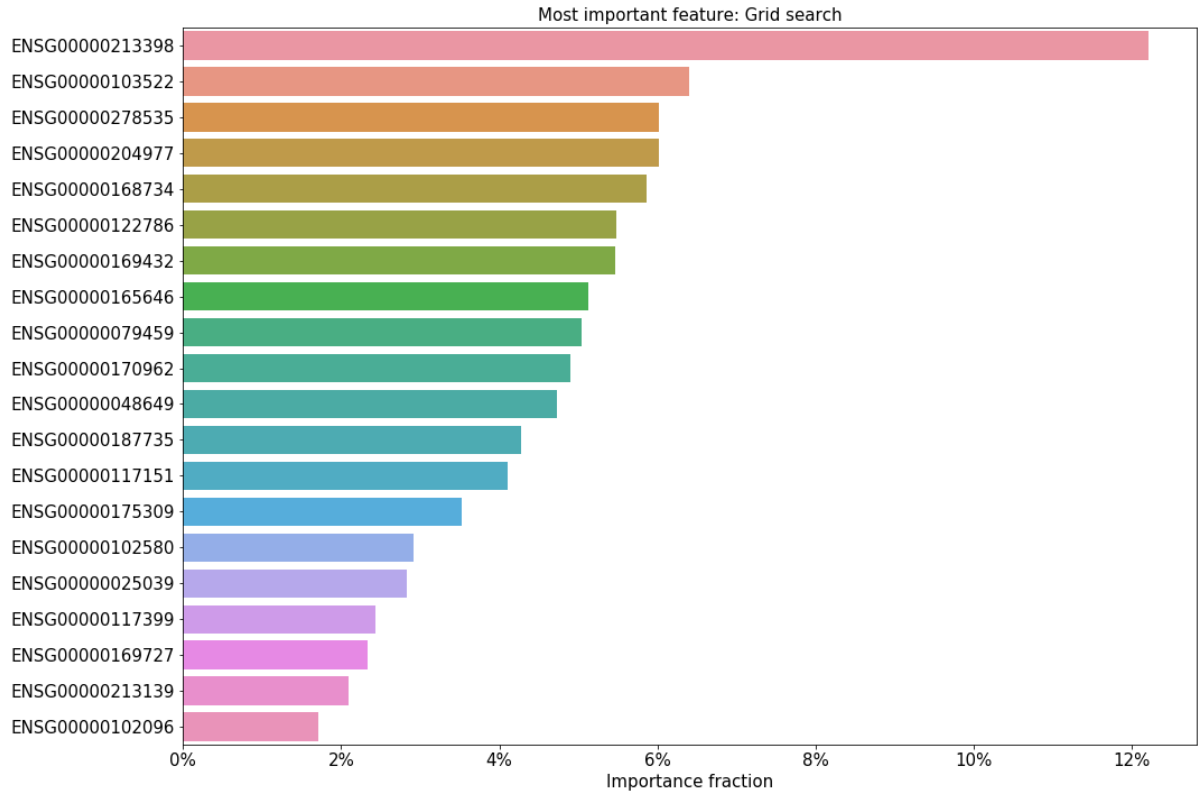


Figure 4.9: Grid search feature importance ranking

Only 22 (between 19 and 22 depending on model runs) features out of the 100 available capture over 95% of the importance. This is illustrated in **Figure 4.10**. Hence, running the model with only these 22 features barely increases the mean squared error of the result (less than 10^{-5}) from 0.00371%, though it decreases the r^2 by 35%, from 0.0021145420781890634 to 0.0013698372767038114. These values are very close to 0 however, which indicates that the model is barely better at predicting the output than if it was giving the expected value of the output based on training data.

Random search

Much like for the Grid search, a feature importance analysis from the random forest regression model shows consistently that target `ENSG00000213398` (lecithin-cholesterol acyltransferase) has an importance level of about 14%. The next 18 most important features range from 2 to 8%, as illustrated in **Figure 4.11**.

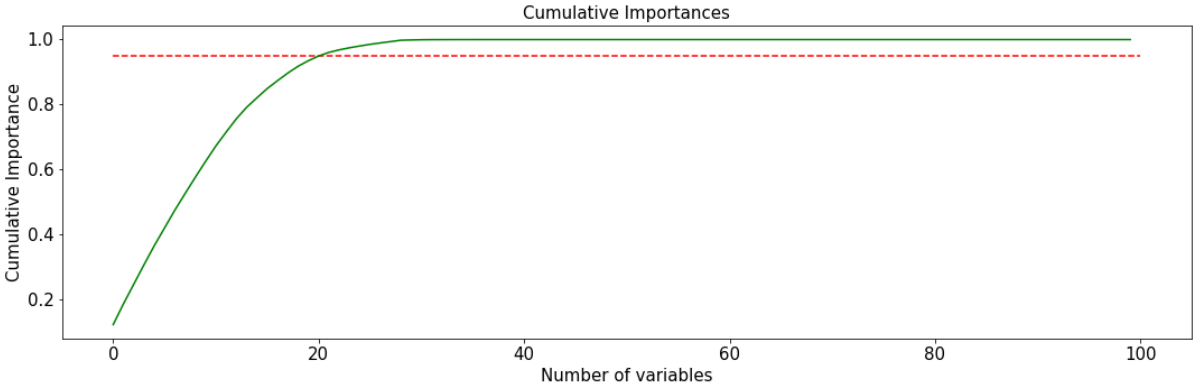


Figure 4.10: Grid search cumulative importance of features with a 95% threshold

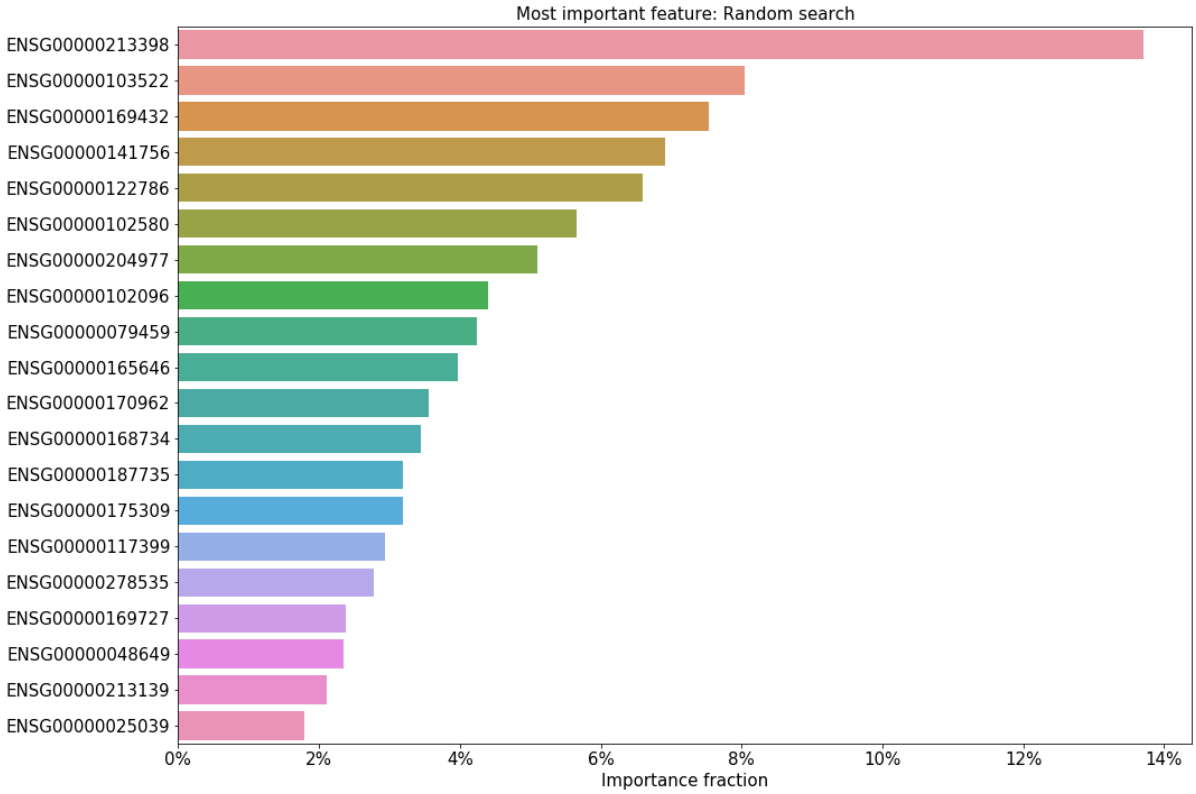


Figure 4.11: Random search feature importance ranking

Only 21 (between 19 and 22 depending on model runs) features out of the 100 available capture over 95% of the importance. This is illustrated in **Figure 4.12**. Hence, running the model with only these 21 features barely increases the mean squared error of the result (less than 10^{-5}) from 0.00375%, though it decreases the r^2 by 47%, from 0.0009199693393540098 to 0.0004879729800447707 4. These values are very close to 0 however, which indicates that the model is barely better at predicting the output than if it was giving the expected value of the output based on training data.

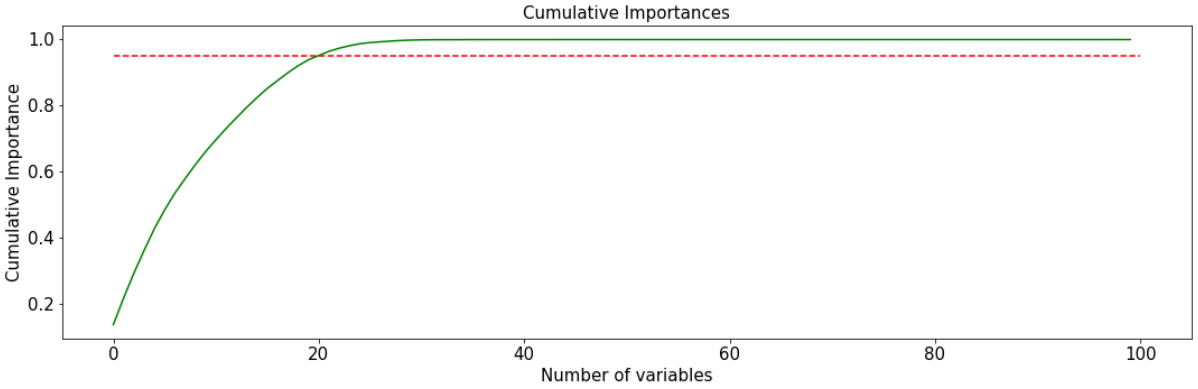


Figure 4.12: Random search cumulative importance of features with a 95% threshold

4.2 Benefits for the research on the virus

The network analysis was able to rank diseases and genes related to COVID-19 based on centrality analysis. The nearest diseases are bone disease, psychiatric disorder, biological process, protein measurement and diabetes mellitus as can be observed in **Table 4.3**.

The random forest regression has achieved the same results for drug targets, listed in **Figures 4.9 and 4.11**.

Identifying links can help highlight paths that may help reduce the costs of drug development. By highlighting targets associated to COVID-19, it is possible to focus research on these specific targets and thus reduce the field of investigation or time required for drug development. Tests will also be able to target these risk-factors specifically instead of a larger panel, thus reducing their cost. As a result, countries and individuals will be able to buy more tests, at a cheaper price, hence enabling deeper testing campaigns and reducing the risk of large-scale infections by identifying people at risk early on and taking measures to protect them and those around them. In this regard, a gene ranking has been established from the network analysis in **Table 4.4**. As the network analysis is, as of yet, more complete than the random forest regression model, it should be taken as a reference. It is interesting to note however that results concord between the network analysis and the random forest regression model. In particular, `LCAT`, also called `ENSG00000213398` is ranked first in both models.

4.3 Limitations and improvements

4.3.1 General

The cornerstone hypothesis used in this study is that genetics and genomics play a role in the impact on COVID-19 [9]. This has not yet been verified and is therefore a mandatory check before using any results produced by this model.

Moreover, the association score between diseases and COVID-19 was built on a partial data-set. Hence, association scores might vary with the update of the underlying (Kaggle) model. It should be run on all CORD-19 papers to be improved. One could hope

that it would therefore also increase the number of diseases with a non-zero association score to COVID-19 and hence the amount of data available to create the data-set used in this study. Indeed, as shown in **Figure 2.2**, there are few diseases with high association scores. Therefore, it can be expected that the currently produced model is not adequate for high association scores, which has yet to be verified.

4.3.2 Network analysis

The main limitation of the analysis is in the amount of data used. Only 4'000 diseases and 100 targets are used. The impact relies on the results that have been produced In the IDS2 data-set. Respiratory diseases and illnesses generally affecting overweight people were mainly expected. Coincidentally, diabetes appears at the fifth place in the ranking.

4.3.3 Random forest regressor

A major limitation of the produced random forest regressor model is that it is built on a small subset of identified targets. Hence, the most important targets may not be included in the data-set. This is stressed by the low r^2. Indeed, its low value demonstrates the low predictive power of the model. To improve the accuracy and the validity of the model, it should therefore be run on the whole panel of available targets. Further hyperparameter tuning may also increase the predictive capacity of the model.

Conclusion

This study is based on multiple hypotheses, first and foremost of which is that genetics and genomics play a role in the impact on COVID-19. Moreover, the present study was conducted on the basis of partial data-sets. Nevertheless, models have been run and have produced conclusive results.

Through network analytics, this study has identified the diseases and drug targets closest to COVID-19. As a result, if hypotheses are proven valid by other studies, researchers will be able to reduce the time spent on test and vaccine research, thus reducing their cost and enabling countries and individuals to purchase more tests and vaccines, at a lower price. The identification of principal target risk-factors will also enable the development of tests to define whether individuals are at risk for COVID-19. This will enable governments in turn to make proportionate decisions to protect their population by mitigating risks of infection and balancing measures taken with economic concerns.

Multiple improvements may yet be made to the proposed model. Among others, the underlying model may be run anew with many more papers to improve association scoring between diseases and COVID-19. The data-set created in this study may also be expanded to include the full range of targets from the Open Targets Platform. Finally, further hyperparameter tuning may be executed on both models to improve accuracy and decrease computational requirements.

Overall, this paper has provided encouraging results which, if proven true by further testing and improved underlying data-sets, will have a large and global economic impact.

Bibliography

[1] Coronavirus : la France lance une grande campagne de dépistage.

[2] Coronavirus (COVID-19) Testing - Statistics and Research.

[3] Coronavirus Update (Live): 3,834,038 Cases and 265,238 Deaths from COVID-19 Virus Pandemic - Worldometer.

[4] Covid-19 open research dataset challenge(CORD-19).

[5] Mondo Disease Ontology.

[6] Q&A on coronaviruses (COVID-19).

[7] World Population Clock: 7.8 Billion People (2020) - Worldometer.

[8] Joan Ballester, Jean-Marie Robine, François R. Herrmann, and Xavier Rodó. Effect of the Great Recession on regional mortality trends in Europe. *Nature Communications*, 10, February 2019.

[9] Christophe Guéret. Covid-19 risk factor predictor, April 2020.

[10] William A. Haseltine. Tests For COVID-19 Are Expensive, But They Don't Have To Be.

[11] IMFBlog. The great lockdown: worst economic downturn since the great depression.

[12] Kaisha Langton. Lockdown: Which countries are in lockdown? How many people?, May 2020.

[13] Dan Mangan. Coronavirus: New plan would test 30 million per week and cost up to $100 billion, but 'we've got to do it', April 2020.

[14] Federal Statistical Office. Population.

[15] Wikipedia contributors. Quantitative trait locus — Wikipedia, the free encyclopedia, 2020. [Online; accessed 9-May-2020].

[16] Wikipédia. Kaggle — wikipédia, l'encyclopédie libre, 2020. [En ligne; Page disponible le 19-janvier-2020].