Maspredicciones by Masbaratillas

Clinical-microbiological characterisation of SARS-CoV-2 infection in the paediatric age

Cleaning the Dataset

- The dataset contains many fields that are not relevant to us: IDs, Type of Interview, Interviewer Initials
- There are also many data that we cannot treat in this event: Postal Code, School Name and Position, Date of the symptoms, Sports Activities
- There are fields that contain text that would require text mining approaches
- Other columns have data related with PCR and Antigens test, that are only filled if the subject is positive in COVID-19

- We have deleted these fields

Cleaning the Dataset

- Many fields depend on others. If there is no headache, then headache_first is not filled → Fill missing data as no symptoms
- Some values have arbitrary numbers: → Normalize the data to [0,1]. 1 means yes
    - Yes represented as 1
    - No represented as 2
- Other cases the value is not known. Those values are not accepted by our models → Change value to 0.5
- Those cases that are suspected of COVID but are not confirmed → Remove the observations

Other possible data treatments

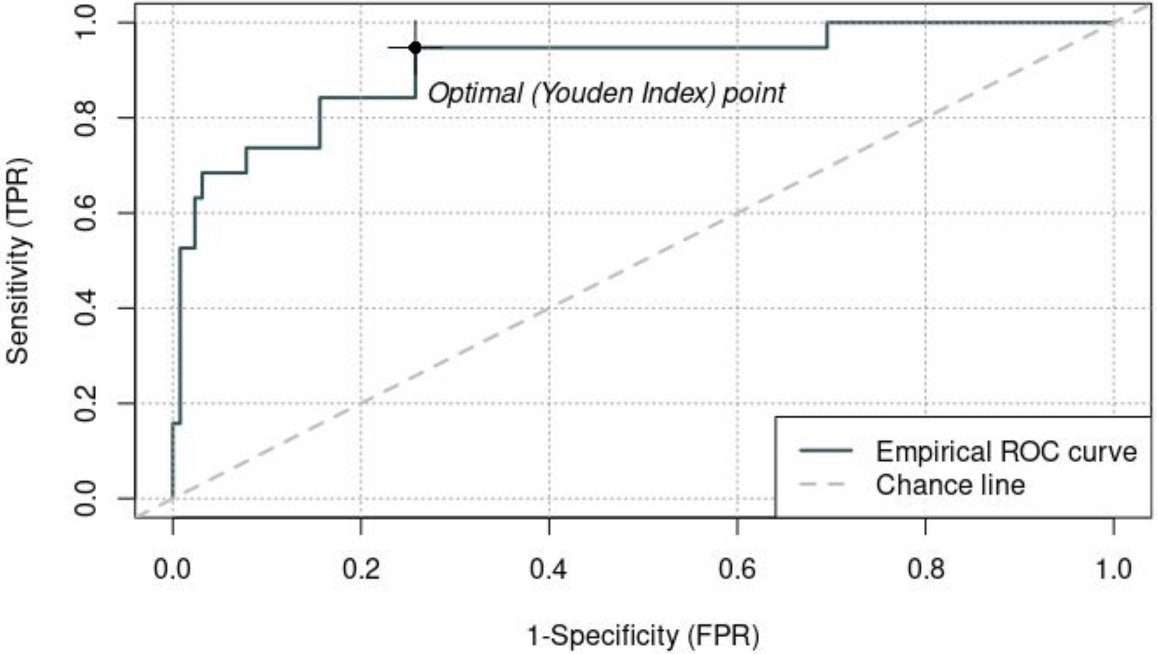
- Instead of 0.5 we could use the data distribution of each variable. For example we know that 60% of childs have no smokers at home
- We could perform data mining on the comments and observations field to create a dictionary of words
- We could add the Accumulated Incidence of each postal code to each observation

Linear Regression

- With the 90% of the cleaned data we have trained a Linear Regression Model in R
- This models tells us that what are the variables that are more relevant:
  - Environment cases: positive or symptoms at home
  - Modification in smell perception
  - Presence of fever
  - Headaches and shortness of breath
- Surprisingly the models tells us that some indicators are not good:
  - Presence of cough
  - Fatiga
  - Nasal congestion

Linear Regression

- The area under the ROC curve is 0.909

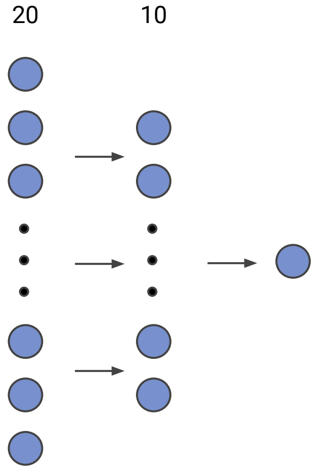


Bayesian Linear Regression Model

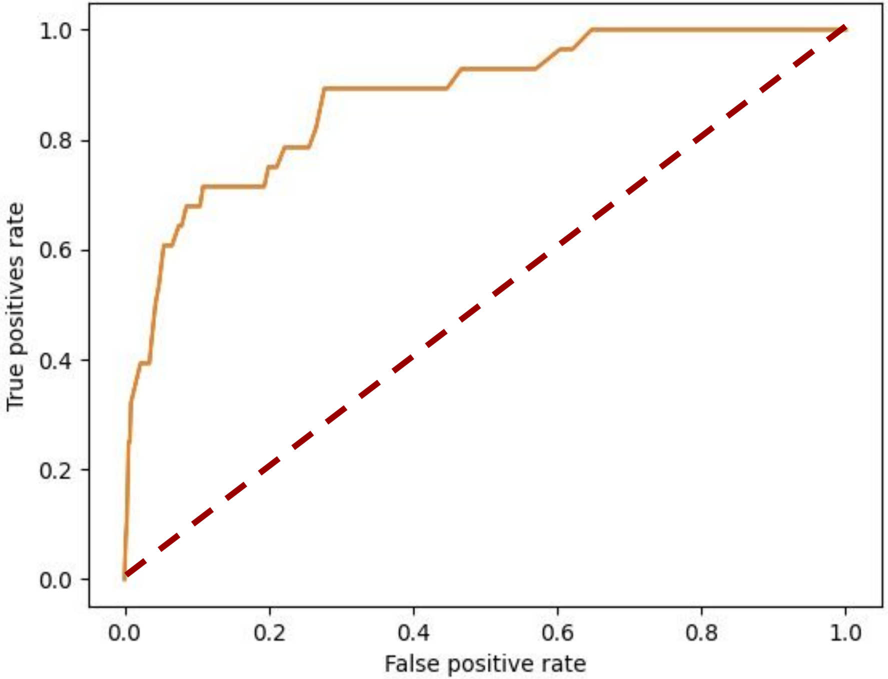
- We use Bayesian Regression model to predict if a case is COVID or not
- The bayesian network uses the probability of each variable and the relations between them to predict a final result
- We use a bayesian linear regression from sklearn library

Neural Networks

- Keras library
- Train set: 120 positives, 120 negatives
- Validation set: 30 positives, 30 negatives
- Test set: the rest
- Softplus activation for hidden layers
- Sigmoid activation for output layer
- Binary crossentropy error
- We tried different configurations



Neural Networks: ROC curve

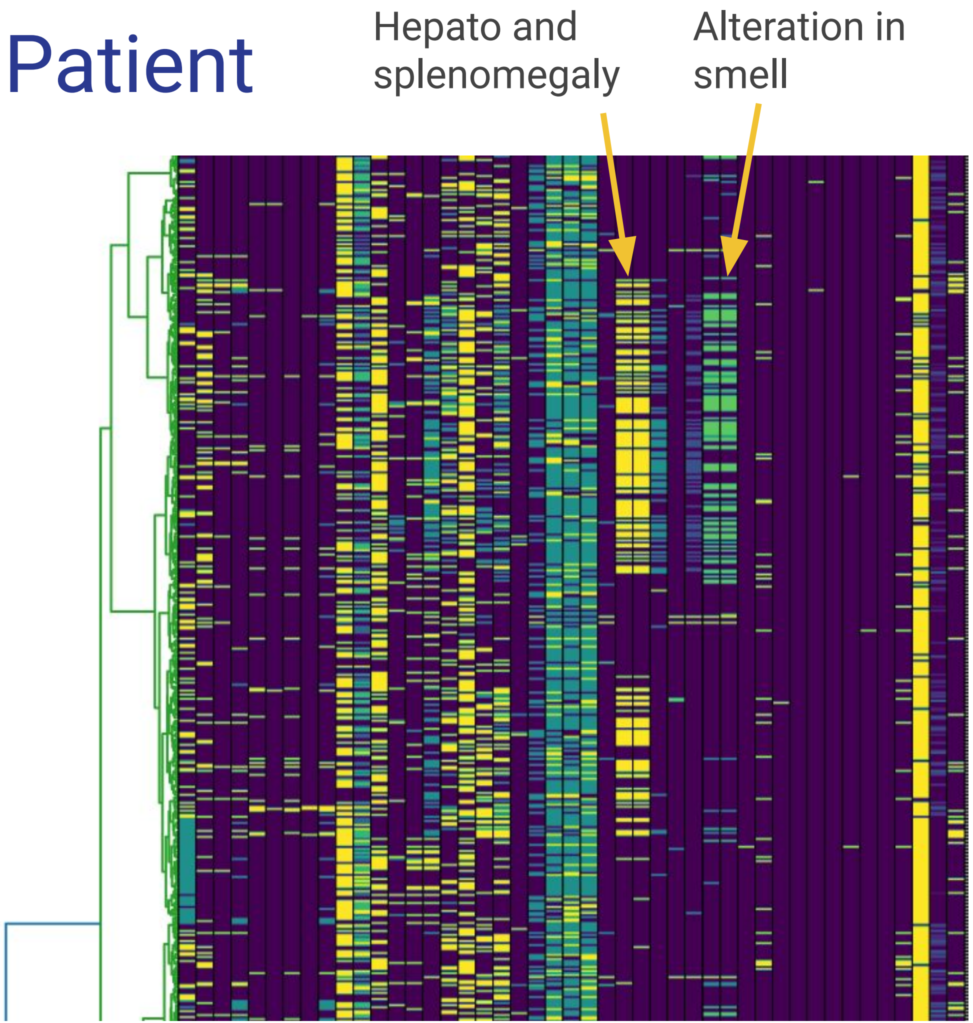


The COVID-19 Positive Patient

- We have performed a K-means clustering and hierarchical clustering on the data
- We have found that
  - 50% have a COVID-19 positive at home or with symptoms
  - 60% have fever and mostly under 39ºC
  - 20% have alterations in the smell or taste
  - 40% have cough
  - 50% have diarrhea or vomits
- The clustering approach allows us to create groups of positives that are similar
- And allow us to visualize the data

The COVID-19 Positive Patient

- Here we can see part of the visualization of the clustering
- We can see codified in color the values of the patients
- We can see that many patients with Hepatomegaly problems have alterations with smell



Interface for Our Models

- We use Django connecting python with html
- The user has to fill a form.
- Then in the screen will show up several results
  - Index of the bayesian regression
  - Index of the neuronal network
  - Clustering

Thank You

Isaac Sanchez
Rubén Langarita
Sergio Langarita
Víctor Soria