Pedretti Beatrice, 943805

Classification assignment

The script implemented for this project follows three main parts:

- Feature analysis
- Feature engineering
- Modeling

1. Preliminary analysis: loading and checking data

Firstly, I imported some useful tools and loaded the .csv file in the Jupiter Notebook[1].
Thanks to the functions `data.info()` and `data.describe()` we are able to observe that:

- There are 13741 observations in the dataset
- Out of 11 features, 8 are categorical and 3 are numerical

2. EDA

Outliers

After loading the data, since outliers can have a dramatic effect on the prediction, I implemented Tukey's method for outllier detection (Tukey JW., 1977) in a function that takes a dataframe and returns a list of the indexes corresponding to the observations containing more than n outliers.
As for the selection of the parameter n, since setting n=0 resulted in more than 2800 outliers, and those outliers were in the range of reasonable values for the corresponding features, in order to not lose in generality I decided to consider as outliers the rows that have at least two outlied numerical values, and thus none.

Missing data

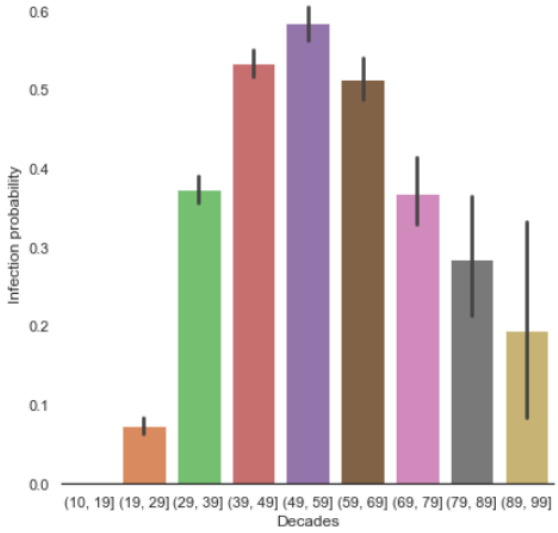
Afterwards, I checked for missing values with the function `data.isnull().sum()`, which didn't return any Nan values. Although, on further inspection of the features, I found out that missing values had been filled with the symbol '?' in the columns 'wing', 'country', and 'occupation'. By inspecting the rows containing this character I was able to infer how the symbol '?' has the same cardinality (711) and can be found in the same rows fow both the features 'wing' and 'occupation'. This lead me to believe that the symbol might probably code corrupted data referring to the patients and medical staff in the East wing ('E'), which is the only missing wing in the column. As for the country, with the present knowledge it is impossible to know for certainty why the data is missing.

[1] See 943805.ipynb

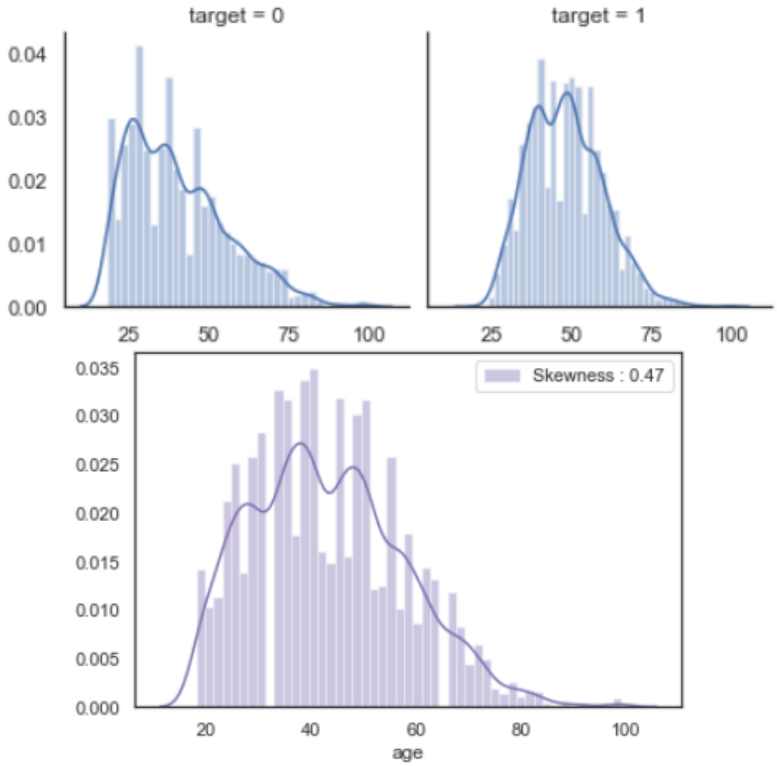
Numerical variables: 'age'

From the correlation matrix between the numerical variables it appears that the only variable with a significant correlation to the target would be 'speciality_years'. Anyway, it doesn't mean that the other features are not useful.

After binning the 'age' feature for easier legibility, it appears that the people between 40 and 70 years of age tend to be the majority of the analyzed cases.



From further inspection, we see that younger people have a higher probability to not be infected, while for the infected patients the distribution tends to be centered around the 40s-50s decades. The age feature also has a relevant skewness.



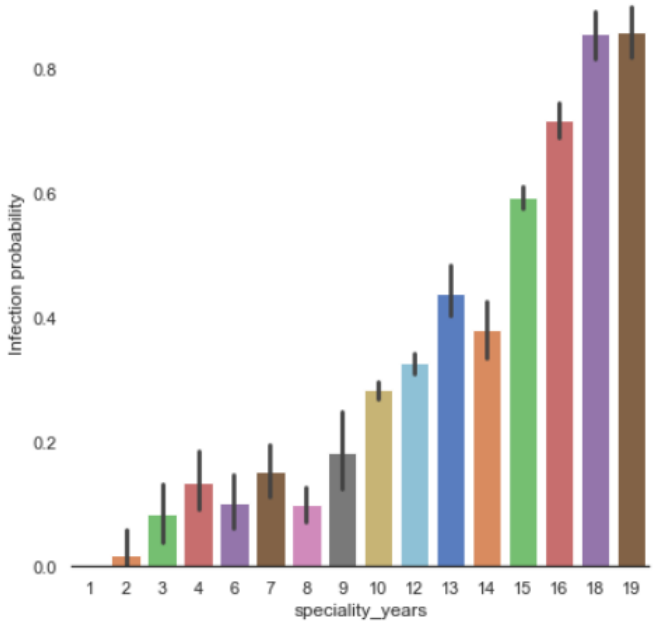
Numerical variables: 'speciality_years'

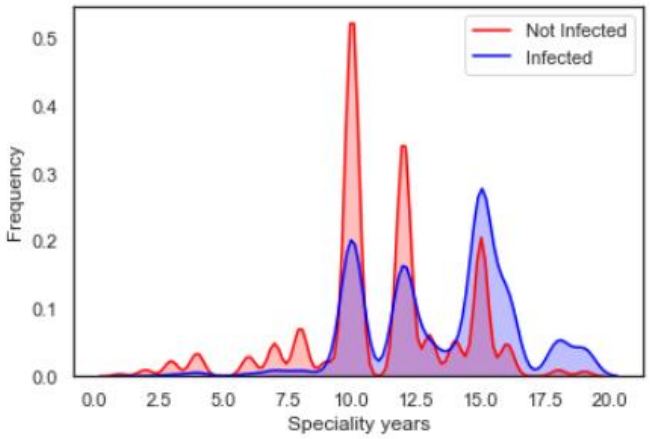
After computing the value counts for the features 'speciality' and 'speciality_years', we are able to fill the folowing table:

Speciality	**Speciality_years**	**Number of istances**
Psychiatry	10	4163
Emergency	12	2891
Pulmonary	15	2645
Pediatrics	16	936
Ophtalmology	13	571
Gynecology/Urology	14	467
Orthopedic	8	439
Neurosciences	18	345
Nephrology	7	300
Endocrinology	19	259
Oncology	4	215
Radiology	6	177
Gastroenterology	9	137
Anaesthesiology	3	129
Cardiology	2	51
Aesthetic	1	16

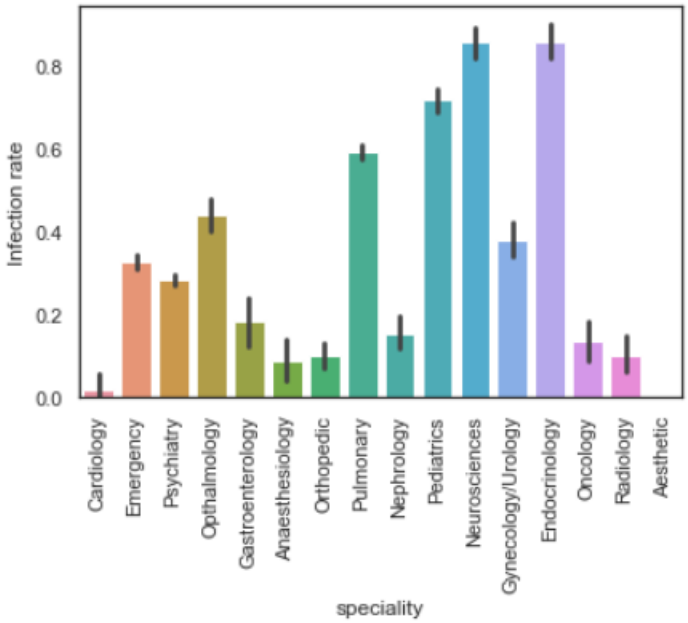
The feature 'speciality_years' reveals some interesting information: by a simple inspection it is evident how the people in the older specialities tend to have an almost certain probability to be infected, while the newer the speciality, the less the infection rate.

It is also apparent that the vast majority of instances corresponds to the Psychiatry ward.



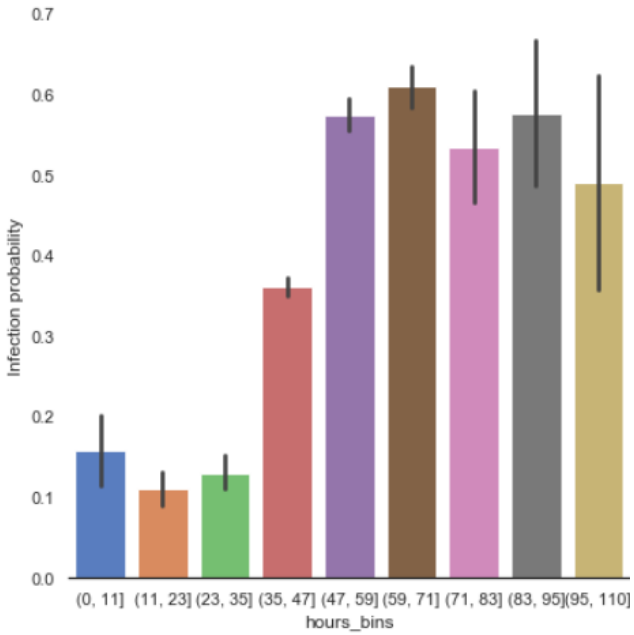


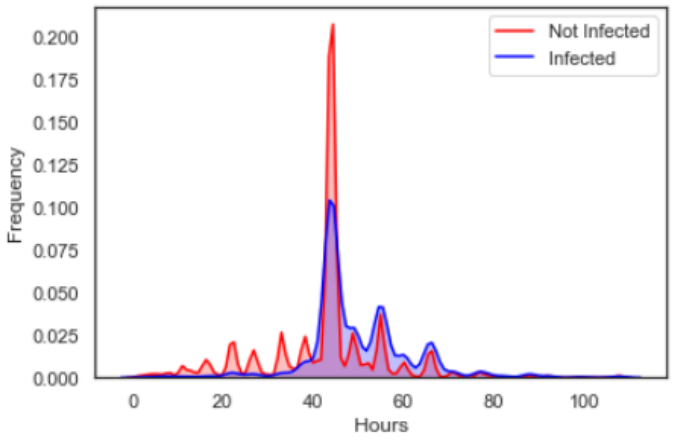
If we look at the speciality feature distribution it is indeed evident that the older wards such as Neurosciences, Endocrinology and Pediatrics have a higher probability of infection, while the newest have a very low infection rate.



Numerical variables: 'hours'

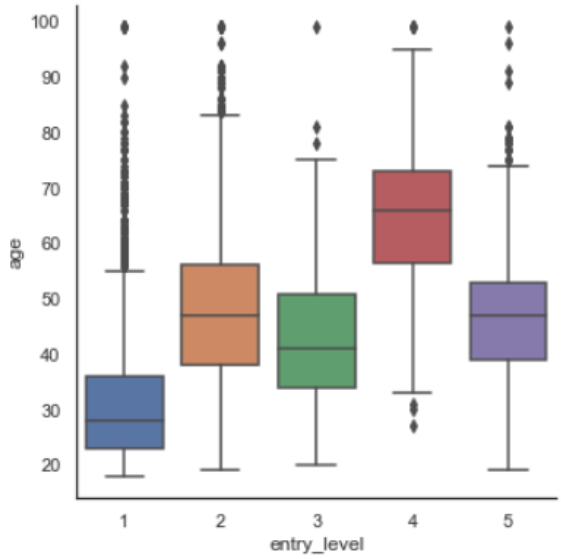
The infection seems to have started around the 109 hours before the creation of the dataset. The hours since the infection have an important peak towards the value 44, which is when the majority of the observations was collected. The probability of being infected also decreases for the more recent observations, meaning that the virus is probably being contained.





Categorical variables

- Speciality: the speciality feature has already been discussed with respect to the speciality_years numerical variable
- Wing: the ilness probability in each wing seems equally distributed except for the NE wing. It has already been discussed how the '?' value probably refers to the E wing, which is missing.
- Country: although there are some countries whose citizens have a very high infection rate, the vast majority of the subjects seems to come from the US, with 12375 observations.
- Occupation: a lot of dietitians have been infected, and also the majority of the gift shop staff, although the observations pertaining to the gift shop are very few.
- Intervention: it seems that the majority of infected patients have been treated with either a surgical or therapeutic intervention.
- Entry level: since the entry level feature is an ordinal categorical variable, I decided to label it by seriousness in a growing sequence.



The patiens presenting the more serious conditions appear to be older.

- Ethnicity: highest rates of infection between 'Asian-Pac-Islander' and 'White' ethnicities.
- Sex: roughly 50% of the men have been infected, while for the women this values falls at around 20%.

Reference graphs are in the 943805.ipynb file.

Feature engineering

Firstly, based on the earlier observations about frequency of infection in the single variables, I decided to label the majority of features in order to perform a better encoding of the categorical variables. I also decided to keep the binned numerical variables and perform a One Hot Encoding on the single bins since I observed (by trial-and-error of many combinations) that this is the combination of variables that gives a better performance on the prediction model.

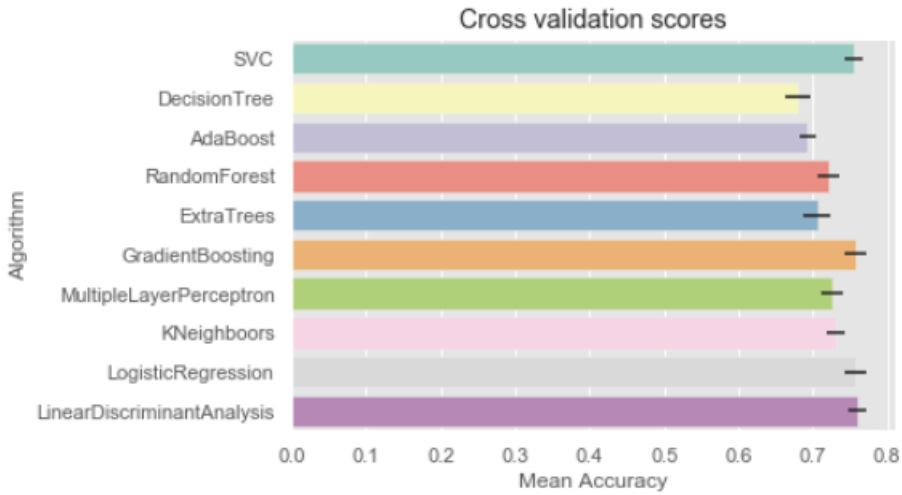
As said before, I mapped the entry level feature to a numerical ascending scale, assigning to the mildest condition the value 1, and to the most serious the value 5.

Then I mapped all countries into a binary distribution since the 90% of the patients/staff come from the United States. I also mapped the '?' value into the new value 'Other-country'.

As for occupations, I decided to map 'Delivery' and 'Cafeteria' into the same category because of the low number of instances and low rate of infection in both classes. I considered putting 'Gift shop' in the same category, but didn't because of the high rate of infection.

Modeling

After some feature engineering, I divided the dataset thanks to the `train_test_split` function, and then I implemented a comparison of the most popular modeling methods that, as output, gives a visual score for each algorithm. In the end, what I got was this:



I decided to try the Logistic Regression, Gradient Boosting, and the Support Vector Machines Classifiers. For each model, I then divided the original dataset in two random subsets, of which I used only the first to train the model, as an additional assesment. After applying the same feature engineering as before, I trained the model on the first subset and used it to

predict the second one, and I found out that the Gradient Boosting, although it seemed better performing (with a f1 score of roughly 0.81), it tended to overfit the training dataset.

Eventually the overall best performing model, although computationally costly, proves to be the SVC, that did not seem to overfit the model even in light of the second evaluation I decided to try, with a f1 score = 0.7678467788740568 and the subsequent confusion matrix and ROC curve, and a very high AUC (89%).