Fostering Machine Learning Tasks for the Data Science and Engineering Club UAB

Jaume Gerard Estany González

Resumen– Hay una creciente necesidad para las empresas de contratar a profesionales especializados en el campo de las ciencias de datos. Por eso, animar a los estudiantes a utilizar plataformas con la intención de enseñar sobre él es una prioridad para los docentes de estos campos. El objetivo de este proyecto es el estudio de diversos problemas de Machine Learning y la composición de blogs que sirvan para explicar como afrontar las situaciones que suelen darse a la hora de analizar un problema. Además, se participará en una competición pública y se examinarán las estrategias con los mejores resultados.

Palabras Clave– Aprendizaje Computacional, Aprendizaje Profundo, Análisis de Datos, Machine Learning, Deep Learning, Data Analysis

Abstract– There is a raising need for companies to hire professionals specialised in the data science field. Therefore, to encourage students to use platforms with the purpose of teaching about it is a priority for the teachers of these fields. The objective for this project is the study of different Machine Learning problems and the composition of blogs that explain how to deal with situations that often happen when analysing a problem. In top of that, there will be a participation to a public competition and the top-scorer strategies will be examined.

Keywords– Machine Learning, Deep Learning, Data Analysis

1 Introduction

Recently, Machine Learning has been becoming more and more important. This can be seen even by people without a technical studies on the subject, because of the impressive technologies that have been shown recently. Is that so, that companies have demonstrated a growing need for people with knowledge in these fields.

Despite that, these technologies are still very young and there aren't that many professionals that can cover all this demand. This is why the purpose of this project is to offer Machine Learning students a platform where they can learn how to solve common problems of these fields by example and encourage them to participate in online Machine Learning competitions.

The target platform is the DataUAB webpage [16]. It is a page where there are some blogs showing examples of machine learning problems. In order to make the platform more interesting, there will be a study on some Machine Learning problems and blogs explaining the work that has to be done in each one of them. In top of that, in order to encourage the students to participate in public Machine Learning competitions, there will be a participation in a public competition from Kaggle [12]. Kaggle is a webpage that allows data scientists to upload datasets and to publish kernels and open competitions about them.

The DataUAB blog works by now, but it still has room to improve. Because the target of this project is to add more content to the blog, in top of adding new blogs, some functionality improvements have been done.

2 Dataset Selection

There is a selection of datasets that has been made taking into account the overall properties of the datasets. This selection has been made trying to repeat the least possible and to explain different key concepts everytime.

1. Suicide Rates between 1985 and 2016 [10]: this dataset contains information about all the suicides that happened in different countries and years. Given all the combinations of country, year, gender, age and generation considered, it gives the number of inhabitants for each one and the number of suicides that hap-

• E-mail de contacte: jaume.estany.gonzalez@gmail.com
• Menció realitzada: Enginyeria de Computació
• Jordi Gonzalez (Ciències de la Computació)
• Curs 2018/19

pened in that part of the population. For this dataset, the model will try to guess the number of suicides that happened in a part of the population given all of their properties. This will be a regression problem.

2. Flight Delays and Cancellations of 2015 [11]: it is a registry of plane flights that happened during 2015 in the USA. The registry contains data like the origin airport, the destination airport and the airline of the flight. In top of that, it has registered if the flight was cancelled or not and it also contains detailed information about the delay of the flight. For this dataset, the model will try to guess whether a flight will delay or not given its properties. This will be a classification problem.

3. Malaria Cells [13]: this is an image dataset. It contains images of human cells. They are divided in two classes: the healthy ones and the malaria infected ones. Here, the model will try to guess whether a cell is infected with malaria or not. Precisely because it's an image dataset, this will be a Deep Learning classification problem.

3 Blog Posts

A blog post has been built for each one of this datasets, and all of them are composed by Python 3.6 [9] code and markdown language chunks. These blogs have been developed using Jupyter Notebook tools [15] and have been exported to HTML [17] and added to the DataUAB blog [16].

3.1 Suicide Rates between 1985 and 2016

For this problem, the dataset that has been used is the "Suicide Rates between 1985 and 2016" [10] dataset. This dataset considers different parts of the population and gives the number of inhabitants and the number of suicides for that set of people.

Our goal here is to guess the number of suicides that happened on that part of the population, and this will be achieved using a regression model.

3.1.1 Data Analysis

In the blog, the first thing that is explained how to take care of are the missing values. The blog emphasises in properly analysing the missing values across the columns to minimise the data loss. As explained, the missing values must be either deleted or modified. This is why carefully choosing the features for the model is important to avoid losing data unnecessarily.

The next step in the blog is to see the relations between all the features of our dataset. This is best done by plotting the Pearson's correlations between all our columns. The blog's image that shows all the correlations can be seen at the figure 1.

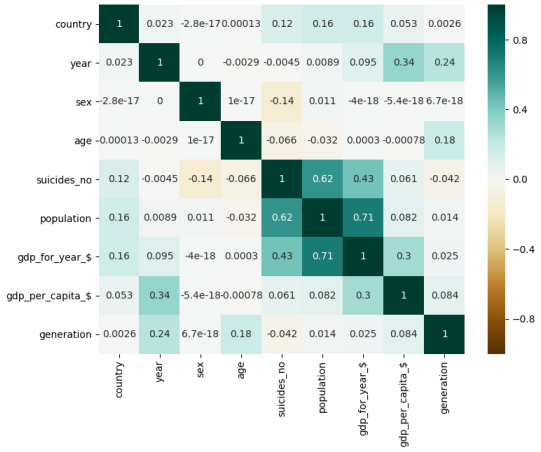


Fig. 1: Heatmap of the features correlation.

Some dependencies stand out immediately. There are some heavy dependencies between the features "population" and "number of suicides" and "population" and "GDP for year".

These two relations seem a consequence of the nature of the features. The number of suicides depends on the total population because, generally, the bigger a set of population is, the more probable it is for someone to commit suicide.

In addition, the GDP for year also heavily depends on the total population. This is because the GDP for year collects the value of all the goods of the population in that year. This generally means the more people there is, the more goods will be taken into account and therefore the bigger the sum will get.

3.1.2 Machine Learning

In this part, the model chosen for this problem is a polynomial regressor. A polynomial regressor is a regressor that, given a degree, finds a regression function that is a polynome of said degree that tries to fit the data.

Because the degrees for the regressor have to be chosen manually, the range considered is from degree 1 to degree 6. An evaluation of the R^2 and MSE metrics depending on the degree chosen can be seen at the figure 2.

Here the blog starts by explaining some concepts of normalisation. Normalisation is very important in the process of training a machine learning, because it removes the magnitude of the different features. It consists in scaling the data to remove the sense of magnitude.

In top of that, it also benefits the gradient descent process, because in the case of an extremely big or small magnitude, the gradient descent could take too long to finish or skip minimums completely.

There are a variety of methods to scale the data, but the blog post goes over two of them: the minmax and the standard.

The blog continues by explaining the regression metrics that have been used to evaluate the final model. Those would be the R^2 score, the Mean Squared Error and the Root Mean Squared Error.

Finally, the blog explains how to create and train the polynomial regressor given a specific degree.

3.1.3 Results

In this section of the blog, some of the final results of the trained model are shown. A sample of that can be seen at the figure 2. The figure shows the R^2 and MSE metrics for the training and validation set. This plot varies with the degree chosen for the polynomial regressor.

This plot enables us to choose a degree for our polynome by showing us the metrics of each of the models. Generally, when raising the degree of our model, both the scores for train and validation will raise.

However, this improvement has a limit. This limit comes when the model scores better with the training set (i.e. data that already has been exposed to), but scores worse with the validation set (i.e. data that has never been exposed to).

This is because of overfitting [19]: the model does good with data from the training set, but is far worse when predicting samples that it hasn't used during the training phase.

In order to avoid overfitting, we choose the model that has scored better with the validation set (i.e. degree 5).

Also, some properties of the model have been observed. The feature importances have been extracted from the model and can be seen here 3. The three most determining factors to raise or lower the estimation of suicides are the number of inhabitants, country and sex.

It stands out that the number of inhabitants is really important, because statistically, the more people there is, the more probable it is for a suicide to happen.

Also, males have suicide rates that are generally between 3 and 5 times as much as women's rates, depending on the country. This phenomena can be observed at the figure 4. This figure shows the frequency of the number of suicides by sex.

It can be observed that men have always superior frequency in the right part (high number of suicides), while women have a lot more frequency in the left part. The most left part of the figure represent 0 suicides, so the more people there is close to the left limit, the less suicides that part of population suffers.

In top of that, the country is also really influential, because of the suicide past record that some countries have. When talking of relative frequency, the three more significant are Lithuania, Sri Lanka and Russian Federation.

These three countries have shown heavy past records of suicides, which can be seen here [1] for Lithuania, here [20] for Sri Lanka and here [18] for Russia. The average of suicides by country every 100k people can be seen at the figure 5.

In Lithuania, apparently, there have been high unemployment rates in the country in the past years, and that has led to high suicide rates in the rural areas, specially among men [1].

In Sri Lanka, apparently, there were some high suicide rates in the past because of the easy access anyone had to pesticides. The pesticides used in the past were very harmful and resulted almost always fatal. However, some particular toxins were banned to use as pesticides from 1984 to 2011, which successfully stopped the growth of the rates.

In Russia, the suspicions revolve around the high alcohol consumption among the teenagers and the easy access they have to lethal methods. Also, Russia is one of the countries where the suicide rate difference between men and women is incredibly high.

Despite the people, sex and country being the most important features for the regression, the wealth is also very important in the analysis of the data. This can be seen through the features anual GDP and GDP per capita at the figure 6.

3.2 Flight Delays and Cancellations of 2015

For this problem, the dataset that has been used is the "Flight Delays and Cancellations of 2015" [11] dataset. This dataset contains data about numerous flights that happened in 2015 between airports in the United States.

Our goal here is to guess whether a flight will delay or not. For this problem a total delay of more than 30 minutes has been considered actually delayed. The rest of flights have been considered on time.

3.2.1 Data Analysis

In this blog, just as the previous one, the first thing that explains is how to take care of are the missing values.

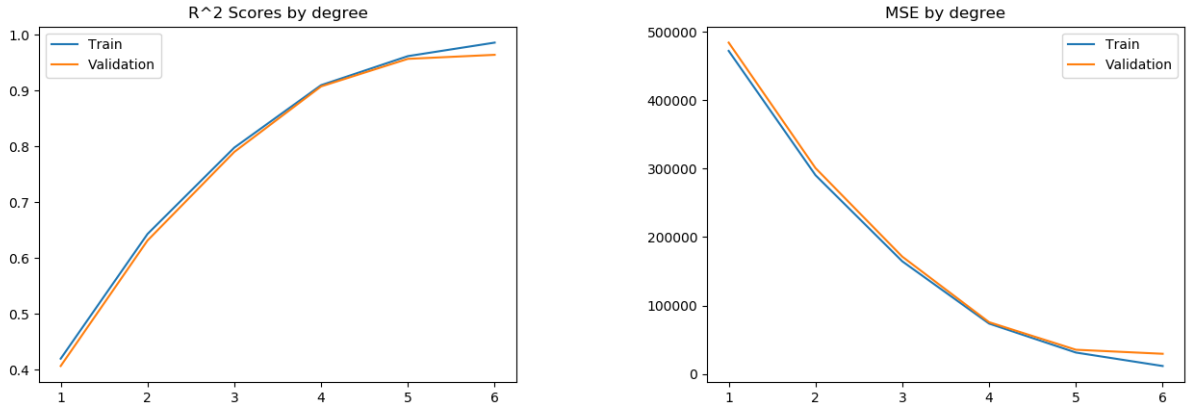
The next step in the blog is to see the relations between all the features of our dataset. This is best done by plotting the Pearson's correlations between all our columns. The blog's image that shows all the correlations can be seen at the figure 7.

There are two correlations that stand out in the heatmap. The first correlation is between the time scheduled for the plane departure and the time scheduled for the plane to arrive. This makes sense, because the arrival time will always depend on the departure time, specially for flights between airports of the United States.

The second correlation is between the origin airport and the destination airport. This makes sense, because the flights between some airports will be more common because of closeness and stops, for instance.

The blog then talks about the importance of checking the feature distribution in order to give meaning to the data. Because of that, some code has been included on how to plot the histograms for each feature. A sample can be seen at the figure 8.

In the distance histogram we can see that most of the flights are from close distances. It makes sense for the higher number of flights to be between close cities and not to be very long journeys.



(a) (b)

Fig. 2: R^2 and MSE metrics depending on the degree chosen for the polynomial regressor.

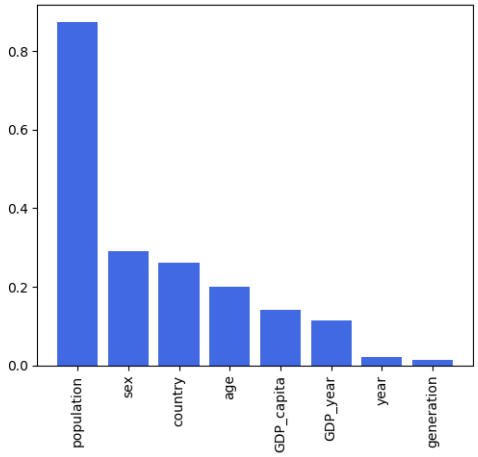


Fig. 3: Feature importance for the regression of the Suicide Rates dataset.

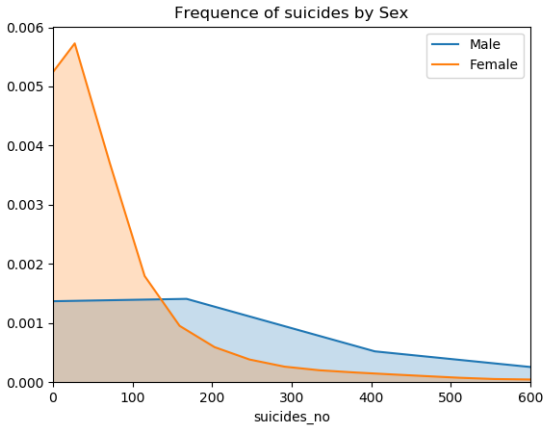


Fig. 4: Frequency of the number of suicides by sex.

Also, the scheduled arrival histogram shows that a very huge proportion of flights arrive from 8:00 to 22:00, having almost no arrivals on the rest of time.

3.2.2 Machine Learning

For this section, the model chosen is a logistic regressor. Despite being called a regressor, this model is a classifier. It's called a logistic regressor because it uses a regression function to determine the probability of the samples belonging to one class or another.

In this section of the blog, the first to be clarified is the metrics that will be used to evaluate the model. Because this is a classification problem, we will use the confusion matrix, along with the accuracy, precision and recall metrics. These metrics are only used in classification and are some of the more useful tools in order to evaluate classification models. These metrics will also be seen in other sections.

After that, the blog explains what the K-fold Cross Validation method is and what advantages does it have over other validation methods. Then, it includes how to train a logistic regressor using K-fold Cross Validation.

The method of K-fold Cross Validation consists in dividing the whole dataset in equally-sized parts and, one by one, train the model with all of them but one, and to test it with this remaining one. This is done once for every batch.

To this day, this is one of the most robust and efficient ways of testing a model.

3.2.3 Results

In the results section of the blog, some of the final results of the classifier are explained and shown. When training a classifier, one of the most useful plots is the Precision-Recall curve. The resulting plot can be seen at the figure 11.

This figure shows the balance between the precision and recall of the model. This means that by looking at the function points, it can be estimated what the precision will be when demanding a certain recall.

Therefore, the optimal function point is to be at the top right corner, where both precision and recall are 1. This plot

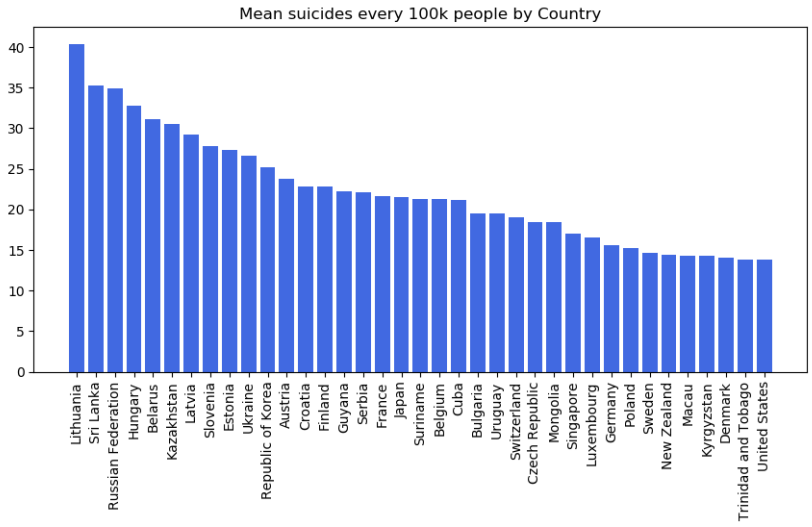


Fig. 5: Frequency of the number of suicides by country every 100k people. Limited at a minimum of 13 people.

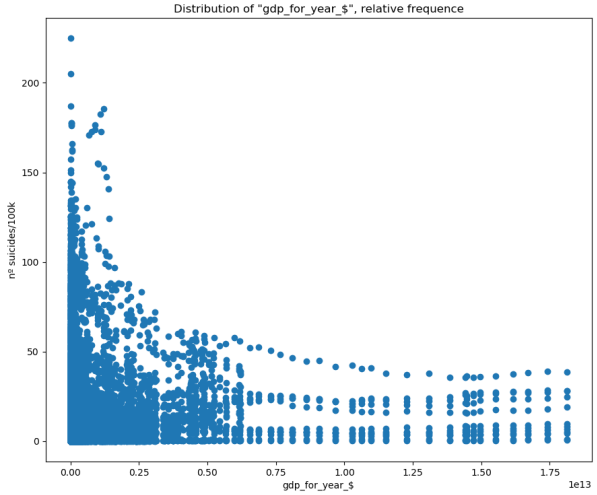


Fig. 6: Data distribution relating the nº of suicides and anual GDP and GDP per capita.

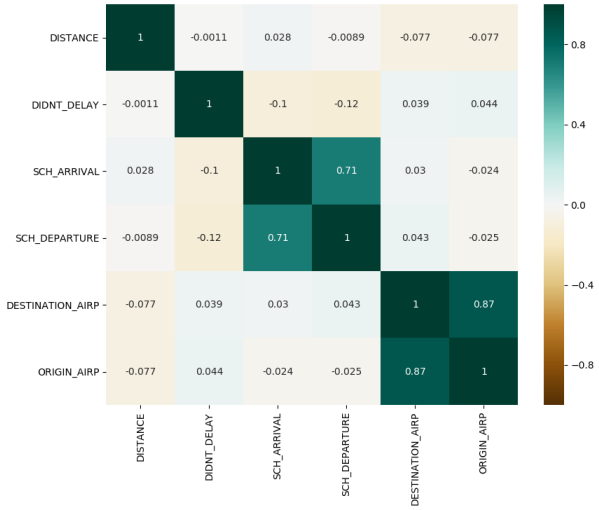


Fig. 7: Heatmap of the features correlation.

shows that the model performs well, because is really close to the top right corner.

Also, some properties of the model have been observed. The feature importances have been extracted from the model and can be seen here 9.The three factors that add the most information to the model are the scheduled arrival, scheduled departure and distance.

The model has shown that the most useful feature is the time of arrival. When getting the proportion of late flights over the time of the day, we get an interesting result. It corresponds to the figure 10.

The X axis corresponds to the time of the day in 24 hour format and with the colon taken out. This is what creates the little steps in the axis.

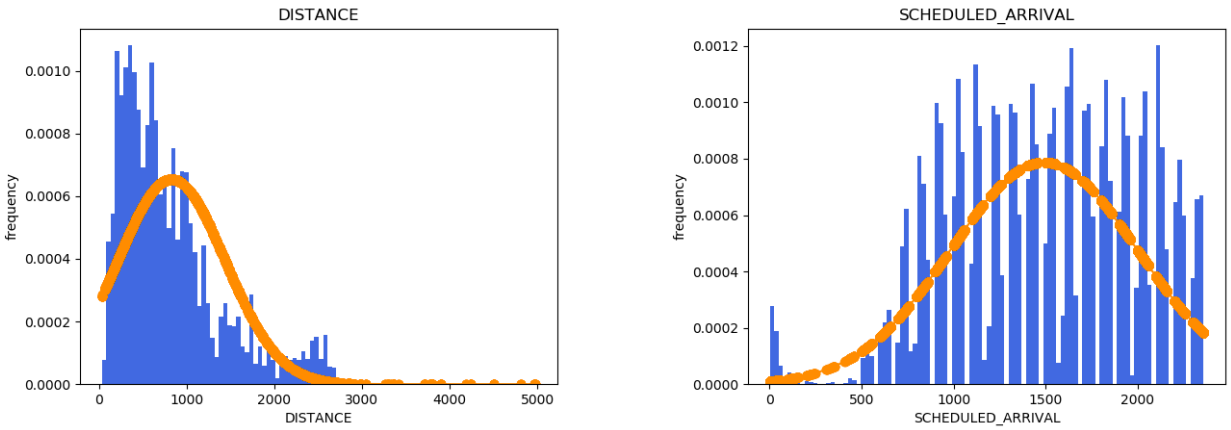
The growth we see in the Y axis is a high proportion of late flights that arrive from 02:00 to 05:00. Then, there's the absolute minimum of the function at about 7:30, where little proportion of the flights are late. This minimum has a value of about the 5% of the flights being late. After said minimum, the proportion grows approximately linearly until it reaches a 17.5% of the flights at 2:00.

This is very interesting, because if we rearrange the hours of the day to go from 07:30 to 7:30 of the next day we obtain a slow growth from 7:30 to 2:00, a fast growth from 02:00 to 03:30 and a fast reduction from 03:30 to 7:30.

3.3 Malaria Cells

For this problem, the dataset that has been used is the "Malaria Cells" [13] dataset. This is an image dataset, and it contains images of human cells. Those are distributed in two different classes: the healthy ones and the malaria infected ones.

Our goal here is to create a model that tells apart the infected cells from the rest.



(a) (b)

Fig. 8: Histograms for the features of flight distance and time of scheduled arrival with an overlayed gaussian.

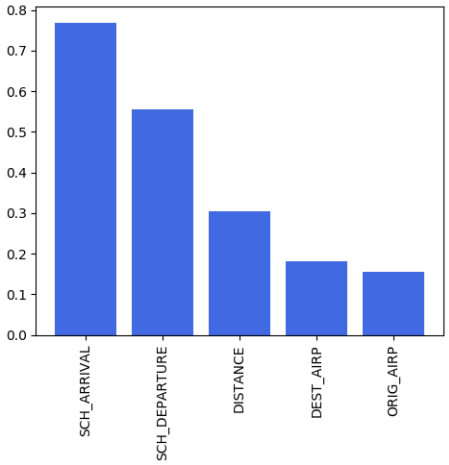


Fig. 9: Feature importance for the classification of the Flight Delays dataset.

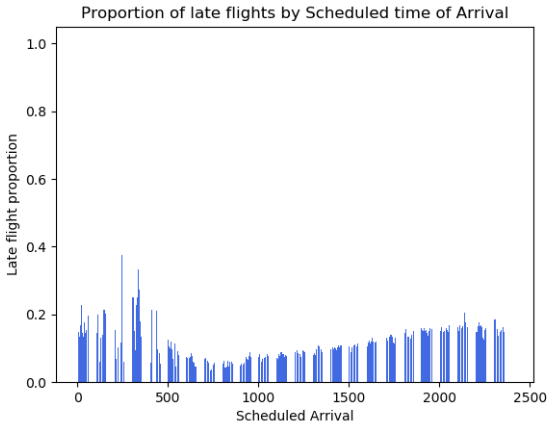


Fig. 10: Proportion of late flights depending on the Scheduled time of Arrival.

3.3.1 Data Analysis

The dataset contains 13800 samples of each class. The images have different sizes, with both the height and width being between 80 and 160 pixels.

Because of the different image sizes, all of them are randomly cropped and scaled to a 64x64 pixel image, using either 3:4 or 4:3 ratios. This, in a way, is a tool for data augmentation, but in this model it's being used to process the images before feeding them to the network.

In this section of the blog the class distribution is shown and some examples of images can be seen. Both uninfected and infected samples can be seen at the figure 12.

3.3.2 Machine Learning

Because the desired model must take an image on the input, the model that will be used is a convolutional neural network. This section of the blog tries to show the global composition of the convolutional neural network used and the purpose of each of its parts.

A diagram of the structure of the network can be seen at the figure 13.

The neural network is composed by some convolution and max pooling layers in order to extract as much features as possible from the image. Then, there are some fully-connected ReLU layers to learn the proper weights of the images.

However, in order to reduce the overfitting of the network and to raise the model's accuracy, some dropout layers have been added between all the convolution and linear blocks.

3.3.3 Results

Finally, in this section, the blog explains what properties a model like this should have in order to be acceptable. It is really important to understand the quality standards this model must have on its output, because this could be used as a medical tool, and some people's health would depend on

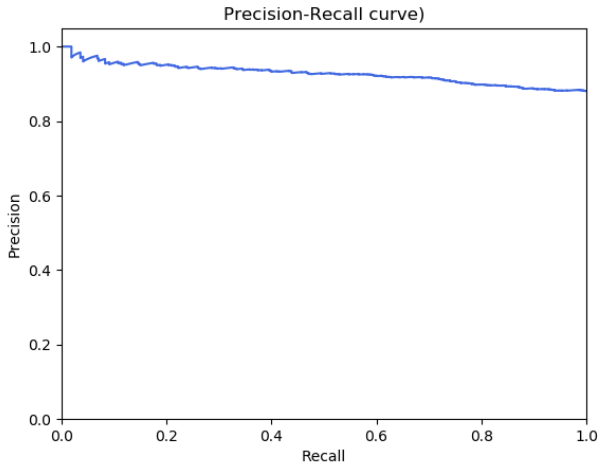
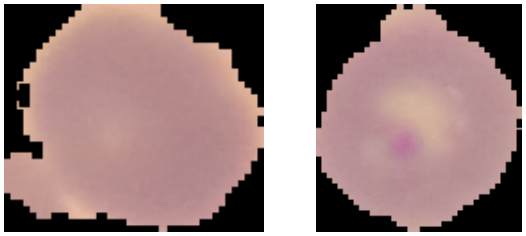


Fig. 11: Precision-Recall curve for the Flight Delays model.



(a) (b)

Fig. 12: Samples of a healthy cell (left) and a infected cell (right).

it. This is why a high recall on the model is desired, to reduce the false negatives as much as possible. The false negatives here would mean telling someone he's not infected while he indeed is.

In addition, while this model gets an accuracy of 0.86 and a recall of 0.93, the best solution found for this dataset has both accuracy and recall of 0.97. However, this solution implies using the previously defined ResNet34 within the PyTorch library. This, of course, doesn't hold as much value as creating a neural network from scratch for the blog. The solution found can be checked here [3].

Also, some properties of the model have been observed. There are some traits that determine with ease whether a cell is infected, and there are also some traits that all infected cells share but some healthy cells have too.

- The Shape: While healthy cells usually have a very round shape or, at least, a normal stretched shape, most of the infected cells have a shape that looks like if it had been torn apart. This makes the model mistake the healthy weird-shaped cells for infected ones and struggle to detect the infected round shaped cells. Cells of example can be seen at the figure 14.
- The Colour: The colour of the cells is also really important in order to determine whether it's infected or not. There are a total of three colours and their mixes: brown, pink and grey. While the grey colour is more common in the infected cells class, the other two are split between the two classes. Despite that, the brown colour is still more common in the healthy class. This is reflected in the model having special trouble when given a pink sample. Examples of the colours can be seen at the figure 15.
- The Stains: The cells of the dataset often show little purple stains. These are mostly present in the infected class. However, there are some uninfected samples that also contain these stains. This means that while the model can detect the infected cells with stains easily, the healthy cells that also have these stains are more likely to be mistaken. Examples can be seen at the figure 16

4 The Competition

The competition that has been chosen for this section uses the dataset "Titanic: Machine Learning from Disaster" [14].

This competition is public in the Kaggle webpage [12] and is a competition that has no prize attached to it. It has been set as permanently open, so it's the perfect target to use as a participation example.

The target of the competition is to guess whether a person survived or died in the Titanic disaster.

4.1 The Dataset

For this dataset, the first that should be explained is the feature meaning of the features.

- PClass: This feature contains a number that states the class the passenger travelled on. The lower the number is, the more luxurious the class they travelled on is.
- Sex: This feature is the sex of the passenger. It's either "male" or "female".
- Age: This feature is the age of the passenger in years.
- SibSp: This feature is the number of the passengers on board that were either their sibling or spouse.
- ParCh: This feature is the number of the passengers on board that were either their parent or child.
- Fare: This feature is the fare paid by the passenger.
- Embarked: This feature is the letter code referring to the port where the passenger embarked.
- Survived: This is our class and it's either 0 or 1, depending on if the passenger finally survived.

The next step is to see the relations between all the features of our dataset. This is best done by plotting the correlations between all our columns. The blog's image that shows all the correlations can be seen at the figure 17.

There are two correlations that stand out in the heatmap. The first correlation is between the "survived" class and the "sex" feature. This feature is given to the model with female

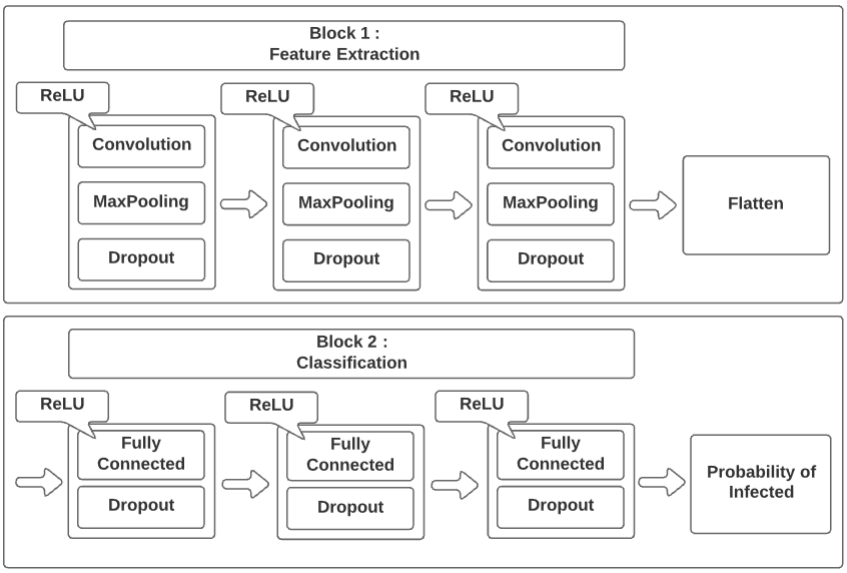
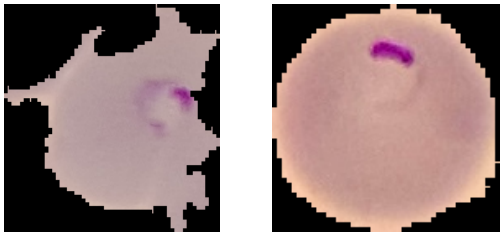
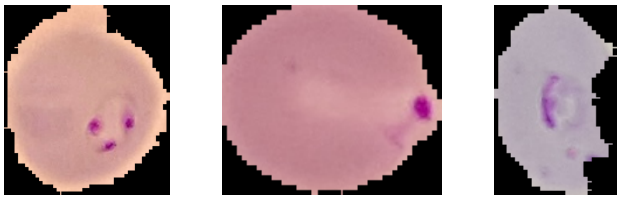


Fig. 13: Diagram showing the Neural Network structure.



(a) (b)

Fig. 14: Samples of infected cells. A torn apart cell (left), and a round-shaped cell (right).



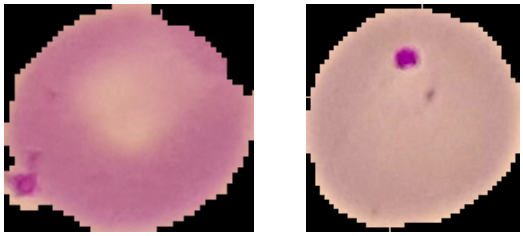
(a) (b) (c)

Fig. 15: Three samples of infected cells that differ in colour (brown, pink, grey).

being 1 and male 0, so it means that a lot more women survived than men. This makes sense, because many years ago, women would have preference to get safe with their children. Then, it makes sense that men have a lower chance of surviving the accident.

The other correlation is between the class the passenger travelled in and the fare paid by the passenger. This correlation makes sense, because the more luxurious the class is, the more expensive it will be to travel.

It's also important to check the feature distribution in order to give meaning to the data. Because of that, the histogram of the features should be generated. An example can be seen at the figure 18.



(a) (b)

Fig. 16: Uninfected cell (left) and infected cell (right). Both showing similar stains.

4.2 Own Solution

The purpose of this section is to apply the logistic regressor used in the problem of "Flight Delays and Cancellations of 2015".

This model gives us an accuracy of 0.77, a precision of 0.77 and a recall of 0.71. This information can be further inspected with the ROC curve [4]. The ROC curve of the model can be seen at the figure 19.

The ROC curve (i.e. Receiver Operating Characteristic) is a performance measurement that allows us to see the balance between the True Positive Rate and the False Positive Rate (TPR and FPR).

The best part to reach with ROC curve is the top left corner, because that means a TPR and FPR of 1, and that means a perfect classification. The figure also depicts the random guess curve, which would mean missing half of the guesses.

The ROC curve also features the AUC (Area Under the Curve). It's a number that can go from 0 to 1 and it indicates the quality of the model. Again, with the curve being in the top left corner, the area would be 1.

The ROC curve for this model, 0.85, is really good, but it has still room for improvement. The next subsection will talk about how to improve the metrics.

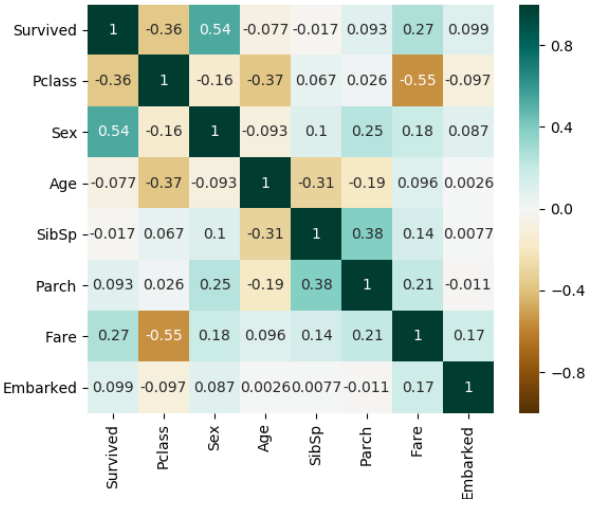


Fig. 17: Heatmap of the features correlation.

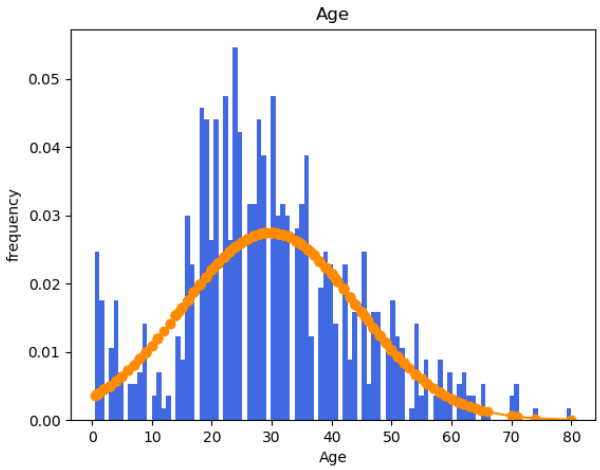


Fig. 18: Histograms for the age feature with an overlayed gaussian.

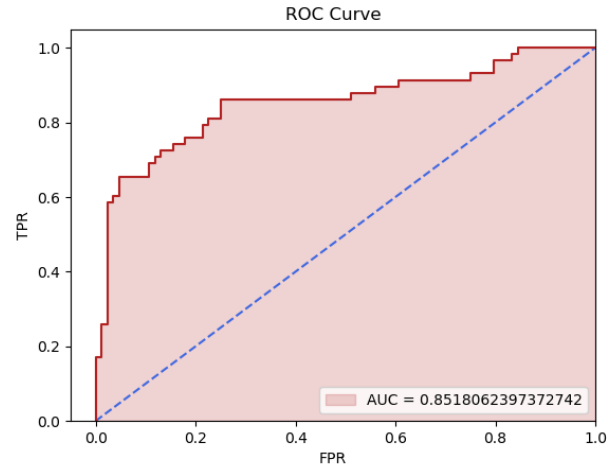


Fig. 19: ROC curve for the Titanic Disaster model.

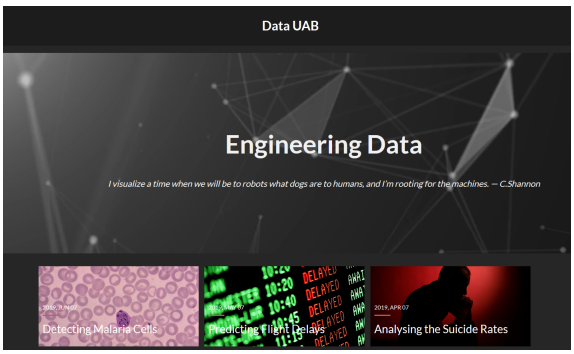


Fig. 20: DataUAB home page after the upload of the blogs.

4.3 The Best Solution

Despite the previous solution being good, it's far from the best solutions found.

In the submission part of the competition there are a lot of repeated kernels. The highest scores use a genetic algorithm for this problem, and they achieve an accuracy of 0.89. The problem is that the code of these kernels is useless because of it having been autogenerated and minified. An example of this genetic algorithm can be seen here [2].

However, there are other kernels that do offer valuable information. An example can be seen here [5]. This kernel uses different classifiers from the sklearn and xgboost libraries.

Said classifiers include AdaBoost, Bagging, ExtraTrees, GradientBoosting and RandomForest.

After testing these algorithms, the best test accuracies are from the XGBClassifier and the RandomForestClassifier. These two alone have an accuracy of 0.82.

However, the post from the kernel takes the DecisionTreeClassifier model (which gave an inferior accuracy) and decides to optimise it in order to improve its metrics. It does that by searching for the optimal hyper-parameters by using a grid search.

After obtaining the final hyper-parameters, the Decision Tree model has an accuracy of 0.89.

5 The Blogs

The content of the blogs added to the DataUAB page will surely be useful to future Machine Learning students. The blogs are at access here [16].

The project has been successful in the matter of adding more content to the DataUAB page. The blogs have been added through HTML generated with a Jupyter Notebook tool.

The blogs are the following: Analysing the Suicide Rates [6], Predicting Flight Delays [8] and Detecting Malaria Cells [7], and an image of the page after the upload can be seen at the figure 20.

6 Conclusions and Future Work

The main goal of this project has been to improve and refine the blogs and tools at the DataUAB blog [16] and to encourage the Machine Learning students to participate in competitions of these fields.

The biggest contribution has been the expansion of the knowledge by adding more blogs. Also, some issues of the platform have been fixed in order to make the platform more usable.

For the future work in this project, I think that the most important for the platform is to keep having contributions and growing. Even if the platform is not ready right now, it will soon become more important within the Machine Learning students. The most important, however, is to keep encouraging people to join and use the blog.

A good evolution for the blog would be to have a big community of students that uses it frequently.

Acknowledgements

I would like to express my appreciation and thanks to my advisor Jordi Gonzalez, for introducing me to the field of Computer Science, which I love, and for being of incredible help during the development of this project.

I would also like to thank my family for always supporting and helping me during the course of this degree.

Finally, I would like to express my gratitude to Laura, my partner, who has unconditionally been there, supporting me no matter what.

References

[1] Nerijus Adomaitis. Suicides in lithuania show social pains persist. https://www.reuters.com/article/us-lithuania-suicide/suicides-in-lithuania-show-social-pains-persist-idUSL10879374620080409. Last Visit: 2019-06-29.

[2] akshat113. Titanic dataset analysis. https://www.kaggle.com/akshat113/titanic-dataset-analysis-level-2. Last Visit: 2019-06-29.

[3] Danielh Carranza. Malaria detection with fastai v1. https://www.kaggle.com/ingbiodanielh/malaria-detection-with-fastai-v1. Last Visit: 2019-06-29.

[4] Kaggle Community. A data science framework: To achieve 99% accuracy. https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5. Last Visit: 2019-06-29.

[5] Kaggle Community. A data science framework: To achieve 99% accuracy. https://www.kaggle.com/ldfreeman3/a-data-science-framework-to-achieve-99-accura scriptVersionId=2051374. Last Visit: 2019-06-29.

[6] Jaume Estany. Analysing the suicide rates. https://datauab.github.io/suicide-rates/. Last Visit: 2019-06-29.

[7] Jaume Estany. Detecting malaria cells. https://datauab.github.io/malaria-cells/. Last Visit: 2019-06-29.

[8] Jaume Estany. Predicting flight delays. https://datauab.github.io/flight-delays/. Last Visit: 2019-06-29.

[9] Python Software Foundation. Python.org. https://www.python.org. Last Visit: 2019-06-29.

[10] Kaggle. 1985-2016 suicide rates. https://www.kaggle.com/russellyates88/suicide-rates-overview-1985-to-2016. Last Visit: 2019-06-29.

[11] Kaggle. 2015 flight delays and cancellations. https://www.kaggle.com/usdot/flight-delays. Last Visit: 2019-06-29.

[12] Kaggle. Kaggle. https://www.kaggle.com. Last Visit: 2019-06-29.

[13] Kaggle. Malaria cells. https://www.kaggle.com/iarunava/cell-images-for-detecting-malaria. Last Visit: 2019-06-29.

[14] Kaggle. Titanic: Machine learning from disaster. https://www.kaggle.com/c/titanic. Last Visit: 2019-06-29.

[15] Jupyter Notebook Team. Project jupyter. https://jupyter.org/. Last Visit: 2019-06-29.

[16] UAB. Data uab. https://datauab.github.io. Last Visit: 2019-06-29.

[17] W3Schools. Html5. https://www.w3schools.com/html/default.asp. Last Visit: 2019-06-29.

[18] James Watkins. The story behind russia's male suicide problem. https://www.ozy.com/acumen/the-story-behind-russias-male-suicide-problem 76845. Last Visit: 2019-06-29.

[19] Wikipedia. Overfitting. https://en.wikipedia.org/wiki/Overfitting. Last Visit: 2019-06-29.

[20] Walter Wuthmann. Suicide in sri lanka. http://www.dailynews.lk/2017/10/23/features/132102/suicide-sri-lanka. Last Visit: 2019-06-29.