Millora d'algorismes de predicció de propagació del Dengue amb eines d'intel·ligència artificial

A Degree Thesis

Submitted to the Faculty of the

Escola Tècnica d'Enginyeria de Telecomunicació de Barcelona

Universitat Politècnica de Catalunya

by

Victor Fayos i Pérez

In partial fulfilment

of the requirements for the degree in

TELECOMMUNICATIONS TECHNOLOGIES AND SERVICES ENGINEERING

Advisor: Mercè Vall-Llossera Ferran

Barcelona, June 2021

Abstract

It is well known that Dengue is a virus mainly spread by mosquito, the Aedes Aegypty, that is common in the area of Brazil. This country had been very affected by dengue disease, as well as ZiKa and chicunguña. This mosquito was originated in Africa, but now is found in tropical, subtropical and temperate regions throughout the world. Climatological conditions are very related with the number of mosquitoes and the viruses spread. The availability of Earth Observation mission that periodically and globally provide of climatological variables such as temperature and precipitations, may help to develop dengue prevention algorithms. The passive remote sensing group at Signal theory and propagation group lately had started to work in the development of such algorithms.

The main objective of this thesis has been the improvement of artificial intelligence algorithms for dengue risk detection in Brazil. For achieving such objective, the original database has been upgraded adding new years of information, with all the information: monthly dengue's cases in in Brazil, as well as all the climatologic and socioeconomic variables associated and necessary for the dengue risk prediction. In addition, the predictive algorithms based in Machine Learning and Deep Learning already developed in a previous project had been adapted with the new information in the database, as well as analyzed and improved.

This document details the information of the database, its analysis and explains the different statistical models that have been used to make prediction of the epidemiological risk of dengue in a month ahead. Moreover, an analysis of the performances of each model and a comparative study among them is presented.

Resum

És ben conegut que el Dengue és un virus principalment transmès per el mosquit, l'Aedes Aegypty, que és comú a la zona de Brasil. Aquest país ha estat molt afectat per aquesta malaltia, juntament amb el Zika i la chicuguña. El mosquit era originari de l'Africa, però ara es troba en àrees tropicals, subtropicals i temperades de tot el món. Les condicions meteorològiques estan molt relacionades amb el nombre de mosquits, i per tant amb la propagació del dengue. El fàcil accés a la missió "Earth Observation" que periòdicament proporciona a nivell mundial variables meteorològiques com la temperatura i les precipitacions, ha ajudat a desenvolupar algorismes de prevenció del Dengue. El grup de obtenció de dades passives al grup de teoria de senyal i propagació del virus ha començat a treballar en la implementació d'aquests.

El principal objectiu d'aquest projecte ha estat la millora dels algorismes d'inteligencia artificial encarregats de la predicció de dengue al Brasil. Per tal d'assolir aquest objectiu s'ha millorat la base de dades original afegint més anys d'informació, amb tota la informació: els casos mensuals de dengue a Brasil, també dades climatològiques i socioeconòmiques relacionades i necessàries per la predicció de dengue. A més a més, els algorismes predictius basats en Machine Learning i Deep Learning que ja han estat implementats en projectes anteriors han estat adaptats amb la nova informació de la base de dades, també com analitzats i millorats.

Aquest document detalla la informació de la base de dades, l'anàlisi d'aquesta i explica els diferents models estadístics utilitzats per fer una predicció a un mes vista del risc epidemiològic del dengue. Addicionalment, per cada model s'ha fet un anàlisi del seu rendiment i s'ha fet un estudi per comparar els diferents models.

Resumen

Es bien conocido que el Dengue es un virus principalmente transmitido por el mosquito, l'Aedes Aegypty, que es común en la zona de Brasil. Este país ha sido muy afectado por esta enfermedad, juntamente con el Zika i la Chikunguña. El mosquito era originario del África, pero ahora se encuentra en áreas tropicales, subtropicales y temperadas de todo el mundo. Las condiciones meteorológicas están muy relacionadas con el número de mosquitos, y por tanto con la propagación de Dengue. El fácil acceso a la misión "Earth Observation" que periódicamente proporciona a nivel mundial variables meteorológicas como la temperatura i las precipitaciones, ha ayudado al desarrollo de algoritmos de prevención del Dengue. El grupo de obtención de datos pasivos al grupo de teoría de señal i propagación del virus ha empezado a trabajar en la implementación de este.

El principal objetivo de este proyecto ha estado la mejora de los algoritmos de inteligencia artificial encargados de la predicción de dengue en Brasil. Para cumplir este objetivo se ha mejorado la base de datos original añadiendo más años de información, con toda esta información: el número de casos mensuales de dengue en Brasil, también datos climatológicos i socioeconómicos relacionados i necesarios para la predicción de dengue. A más a más, los algoritmos predictivos basados en Machine Learning i Deep Learning que ya han sido implementados en proyectos anteriores han sido adaptados con la nueva información de la base de datos, también analizados i mejorados.

Este documento detalla la información de la base de datos, analiza esta y explica los diferentes modelos estadísticos utilizados para hacer una predicción futura del riesgo epidemiológico del

dengue. Adicionalmente, para cada modelo se ha hecho un análisis de su rendimiento y se ha hecho un estudio para comparar los modelos.

Agraïments

M'agradaria expressar el meu agraïment a la Mercè Vall-Llossera, per la oportunita de poder realitzar aquest projecte i el seu guiatge. També a Hellen Gurgel, per compartit les dades necessàries per la correcta elaboració del projecte. A en Joaquim Bauxell per la resolució de dubtes i l'ajuda a l'hora de emprendre amb el projecte. I finalment a la UPC per les eines proporcionades per tal de facilitar el projecte.

Historial de revisions i aprovació

Revisió	Data	Propòsit
0	04/05/2021	Document creation
1	09/06/2021	Document revision
2	20/06/2021	Document revision

DOCUMENT DISTRIBUTION LIST

Nom	e-mail
Victor Fayos i Pérez	victor.fayos@estudiantat.upc.edu
Mercè Vall-Llossera Ferran	merce.vall-llossera@upc.edu

Escrit per:		Revisat i aprovat per:	
Data	04/05/2021	Data	20/06/2021
Nom	Victor Fayos i Pérez	Nom	Mercè Vall-Llossera Ferran
Posició	Autor del projecte	Posició	Supervisora del projecte

Taula de continguts

Llista d'il·lustracions

Llista de taules

1. Introducció

El dengue és una malaltia que es transmet a través de la picadura d'un mosquit infectat. Aquesta malaltia afecta a gent de totes les edats, causant febre, mal de cap, dolor darrere els ulls, dolor muscular o en les articulacions i en els casos més extrems la dificultat per respirar, sagnat greu o danys greus en els òrgans. El dengue té un comportament estacionari, es a dir, és més present en les estacions caloroses i amb pluja. En el cas del Brasil i pràcticament tota Amèrica, excepte Canadà i Xile, el principal propagador és el mosquit Aedes aegypti.

Aquest mosquit és un mosquit domèstic, viu a cases i a prop d'aquestes. Durant el cicle de vida de l'Aedes aegipti, un ou, pot està incubat des d'un període de dies a mesos, les larves viuen a l'aigua i en uns 5 dies es converteixen en crisàlides, les quals en uns 2 ó 3 dies es converteixen en mosquit adult capaç de volar. Com a mosquit adult tenen una vida de 4-6 setmanes. Només propaguen el virus les femelles ja que son aquestes les que necessiten sang per viure y per el desenvolupament dels òvuls. Cal destacar que és un mosquit actiu a la matinada i durant la nit. Els llocs on solen deixar els ous solen ser recipients artificials que continguin aigua, com poden ser barrils, tambors o llantes), aquests ous poden resistir condicions climatològiques adverses durant més d'un any. [1]

1.1. Objectius del projecte

El principal objectiu del projecte és crear un model capaç de predir el risc de dengue al Brasil, amb un mes d'antelació. És important sobretot tenir identificats els rics alts de dengue, és a dir, un mínim nombre de falsos negatius i un número raonable de falsos positius per tal de tenir una predicció fiable sobre el nombre de casos.. Comparar i trobar un model el qual es pugui utilitzar per ajudar combatre aquesta epidèmia, que cada any fa estralls a diversos països i concretament a Brasil.

1.2. Requeriments i especificacions

Els requeriments per aquest projecte són els següents:

- Les prediccions de risc siguin el més acurades possible, prioritzant tenir falsos positius a falsos negatius.
- Comparació dels diferents models utilitzats per la predicció de risc epidemiològic.
- Identificar les variables més útils en la predicció a un mes vista del risc epidemiològic.

Com l'objectiu del projecte és el desenvolupament d'un algorisme de predicció de risc de dengue al Brasil, totes les eines utilitzades son software. Les dades utilitzades han estat creades a partir de mesures de satèl·lit distribuïdes en plataformes gratuïtes d'accés públic. Els programes utilitzats han estat:

- Python en un entorn Anaconda: s'ha fet servir per a la programació de tots els models i per a l'estudi de la base de dades. En aquest entorn s'han instal·lat totes les llibreries necessàries per el bon funcionament dels models i per a la bona representació de les gràfiques.
- QGIS: per la visualització en forma de mapa de les dades utilitzades i per a la representació de les dades predites pels diferents models.

2. Base de dades:

Hellen Gurgel de la universitat de Brasilia, ens va proporcionar les dades del número de casos de dengue registrats al Brasil pel període 2010 – 2017 [2]. El període 2010 – 2013, la base de dades només conté el nombre de casos de dengue per municipi i per mes. El projecte del Joaquim Bauxell [3] , s'havia preparat la base de dades pels anys 2010 – 2013. A aquesta base de dades s'hi va incloure variables meteorològiques mesurades des de satèl·lit i també dades socioeconòmiques, com l'índex de desenvolupament humà municipal (IDHM) que és la variant a nivell municipal de l'índex de desenvolupament humà (IDH). L'IDH és una estadística d'ús estès amb l'objectiu de proporcionar una mesura que representi el nivell o qualitat de vida d'un lloc, es va crear l'any 1990 per la ONU. [4]

En aquest projecte s'ha preparat i afegit a la base de dades tota la informació referent als anys 2014 a 2017, doncs partim de la hipòtesis que el model serà més precís si hi afegim més anys d'informació. En aquest cas s'havia registrat a la base de dades cada cas de dengue amb la informació de l'edat, sexe, etc... Per els nostres models, de moment ens interessava convertir la informació en número de casos de dengue mensuals i per municipi, que és el que s'ha fet com a primer processament de la informació. En aquesta base de dades ampliada, també s'han introduït les següents variables:

- 'id_municip': número que identifica cada municipi de Brasil
- 'lat': latitud del punt central en el municipi
- 'lon': longitud del punt central en el municipi
- 'month': mes de la mostra
- 'year': any de la mostra
- 'idhm': índex de desenvolupament humà, calculat a partir de l'índex d'educació (IE), de longevitat (IL) i de renda (IR):

$$IDHM = \frac{IE+IL+IR}{3} \quad (1)$$

- 'rural_urban': índex indicador de la quantitat d'infraestructures urbanes d'un municipi
- 'state': primers dos números del identificador municipal que ens indiquen l'estat de Brasil d'on és el municipi.
- 'ndvi' (Normalized Difference Vegetation Index): índex de vegetació proporcionat pel satèl·lit Terra de la NASA, es calcula utilitzant les bandes d'infraroig proper (NIR) i vermell (Red):

$$NDVI = \frac{NIR-Red}{NIR+Red} \quad (2)$$

- 'ndwi' (Normalized Difference Water Index): índex d'acumulació d'aigua proporcionat per el satèl·lit Terra de la NASA, es calcula utilitzant les bandes infraroig proper (NIR) i les mitjanes (SWIR):

$$NDWI = \frac{NIR-SWIR}{NIR+SWIR} \quad (3)$$

- 'temp_day': temperatura mitjana durant el dia d'un mes, indicada en kèlvins. S'ha obtingut a partir de les mesures amb l'instrument MODIS abordo de Terra de la NASA. [5]
- 'temp_night': temperatura mitjana durant la nit d'un mes, indicada en kèlvins. S'ha obtingut a partir de les mesures amb l'instrument MODIS abordo de Terra de la NASA.
- 'moist': índex d'humitat del sol proporcionat per el satèl·lit SMOS. Les dades es distribueixen al Barcelona Expert Center (BEC).
- 'prec': precipitació acumulada durant un mes proporcionada pel satèl·lit GPM de la NASA i JAXA.
- 'cases': el número de casos de dengue notificats durant un mes en un municipi.
- 'risk': índex calculat a partir del número de casos per 100000 habitants i normalitzat a 500 casos, on considerarem un risc molt alt.
- 'next_cases': número de casos de dengue notificats durant el mes següent.
- 'next_risk': índex de risc notificat en el mes següent, aquesta característica serà la que anomenarem objectiu i la que serà predita.

2.1. Matriu de correlació

S'ha calculat la matriu de correlació entre les diferents variables que s'utilitzen a la base de dades per analitzar el seu comportament, si aporten informacions diferents o si estan molt correlacionades entre elles.

La matriu de correlació, *Fórmula 4*, calcula el grau de correlació de Pearson entre parelles de variables de la base de dades, els valors obtinguts estan normalitzats entre -1 i 1. Si el valor absolut d'aquesta correlació és proper a 1 llavors hi ha una relació forta entre aquelles dues variables, en canvi si és proper a 0 les dues variables són independents. El signe indica si la correlació és directament proporcional, en cas de signe positiu, o inversament proporcional, en cas de signe negatiu.

$$r_{xy} = \frac{\sigma_{xy}}{\sigma_x \cdot \sigma_y} \qquad R = \begin{bmatrix} r_{11} & \cdots & r_{1m} \\ \vdots & \ddots & \vdots \\ r_{n1} & \cdots & r_{nm} \end{bmatrix} \tag{4}$$

S'ha calculat la matriu de correlació de tota la base de dades preparada pel desenvolupament de l'algorisme de predicció de risc de dengue a Brasil (veure secció 3).

Taula 1. *Columna de la variable "next_risk" de la matriu de correlació de la base de dades respecte totes les variables*

	lon	lat	month	year	idhm	rural_urban	state	ndvi
next_risk	-0.008	0.0217	-0.0803	0.0412	0.0744	0.0750	0.0315	0.0286

	ndwi	temp_day	temp_night	moist	prec	cases	risk
next_risk	0.0268	-0.0340	-0.0346	0.0178	-0.0586	0.0966	0.5207

La *Taula 1* mostra els valors de la matriu de correlació per la variable "next-risk", que és la predicció de risc de dengue pel mes següent respecte de totes les variables/paràmetres d'entrada. Cal destacar que el risc a predir correla molt bé amb el risc del mes anterior. En una epidèmia té sentit que el número d'infectats en un mes estigui molt correlat amb el nombre d'infectats el mes següent.

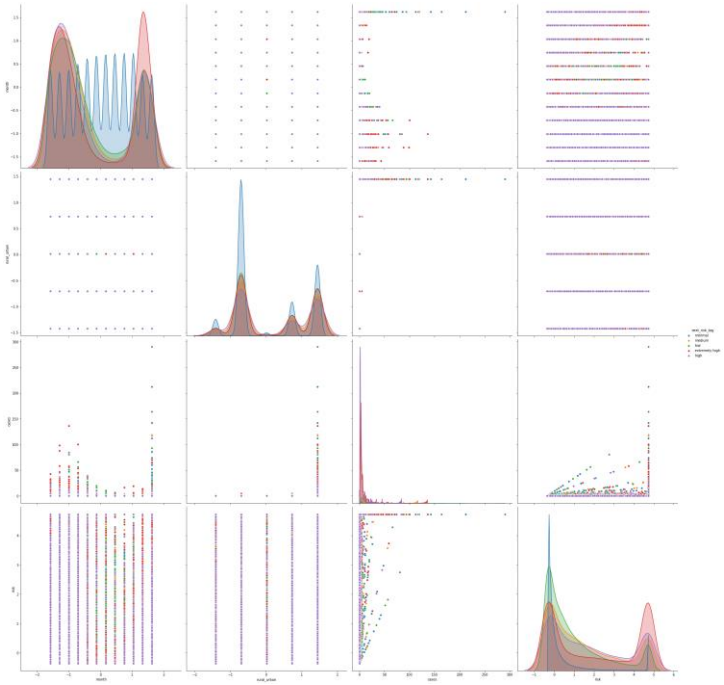
Després del risc, el número de casos, el mes i la classificació segons rural o urbà són les característiques que més alta correlació amb la probabilitat de risc del següent a partir de les dades actuals.

Aquests valors posen en manifest que un sistema de predicció lineal no serà òptim per aquest problema, ja que aquests sistemes funcionen molt bé quan hi ha alta correlació entre les variables de la base de dades.

Pairplot

Un pairplot és un tipus de representació de la informació d'una base de dades. Aquest té estructura de matriu, on a la diagonal hi ha la funció de densitat de probabilitat de la característica i en cadascuna de les cel·les hi ha representades dues característiques en forma de núvol de punts. Aquests tipus de gràfic serveix per identificar visualment si hi ha alguna relació entre parelles de variables que permetin predir fàcilment el paràmetre objectiu.

En l'estudi s'ha fet el pairplot només de les variables que més correlen, per veure si hi havia algun tipus de patró, veure *Il·lustració 1.*



Il·lustració 1. *Pairplot de les variables que més correlació tenen amb el risc del següent mes*

Per la base de dades d'estudi no s'ha identificat cap patró que ens permeti diferenciar els diferents nivells de risc segons parells de variables, no obstant es pot veure una cert patró en la funció de densitat de probabilitat del més, on es veu que hi ha certs mesos amb molta més incidència alta que altres on la incidència alta és inexistent.

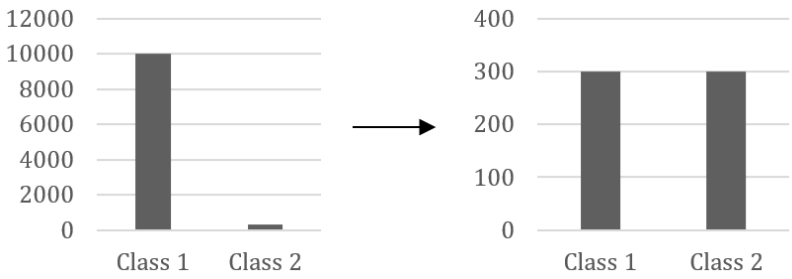
2.2. Data balance

Una base de dades balancejada és aquella que té un nombre igual o molt semblant de mostres de cada classe a predir. Per exemple, en una base de dades de reconeixement de números, hi ha la mateixa quantitat d'imatges de cadascun d'aquests. En el cas de l'estudi la que ens ocupa la base de dades inicial no és balancejada, ja que degut a la naturalesa de les dades hi ha molts més casos de riscs molt baixos (al voltant de 0) que de riscs alts. Per ser exactes la nostra base de dades de risc de dengue als municipis de Brasil té més del 90% de mostres amb el risc entre 0 i 0.2. Durant molts mesos de l'any no hi ha casos de dengue i a alguns municipis no n'hi ha durant tot l'any.

El problema que té tenir aquest tipus de base de dades és que a l'hora d'entrenar qualsevol algoritme de Machine Learning, aquest tendirà a predir valors molt propers a 0, ja que així minimitzarà l'error comès.

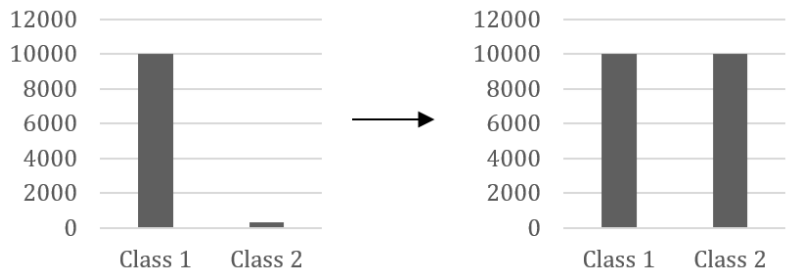
Hi ha diferents maneres d'arreglar aquest problema:

- Undersampling: consisteix en reduir el nombre de mostres de les classes que produeixen que la base de dades no estigui balancejada per tal de balancejar-la. La *Il·lustració 2*, mostra un exemple d'aquest procediment per un cas amb dues classes. Inicialment hi ha un número molt més gran de mostres de classe 1 que de classe 2. Aplicant undersampling, es redueix el nombre de mostres utilitzades de la classe 1 i per tant, també el nombre total de mostres.



Il·lustració 2. *Exemple d'undersampling*

- Oversampling: consisteix en generar artificialment dades de les classes amb menys quantitat de mostres per tal de balancejar la base de dades. La *Il·lustració 3* mostra com s'ha augmentat el nombre de mostres de la classe 2.



Il·lustració 3. *Exemple d'oversampling*

En aquest estudi s'ha aplicat undersampling ja que és un procés menys costós computacionalment i a més, les classes de la base de dades estan molt solapades entre elles i crear artificialment mostres no és una solució fiable.

S'ha aplicat un undersampling per aconseguir el millor rendiment de cada model. Degut a les característiques de cada model l'undersampling és més estricte o ho és menys. [6]

2.3. Dades d'entrenament i de test:

A l'hora de dissenyar un algorisme de Machine Learning o Deep Learning, la base de dades amb la qual es treballa ha de ser segmentada en dues parts: una part per l'entrenament i una altra pel test.

Abans d'això s'ha analitzat el nombre de casos per cada any per a tot Brasil, mostrat en la *Il·lustració 4*. Si un any els número de casos es excessivament baix o excessivament alt aquest s'hauria de descartar de la base de dades, ja que crearia una falsa tendència en el model.

Il·lustració 4. *Evolució de casos de dengue de la base de dades per mes. Dades corresponents a tot Brasil pels anys 2010 a 2017. Les línies verticals identifiquen l'inici d'any*

De l'any 2010 fins l'any 2014 hi ha un patró, un pic al desembre i una baixada de casos al gener. Ara bé a partir d'aquest any, probablement degut a les mesures aplicades per el govern de Brasil per la celebració del mundial de futbol, la tendència canvia i el pic de casos varia i es posiciona a l'abril. També l'any 2017 hi ha una baixada dràstica de casos, per aquesta raó s'ha decidit excloure aquest any a l'hora d'entrenar i testejar la base de dades.

Es conegut que per tècniques de Machine Learning es divideix la informació en dades d'entrenament de l'algorisme i dades de test. Per tal de tenir una predicció fiable s'ha decidit testejar amb un any que la base de dades d'entrenament no ha utilitzat encara. Per tant s'agafarà de 2010 fins a 2015 per fer la base de dades d'entrenament i el 2016 per fer la base de dades de test.

2.4. Base de dades d'imatges

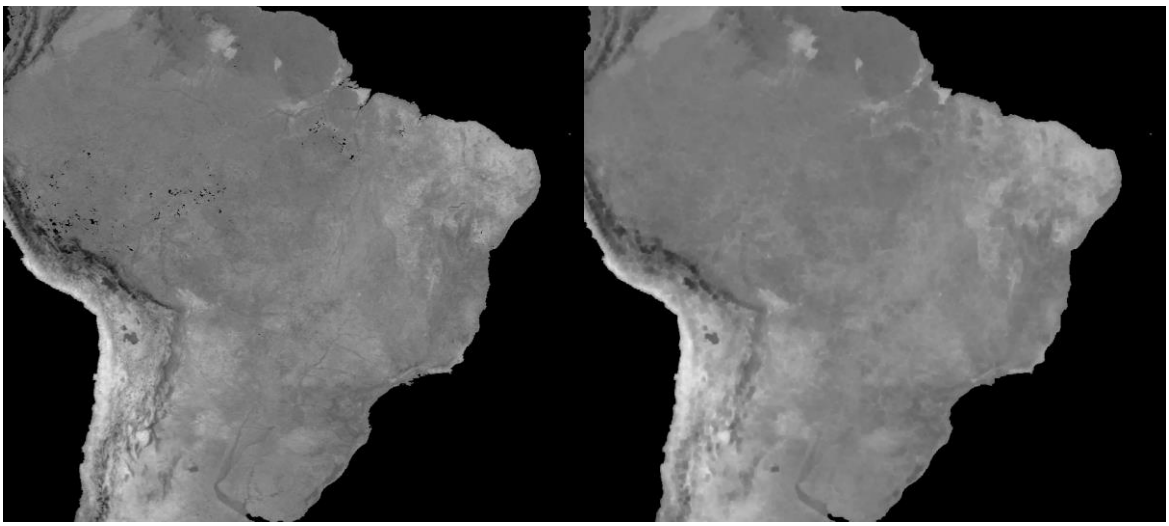
A més de la base de dades en format .csv, per la possible preparació d'un model CNN. S'ha transformat aquesta base de dades original a un format de matriu.

Per poder convertir-ho, s'ha determinat les latituds i longituds per les quals tot Brasil està representat, aquestes han estat: [6.40º,-34.95º] de latitud i [-72.95,-31.25] de longitud.

Per tal d'extreure les característiques en imatge s'ha seleccionat la part desitjada de tota l'imatge proporcionada pel satèl·lit. Degut a que els satèl·lits utilitzats tenen una resolució espacial de 0.05º les imatges creades són de 934x826 píxels.

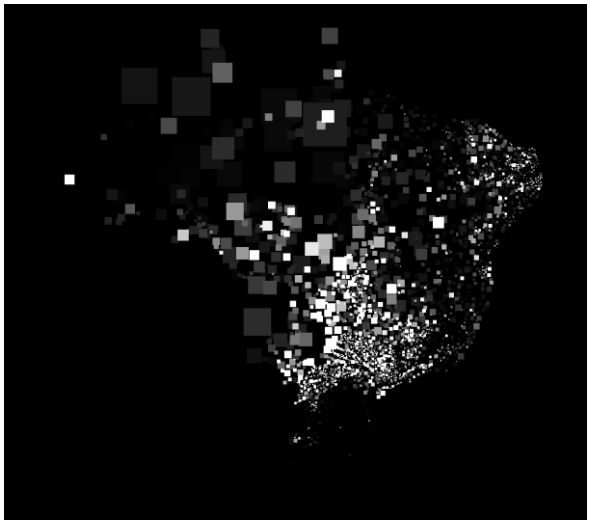
Ara bé moltes de les dades contenien parts on no hi havia lectura de la dada desitjada. Per exemple la temperatura de dia en certes zones de l'Amazones no era possible captar-la i això crea zones a la imatge on no hi ha dades. La manera de solucionar el problema és amb un tancament morfològic amb un element estructurant constant i de mida 5x5. Aquesta operació omple forats a una imatge agafant els valors màxims dins l'element estructurant.

A la *Il·lustració 5* es pot veure un exemple d'una imatge sense tancament i una amb tancament. A la imatge de l'esquerra hi ha alguns punts totalment negres (no hi ha dada) que no apareixen a la de la dreta, després de fer-ne el tractament.



Il·lustració 5. *Imatge de la temperatura de dia abans i després del tancament morfològic*

Per tal de representar el risc, les imatges s'han creat a partir de la posició dels municipis. En aquesta posició s'ha creat un quadrat de l'àrea del municipi que s'està representant. A la *Il·lustració 6* es pot veure un exemple de mostra de risc en aquesta base de dades.



Il·lustració 6. *Mostra de risc a la base de dades d'imatges*

3. Models de Machine Learning i Deep Learning:

Un model de Machine Learning o Deep Learning és un programa que analitza una base de dades (amb informacions de diverses variables) i produeix una sèrie de normes, patrons o pesos per tal de poder predir o classificar altres dades.

Segons la metodologia seguida per entrenar aquests models es poden classificar de la següent manera:

- Supervisats: el model s'entrena sabent quina és la classe o valor a predir de cada mostra d'entrenament.
- No supervisats: el model s'entrena sense saber a quina classe pertany cada mostra. Aquests models solen ser utilitzat pel que en anglès s'anomena 'clustering' on es volen veure possibles relacions entre les diferents variables.
- Semi supervisats: el model s'entrena sabent algunes de les classes o valors a predir de cada mostra.

Aquests algorismes poden ser programats bàsicament per resoldre problemes de dues vessants: problemes de classificació i problemes de regressió.

Els problemes de **classificació** són aquells on cada mostra pertany a una classe i l'algorisme ha d'identificar quina és aquesta classe. Un exemple clàssic seria la classificació d'imatges de gat i gos, on l'algorisme identifica si la imatge d'entrada és d'un gat o d'un gos.

Els problemes de **regressió** són aquells on l'algorisme no prediu una classe però prediu un valor numèric en funció de l'entrada. Un exemple seria l'estimació de les ventes d'una empresa segons els mesos anteriors.

En aquest estudi s'ha optat per desenvolupar algorismes regressius. A continuació, es descriuen els que s'han utilitzat per aquest TFG.

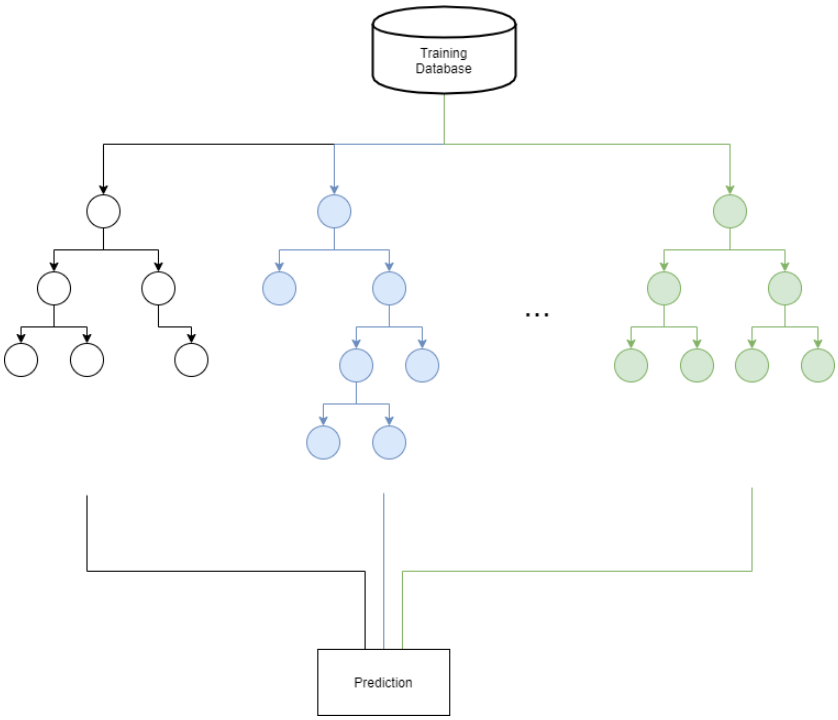
3.1. Random Forest

L'algorisme de random forest és un algorisme de Machine learning molt robust i que pot ser utilitzat per problemes de regressió i per problemes de classificació, sent aquests darrers on té un major rendiment.

El random forest pertany a un grup dels mètodes anomenat 'ensemble method'. Aquest grup es caracteritza per fer la predicció utilitzant moltes prediccions d'un model senzill. En el cas del

random forest s'utilitzen un conjunt de decision trees per tal de fer la classificació o predicció, veure *Il·lustració 7*.

El decision tree és un algorisme de Machine learning el qual genera fronteres de decisió per tal de separar les mostres de la base de dades segons la característica a predir. Aquest tipus de model és molt senzill i tendeix a l'overfitting, per això normalment s'utilitza el random forest per tal d'afegir robustesa a l'hora de la predicció.



Il·lustració 7. *Esquema general Random Forest*

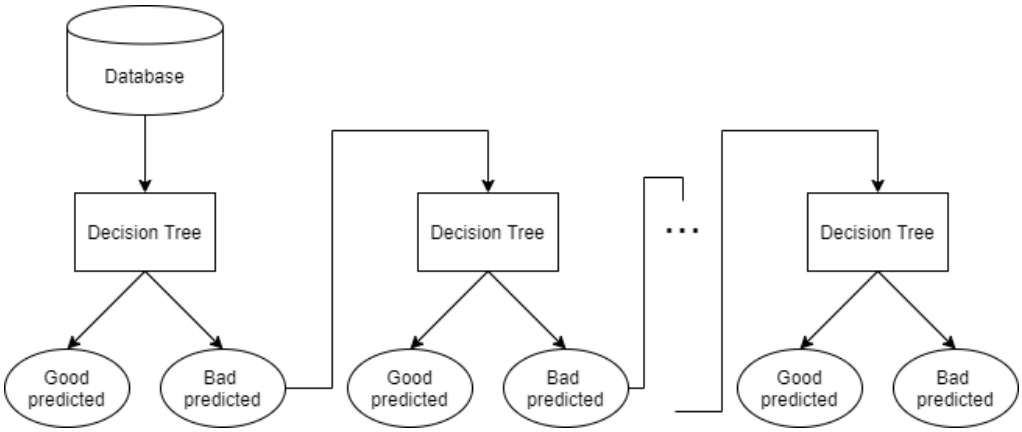
En aquest estudi s'ha utilitzat aquest model per tal d'obtenir els primers resultats. Aquest model és senzill d'aplicar i a més permet fer una anàlisi de les dades que s'utilitzen per la predicció. [7]

3.2. Extreme Gradient Boosting

L'algorisme de Extreme Gradient Boosting és un algorisme de Machine learning utilitzat per tot tipus de problemes, classificació i regressió. És molt robust i molt ràpid d'entrenar.

Aquest igual que el Random Forest està emmarcat dins la classe de 'ensemble method' i també utilitza com a model senzill el decision tree. A més és un algorisme que també pertany a la classe boosting, el que vol dir que els diferents models senzills es van creant a partir dels errors de predicció del model global, és a dir, cada model senzill que s'afegeix al model global intenta

solucionar errors de predicció en el model global. La *Il·lustració 8* mostra un esquema d'aquest tipus de model.

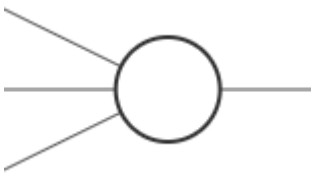


Il·lustració 8. *Esquema Extreme Gradient Boosting*

A més aquest model utilitza una mesura de pèrdua diferenciable i un algorisme d'optimització de gradient descendent, el qual minimitza la pèrdua com si d'una xarxa neuronal es tractés. [8]

3.3. Artificial Neural Network (ANN)

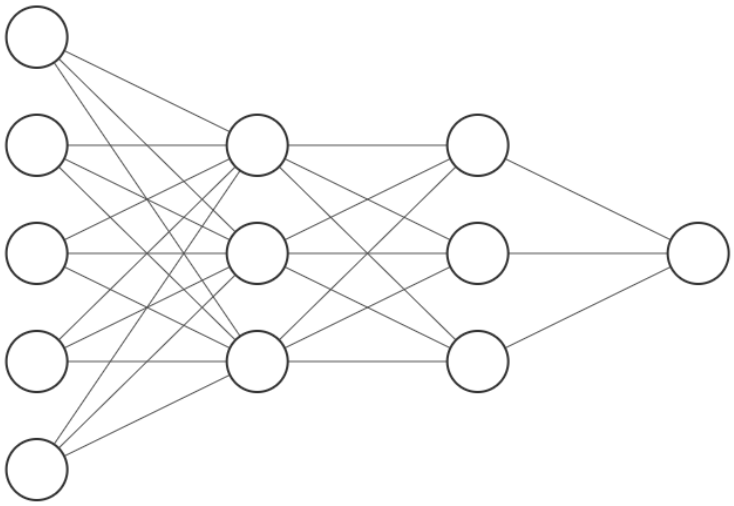
Una xarxa neuronal artificial és un algorisme molt utilitzat tant per problemes de classificació com de regressió. Aquest model esta format per una sèrie de capes de perceptrons, els quals són estructures molt simples d'on a través dels valors de les entrades, que són les sortides de les altres neurones i normalment un valor de biaix, i d'una funció d'activació (RELU, leaky-RELU, sigmoid, ...) es calcula una sortida, seguint la *Fórmula 5*.



Il·lustració 9. *Esquema neurona xarxa neuronal*

$$y_j = g(b_j + \sum_i x_i \cdot w_i) \quad (5)$$

Aquests perceptrons estan connectats per capes, aquestes anomenades capes ocultes, i on tots aquests estan connectats entre ells, veure *Il·lustració 9*. [9]



Il·lustració 10. *Esquema Artificial Neural Network*

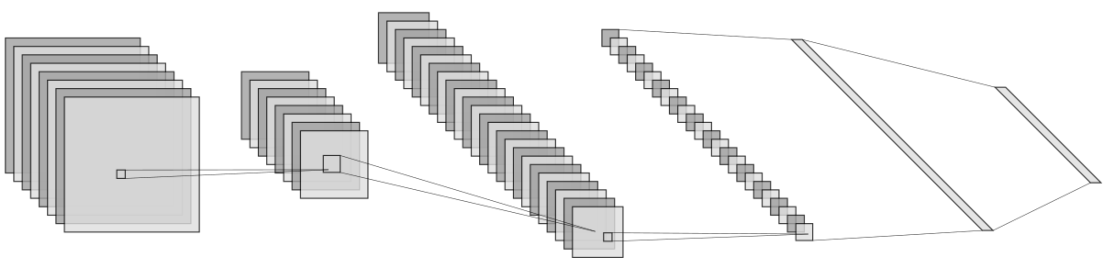
3.4. Convolutional Neural Network (CNN)

Una xarxa neuronal convolucional és un tipus de xarxa neuronal on les entrades no són d'una dimensió, normalment de 2 ó 3. Aquestes per obtenir el valors per tal d'activar la funció del perceptró utilitzen filtres, comunament anomenats kernels, que convolucionen amb l'entrada. Per tant la sortida de cadascuna d'aquestes neurones ve donada per la *Fórmula 6*.

$$Y_j = g\left(b_j + \sum_i K_{ij} \otimes Y_i\right) \quad (6)$$

On la sortida Y de la neurona j és una matriu que es calcula a partir de la convolució de cadascuna de les sortides de la capa anterior, Yi, per el conjunt de filtres Kij més el biaix de cada neurona, tot això dins la funció d'activació g().

A més aquest tipus de xarxes després de cada convolució sol fer un mapeig no causal per tal de reduir la dimensió de la sortida. L'esquema de la figura 8 en resumeix el procediment. [10]



Il·lustració 11. *Esquema Convolutional Neural Network*

4. Resultats

En aquest apartat es presenta primer quines han estat les mesures de qualitat analitzades per cadascun dels models programats, per després poder analitzar i comparar els diferents models programats i millorats.

4.1. Mesures de qualitat

Per tal d'entendre la nomenclatura de les fórmules de les mesures de qualitat utilitzades es defineixen els següents paràmetres:

- TP (True Positives): nombre de mostres ben predites en una classe.
- FP (False Positives): nombre de mostres predites a una classe superior.
- FN (False Negatives): nombre de mostres predites a una classe inferior.

4.1.1. Matriu de Confusió

Aquesta mesura consisteix en una matriu on les files representen el número de mostres predites per una classe i les columnes el número de mostres realment d'aquesta classe.

4.1.2. Precisió

La precisió es el una mesura que ens indica el grau d'encert d'una predicció, va de 0 a 1 i la es calcula de la següent manera:

$$P = \frac{TP}{Total\ Class} \quad (7)$$

En aquesta base de dades no es pot tenir en consideració la precisió total ja que hi ha un número molt gran de mostres de la classe més baixa. Per això la precisió ha estat calculada a partir de la mitjana aritmètica de la precisió de cada classe.

4.1.3. Precisió pessimista

Aquesta mesura ha estat dissenyada per saber el grau d'encert o d'encert d'una classe superior, per tant en el àmbit d'una pandèmia poder optar mesures encara que al final el grau d'incidència real sigui més baix que el predit. Es calcula amb la següent fórmula:

$$PP = \frac{TP+FP}{Total} \quad (8)$$

Degut a que en aquest cas no es te en compte la classe més baixa, ja que sempre ens sortirà 1, només es calcula a partir de de les altres classes en conjunt.

4.1.4. Diagnòstic odd ratio (DOR)

Utilitzat principalment en una mesura de qualitat per a tests binaris, va de 0 a infinit i es calcula de la següent forma:

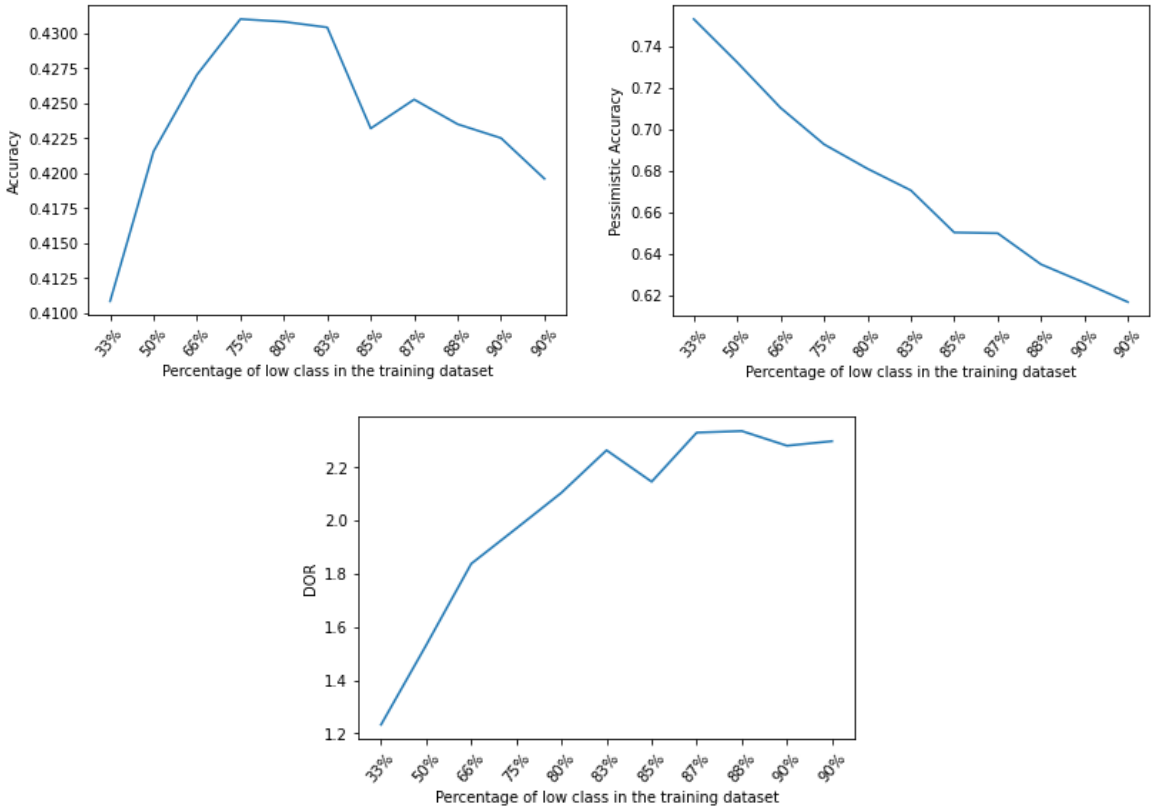
$$DOR = \frac{TP*TN}{FP*FN} \quad (9)$$

Aquesta mesura esta pensada per ser calculada per classe i per tant si volem fer una general la farem amb una mitjana ponderada per el número de mostres. [11]

4.2. Avaluació de resultats

4.2.1. Random Forest

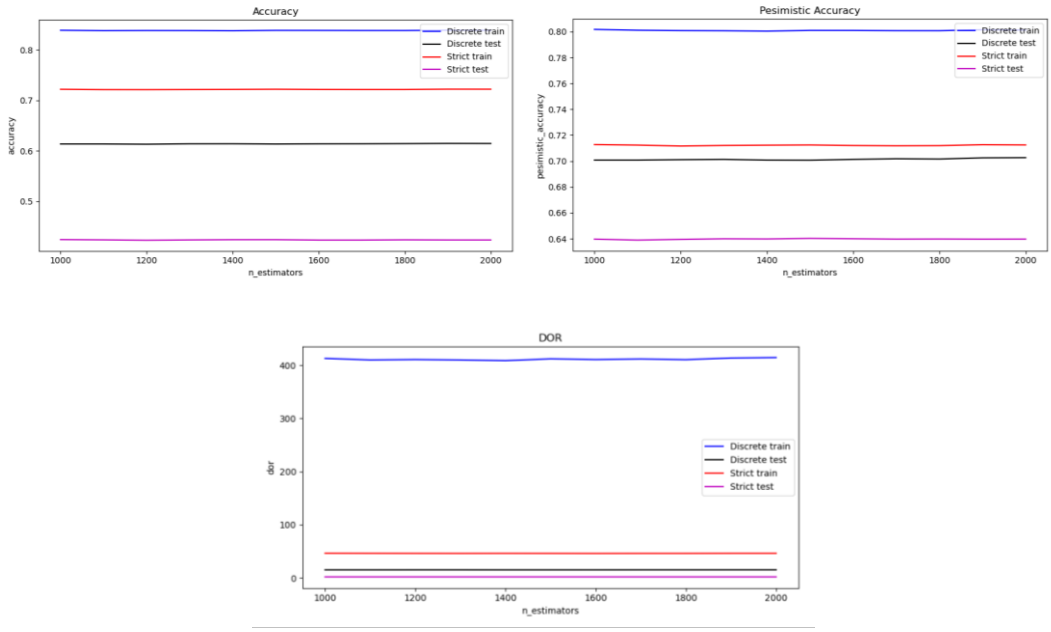
Aquest algorisme és molt robust a l'overfitting. Això permet que l'undersampling sigui menys estricte i encara així obtenir bons resultats. Per obtenir el número de mostres de classe baixa óptim s'ha implementat un algorisme iteratiu on es testeja un model de 1000 arbres amb diferents número de mostres de la classe baixa. En la *Il·lustració 12* es veuen els resultats de les diferents mesures de qualitat.



Il·lustració 12. *Accuracy, Pessimistic Accuracy i DOR d'un model Random Forest de 1000 arbres segons el número de mostres de la classe Baixa*

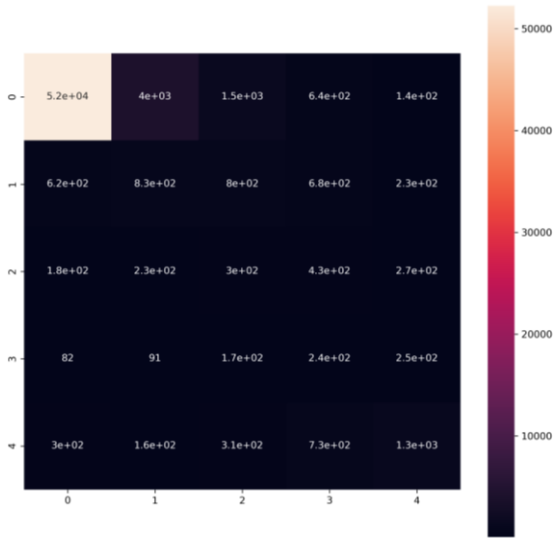
A partir d'aquest resultats, és farà un undersampling per tal que el número de mostres de classe baixa sigui el 87% de la base de dades d'entrenament. S'ha escollit aquest número ja que és el que té un DOR més alt d'entre tots els testejats.

Per tal de trobar el número d'arbres del random forest s'ha fet un testeig exhaustiu amb el número de mostres seleccionat anteriorment. Igualment que abans s'han fet unes gràfiques de les mesures de qualitat en funció del número d'estimadors que componen el model i s'ha representat a la *Il·lustració 13*.



Il·lustració 13. *Accuracy, Pessimistic Accuracy i DOR del model Random Forest*

Per poder analitzar la matriu de confusió, *Il·lustració 14*, s'ha triat el model de 1500 arbres, ja que aquest té un DOR una mica més superior que els altres.



Il·lustració 14. *Matriu de confusió del Random Forest*

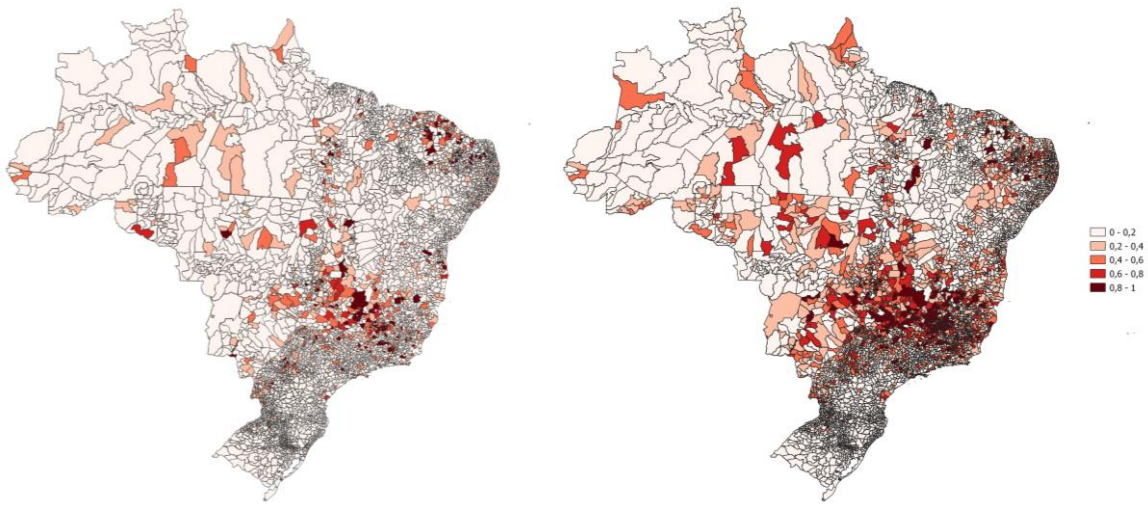
La matriu de confusió, posa de manifest que aquest estimador no representa bé cap nivell. Els nivells 1, 2 i 4 estan representats incorrectament, ja que s'estima el mateix número de mostres a tots el nivells. Per altra banda el nivell 0 i el nivell 3 estan ben representats ja que els valors predits estan a prop dels valors reals, a prop de la diagonal de la matriu. Aquest estimador no té una bona

accuracy, ja que aquesta no arriba al 0.5, té un bon pessimistic, aquest arriba més del 0.5 i el qual pot compensar la manca), si el que interessa és sobretot no perdre cap risc alt. Per la part del DOR, aquest és més gran que 2 i això ens indica un estimador que no és òptim però amb un rendiment mínimament acceptable.

Taula 2. *Mesures de qualitat Random Forest*

Random Forest	
Accuracy	0.4244
Pessimistic Accuracy	0.6402
DOR	2.2086

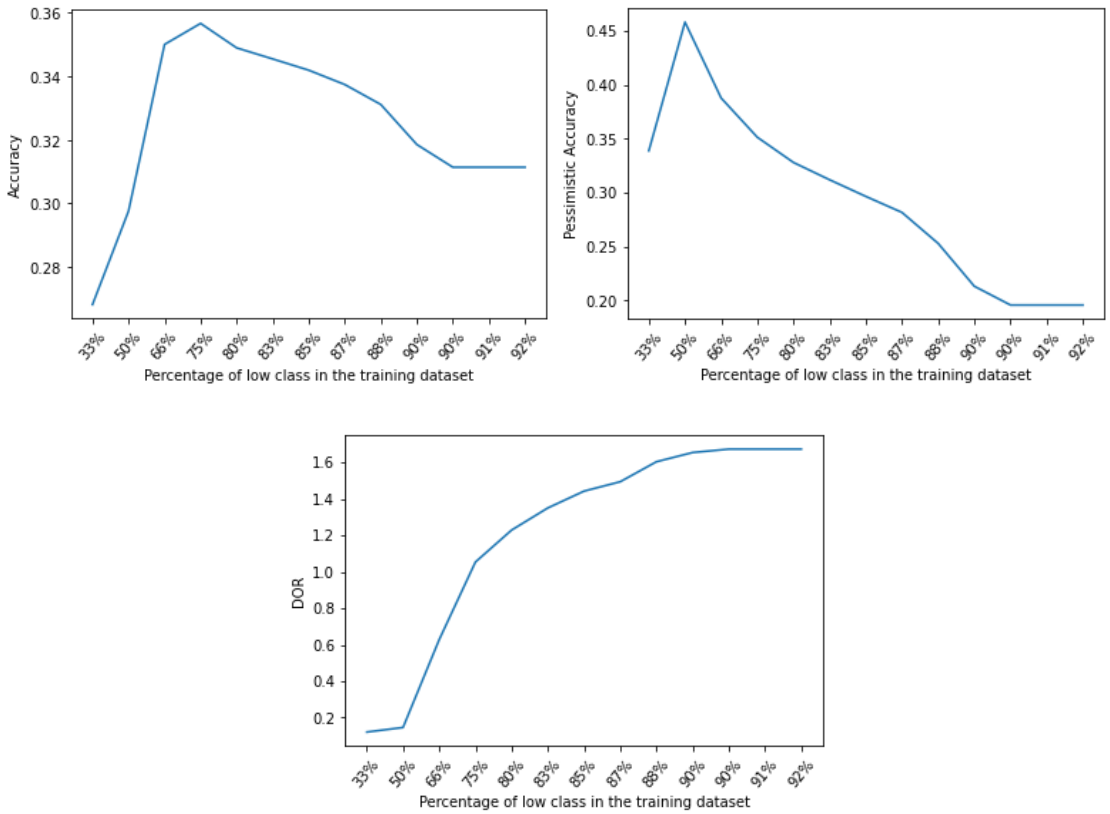
Per poder il·lustrar-ho s'ha representat el risc real i el predit del maig de 2016, *Il·lustració 15.* Risc real vs Risc predit Random Forest. Es veu clarament com els nuclis de més risc estan molt ben predits encara que el valor exacte del risc no estigui ben predit. Els casos on hi ha menys risc no estan tan ben caracteritzats com els casos alts. Aquest model és pessimista, ja que en la majoria de casos sol predir un valor més alt de risc que el real. Tot i no tenir una bona precisió, es podria fer servir per tenir una visió pessimista global del risc al país.



Il·lustració 15. *Risc real vs Risc predit Random Forest*

4.2.2. Extreme Gradient Boosting

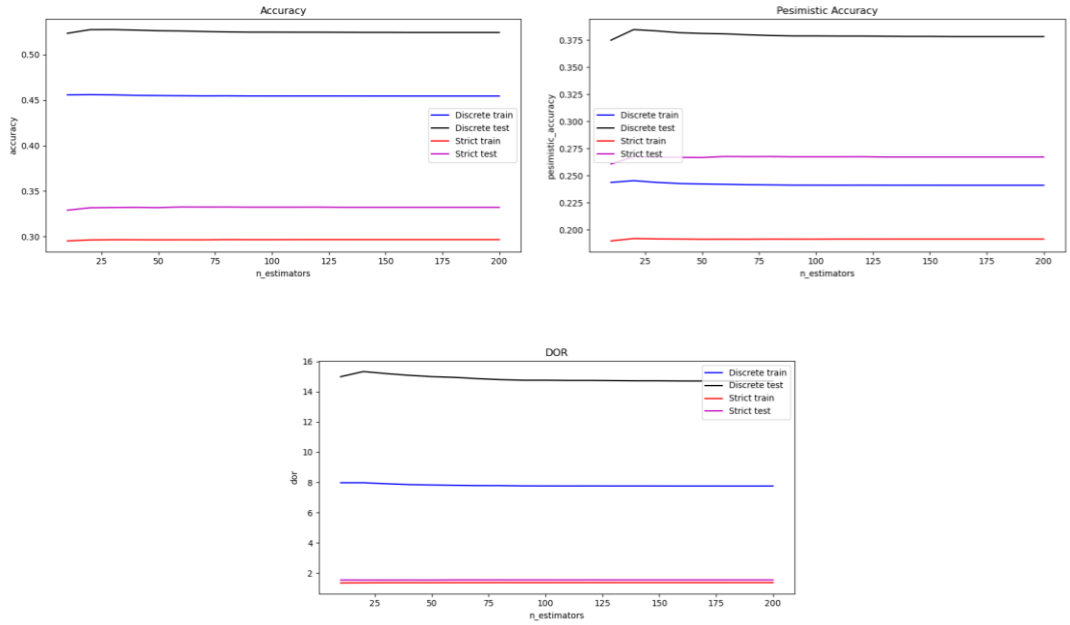
Per tal de trobar el número òptim d'undersampling s'ha seguit el mateix procediment que pel Random Forest, ara bé degut a que aquest model tendeix a fer overfitting s'ha utilitzat un model de 20 arbres per tal d'aproximar el número de mostres de nivell baix òptim, veure *Il·lustració 16*.



Il·lustració 16. *Accuracy, Pessimistic Accuracy i DOR d'un model XGBoosting de 40 arbres segons el número de mostres de la classe Baixa*

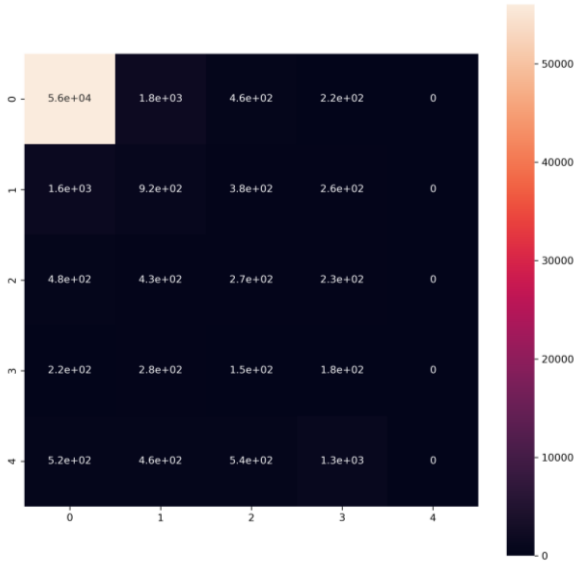
S'observa que tant l'accuracy co, la pessimistic accuracy decreixen si hi posem més dades d'entrenament, ara bé el DOR creix i per tant utilitzarem aquest últim per poder escollir el número de dades òptim. S'ha escollit agafar el 90% de les dades d'entrenament de nivell baix ja que és el que té un número major de DOR amb un número menor de dades d'entrenament.

Tal com ocorria en el random forest per tal de trobar el número òptim d'estimadors s'ha fet un testejat de forma exhaustiva les dades, veure *Il·lustració 17*.



Il·lustració 17. *Accuracy, Pessimistic Accuracy i DOR del model XGBoosting*

Per poder analitzar la matriu de confusió, *Il·lustració 18*, s'ha agafat el model de 120 arbres, ja que aquest el model és el que té el DOR més alt.



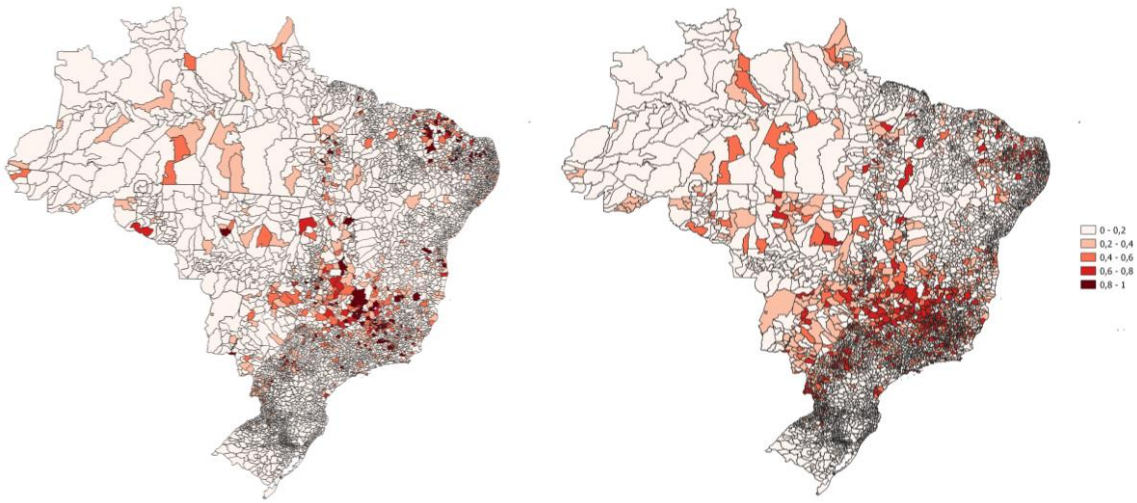
Il·lustració 18. *Matriu de confusió del Extreme Gradient Boosting per 120 arbres*

En aquest model es pot veure com no hi ha cap predicció de nivell alt. Aquesta predicció està feta a 4 nivells i no a 5 com es volia inicialment. Els nivells per això no estan malament caracteritzats, sobre tot el més alt i el més baix, el que ens pot indicar que els nuclis de casos els detectarà bé però amb un nivell inferior a l'esperat.

Taula 3. *Mesures de qualitat Extreme Gradient Boosting*

Extreme Gradient Boosting	
Accuracy	0.3324
Pessimistic Accuracy	0.2674
DOR	1.6205

Per poder il·lustrar-ho s'ha representat el risc real i el predit de maig de 2016, *Il·lustració 19*. Es pot veure com no hi ha cap nivell alt predit, encara que els nuclis de risc estan ben predits no tenen el nivell de risc corresponent. Aquest estimador és optimista ja que prediu nivells més baixos als reals, encara que pot ser fiable per tenir una visió general del risc del país.

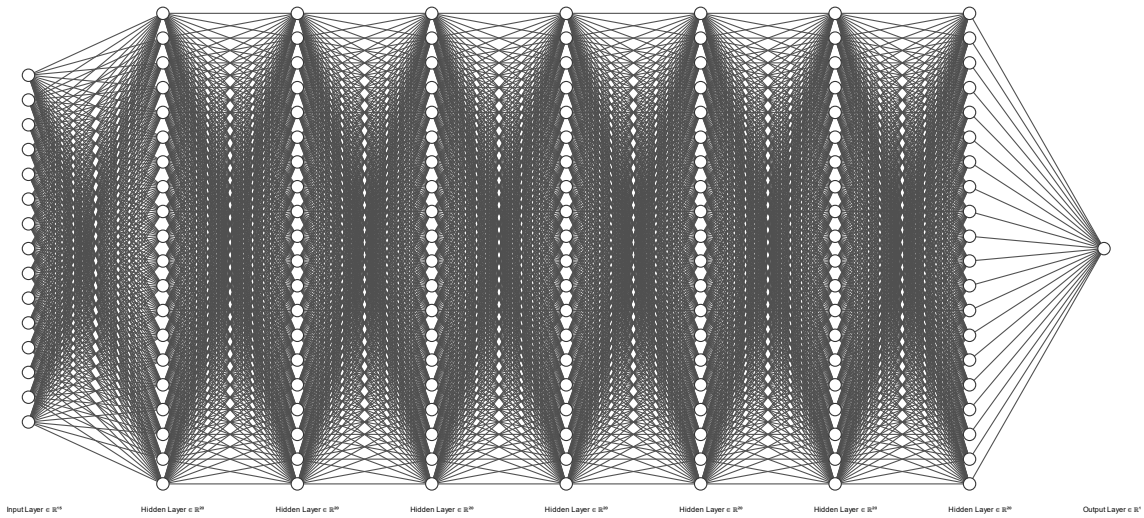


Il·lustració 19. *Risc real vs Risc predit Extreme Gradient Boosting*

4.2.3. Artifitial Neural Network (ANN)

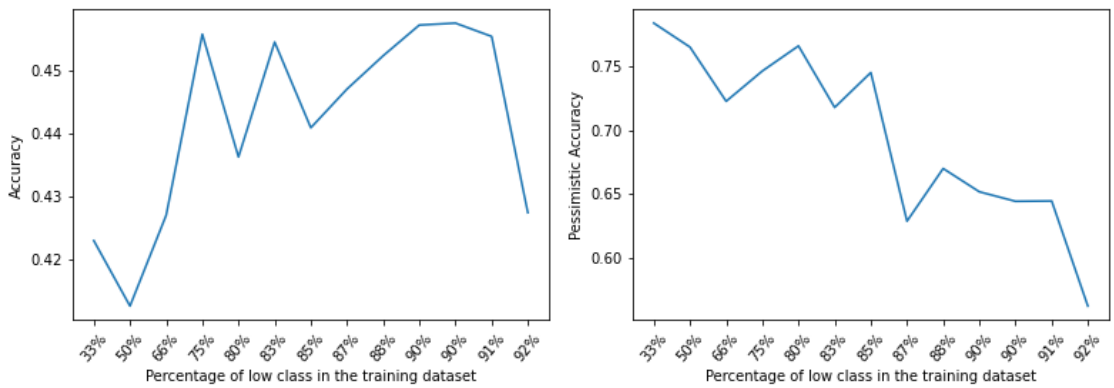
Aquest tipus de model és molt robust a overfitting i, contràriament als dos anteriors, conté no linearitats. Aquesta característica permet al model poder trobar relacions que en el Random Forest o XGBoosting no es podien. També a diferència d'aquest dos models aquest s'ha de dissenyar prèviament l'estructura que tindrà, i depenent d'això el rendiment millorarà o empitjorarà.

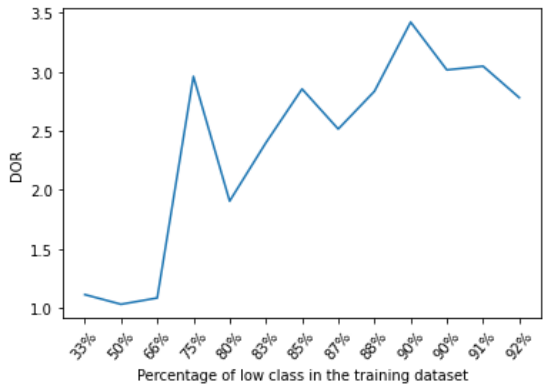
El disseny s'ha fet seguint el pairplot, veure apartat 2.2, on es podia veure com les variables entre elles estan poc correlades i les fronteres de decisió lineals no feien una bona feina. Per això s'ha decidit implementar una xarxa neuronal molt profunda, 7 capes ocultes, i cadascuna amb 20 neurones, representació a la *Il·lustració 20*.



Il·lustració 20. *Estructura de la ANN per la predicció de risc de dengue*

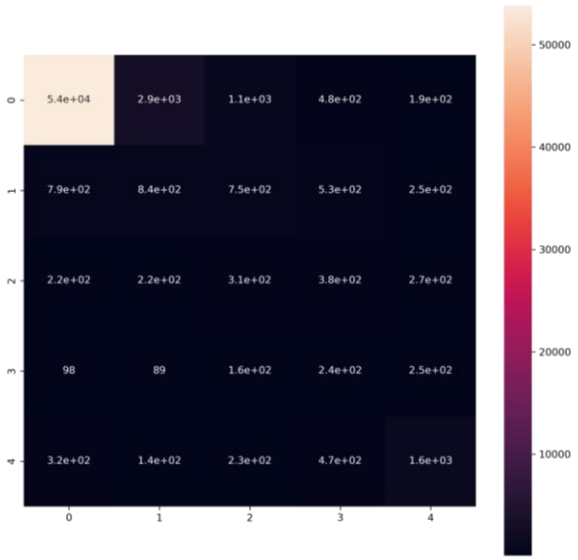
Igual que als models anteriors s'ha testejat exhaustivament per determinar el número de mostres de la classe mes baixa òptim per la regressió, veure *Il·lustració 21*.





Il·lustració 21. *Accuracy, Pessimistic Accuracy i DOR del model ANN segons el número de mostres de la classe Baixa*

S'observa com el DOR i l'accuracy creixen conjuntament quantes més dades de classe baixa hi posem a la base de dades d'entrenament, fins a un màxim al 90% de les dades. Aquest model entrenat és el que s'ha agafat per tal de fer la predicció. A la *Il·lustració 22*, es pot veure la seva matriu de confusió.



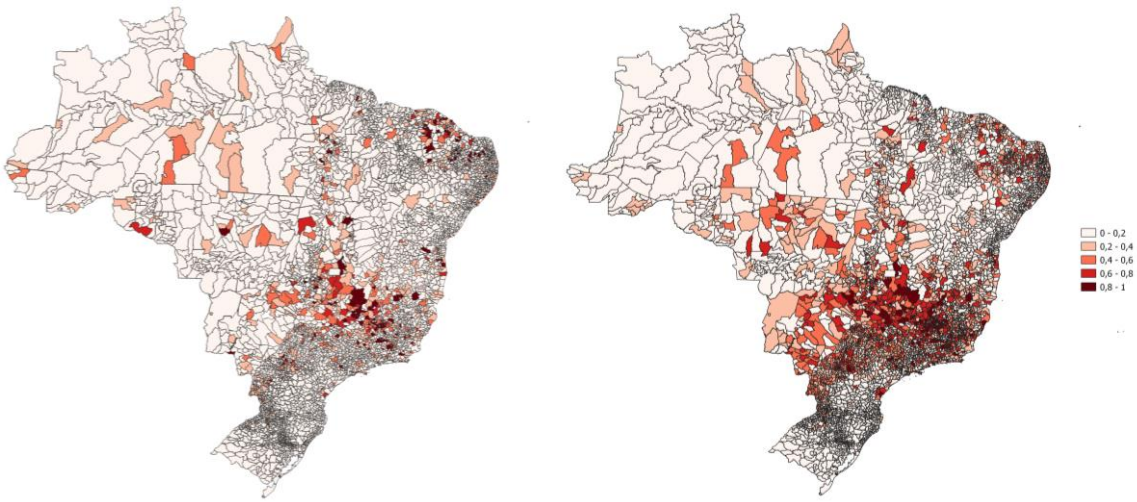
Il·lustració 22. *Matriu de confusió de l'Artificial Neural Network*

Aquest model té una bona predicció de nivells baixos i alts, encara que no ideal. El nivell 2 és l'únic nivell que no està ben caracteritzat però tendeix més a estimar-lo a un nivell més alt que no a un nivell més baix. A la *Taula 4* hi ha les seves característiques exactes.

***Taula 4**. Mesures de qualitat Artificial Neural Network*

Artificial Neural Network	
Accuracy	0.4572
Pessimistic Accuracy	0.6516
DOR	3.4202

Per poder il·lustrar-ho s'ha representat el risc real i el predit de maig de 2016, *Il·lustració 23*. Es veu com el riscos alts, juntament amb els baixos estan ben representats. Els nivells mitjos tendeixen a representar-se amb un nivell més alt. Aquest estimador és pessimista, ja que els nivells mitjos els prediu amb un nivell superior, però els nivells alts els tendeix a predir com a nivells alts i els nuclis de risc coincideixen amb els reals.



***Il·lustració 23.** Risc real vs Risc predit de la ANN*

4.3. Comparació de models

Segons els resultats obtinguts dels diferents models, el millor és el model de la ANN, seguidament el del Random Forest i després el Extreme Gradient Boosting.

Aquesta classificació concorda amb les imatges extretes de les prediccions, ja que la ANN és la que millor prediu les classes baixes i altes i a més era pessimista en les classes mitjanes, sent així un model el qual es podria aplicar en el territori. El Random Forest no té una estimació tan exacta però el ser pessimista, es podria considerar com la predicció del màxim de risc en el que un territori podria arribar durant el mes següent. Per part del XGB, l'estimació sol ser massa optimista i per tant té bastant falsos negatius que fan més qüestionable la seva utilització.

Comparant aquests models amb els implementats a l'inici del projecte, hi ha una millora a l'hora de predir els riscos alts. Els models anteriors caracteritzaven molt bé els nivells baixos, però per part dels nivells alts aquest no estaven gairebé representats. Aquesta millora permet predir una situació més realista.

5. Pressupost

Degut a la naturalesa del projecte aquest no té una gran quantitat de despeses. Aquestes es poden resumir en: cost humà i cost d'equipament.

El cost humà seran els salaris dels integrants del equip d'investigació, en el d'aquest projecte els salaris de l'enginyer i del supervisor del projecte, veure *Taula 5*.

Taula 5. *Pressupost relatiu al cost humà*

Posició	Persones	Compensació	Dedicació	Total Posició
Enginyer	1	10€/h	30h/setmana	6.000€
Supervisor	1	25€/h	2h/setmana	1.000€
			Total:	**7.000€**

El costs d'equipament serà el cost del programari i de l'ordinador en el qual s'ha fet la investigació, veure *Taula 6*.

Taula 6. *Pressupost relatiu al cost d'equipament*

Concepte	Unitats	Preu	Total Concepte
Llicència Excel	1	7€/mes	28€
Ordinador	1	900€	900€
		Total:	928€

S'han utilitzat més programes com el Google Colab y el QGIS, però aquests són gratuïts i per tant no afecten al cost del projecte, també les bases de dades utilitzades han estat proporcionades sense cost addicional o són bases de dades de lliure accés.

El preu total aproximat del projecte és de: **7.928€**

6. Conclusions i línies de futur

L'objectiu principal del projecte ha estat assolit: s'ha generat un model capaç de predir el dengue en Brasil amb el mínim nombre de falsos negatius i un número raonable de falsos positius. Aquest resultats per això no són els òptims i hi ha marge de millora.

La metodologia que s'ha seguit per l'elaboració de cada model ha estat la següent:

- Estudi del model a implementar.
- Preparació de les dades de la base de dades segons al model a implementar.
- Entrenament de diferents models.
- Selecció del millor model.
- Avaluació del millor model.
- Anàlisi del comportament del millor model.
- Representació de les dades i de les mesures de qualitat.

Aquest projecte també ha permès aprendre un conjunt d'eines actuals per a l'anàlisi i creació de models de Machine Learning i Deep Learning, com pot ser el Python amb les llibreries de: pytorch, numpy, seaborn, opencv, ... I eines per la representació espacial dels resultats, com el QGIS.

Per tal de poder millorar els models actuals s'ha fet una llista de possibles tasques a fer que es creu que podrien millorar els resultats dels models:

- Canviar la resolució temporal de la base de dades: en aquesta base de dades la resolució temporal és d'un mes, si aquesta es redueix a dues setmanes pot ser que els resultats milloressin degut a tenir més mostres de classe alta i mitjana.
- Canviar l'arquitectura de la ANN: aquesta arquitectura ha estat dissenyada a partir de la poca correlació entre variables i l'experimentació de diferents arquitectures. Pot ser que una altre arquitectura, ja sigui més complexa o més senzilla, donés millors resultats.
- Model de Recurrent Neural Network: aquest model no ha estat implementat però pot ser un model a tenir en compte ja que és un model amb memòria i per tant pot ser que ja que els casos semblen empitjorar en uns mesos de l'any, que aquest model pugui fer una bona predicció.
- Incorporar altres variables a la base de dades: noves variables podrien ser incorporades a la base de dades, tant climatològiques com sociològiques per veure el seu impacte en la predicció. Per exemple la humitat de l'aire.
- Diferent representació del risc a la base de dades d'imatges: la present representació del risc que es fa a la base de dades d'imatge disminueix dràsticament el nombre de mostres

per poder entrenar un model ANN. El risc es podria representar només per cada municipi i no tots a la vegada i així no es reduiria el nombre de mostres de la base de dades.

Bibliografia:

[1] Paho.org. 2021. *Dengue - OPS/OMS | Organización Panamericana de la Salud*. [online] Available at: https://www.paho.org/es/temas/dengue

[2] Datasus.saude.gov.br. 2021. *DATASUS – Ministério da Saúde*. [online] Available at: https://datasus.saude.gov.br

[3] Bauxell Cornet, J., 2021. *Machine Learning aplicat a la predicció del risc epidemiològic al Brasil*. [online] Upcommons.upc.edu. Available at: https://upcommons.upc.edu/handle/2117/344070

[4] Hdr.undp.org. 2021. *Human Development Report 2004*. [online] Available at: http://hdr.undp.org/en/content/human-development-report-2004

[5] Terra.nasa.gov. 2021. *MODIS Data | Terra*. [online] Available at: https://terra.nasa.gov/data/modis-data

[6] Brownlee, J., 2021. *Random Oversampling and Undersampling for Imbalanced Classification*. [online] Machine Learning Mastery. Available at: https://machinelearningmastery.com/random-oversampling-and-undersampling-for-imbalanced-classification/

[7] Brownlee, J., 2021. *Random Forest for Time Series Forecasting*. [online] Machine Learning Mastery. Available at: https://machinelearningmastery.com/random-forest-for-time-series-forecasting/

[8] Brownlee, J., 2021. *Extreme Gradient Boosting (XGBoost) Ensemble in Python*. [online] Machine Learning Mastery. Available at: https://machinelearningmastery.com/extreme-gradient-boosting-ensemble-in-python/

[9] Nielsen, M., 2021. *Neural Networks and Deep Learning*. [online] Neuralnetworksanddeeplearning.com. Available at: http://neuralnetworksanddeeplearning.com/chap1.html

[10] Brownlee, J., 2021. *How Do Convolutional Layers Work in Deep Learning Neural Networks?*. [online] Machine Learning Mastery. Available at: https://machinelearningmastery.com/convolutional-layers-for-deep-learning-neural-networks/

[11] En.wikipedia.org. 2021. *Diagnostic odds ratio - Wikipedia*. [online] Available at: https://en.wikipedia.org/wiki/Diagnostic_odds_ratio#:~:text=In%20medical%20testing%20with%20binary,does%20not%20have%20the%20disease.

[12] Alexlenail.me. 2021. *NN SVG*. [online] Available at: https://alexlenail.me/NN-SVG/