TEAM MEMBERS

María José García
Daniel Murillo
Paul Garay
Roberto Barrón

PREDICTING BREAST CANCER

with Machine Learning

https://github.com/MajoGarciaMontes/FINAL-PROJECT

CONTENT

01

Objective / main questions

02

Architecture of our ML solution

03

Data gathering and cleansing

04

ML model design

05

API integration

06

Visualization / Dashboard

07

Conclusions / Next steps

OBJECTIVE / MAIN QUESTIONS

CONTEXT

- *'A woman born today has about a 1 in 8 chance of being diagnosed with breast cancer at some time during her life'*
  National Cancer Institute
- *The costs of having more advanced tests done to determine whether a tumor is benign or malignant can go up exponentially, so not every patient will be able to have them done.*

OBJECTIVE

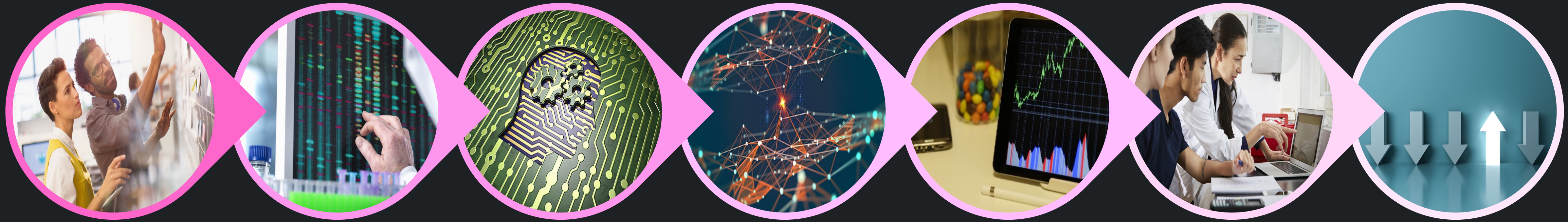
Develop a Machine learning (ML) model and train it with relevant breast cancer variables to predict the probability of developing breast cancer.

MAIN QUESTIONS TO SOLVE

•What classification fits the patient? (benign / malignant tumor).
•Which is the best model to use?
•What is the accuracy of the classification?

ARCHITECTURE OF OUR ML SOLUTION

During the last 6 months we have learned to create data solutions:



PROJECT PROPOSAL

ETL

ML DESIGN

API

DASHBOARD CODING

USER TESTING

PROJECT EVALUATION

DATA GATHERING / CLEANSING

GATHERING

- We were looking for a dataset with relevant information and with a significant size to train the data with ML.

WHERE?
University of Wisconsin (DS) repository

FORMAT
CSV file

SIZE
569 datapoints

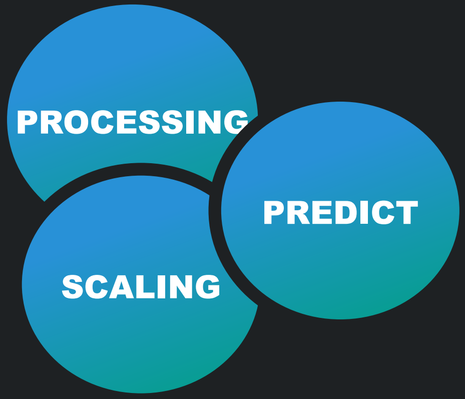
CONTENT
30 variables of characteristics of the tumor and the final diagnosis

CLEANSING

1) Cleansing for database creation
   - Drop of columns without data
   - Check for missing values

2) Cleansing for ML model
   - Pre-process of categorical data
   - Drop of unnecessary columns

ML MODEL DESING



LOGISTIC REGRESSION

	Precision	Recall	F1-score
BENIGN	0.97	0.97	0.97
MALIGNANT	0.96	0.94	0.95
Accuracy	0.97		

RANDOM FOREST

	Precision	Recall	F1-score
BENIGN	0.99	0.98	0.98
MALIGNANT	0.96	0.98	0.97
Accuracy	0.98		

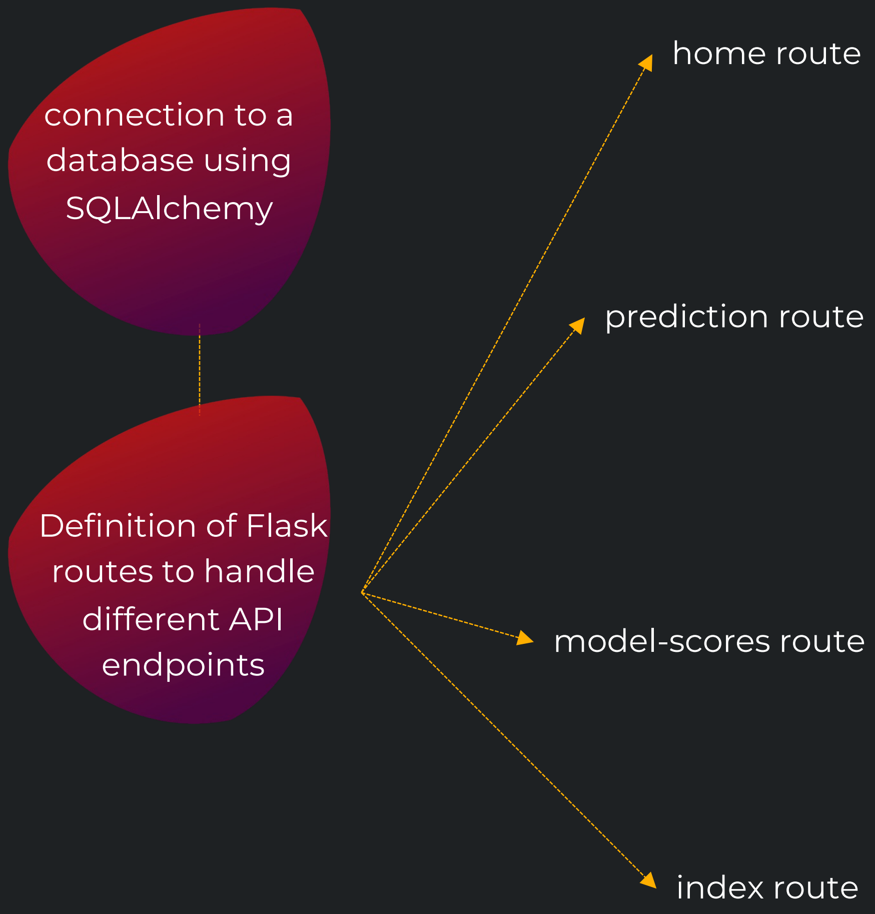
The model is **accurate** for both benign & malignant tumor. Its predictions are nearly always correct with high **precision** scores and the model correctly finds nearly all the true ‘malignant tumors’ as the **recall** scores were extremely high.

WHY IT IS IMPORTANT TO LOOK FOR THE HIGHEST PREDICTION RECALL?

The costs of mis-classifying a 'malignant tumor' as a 'benign tumor' are extremely high. It was not acceptable that this kind of tool could misdiagnose a patient with cancer as a 'healthy patient' and send them home without treatment.

API INTEGRATION

WHAT? Flask API that allows users to submit input data for a breast cancer prediction and returns the scores of the ML model.



connection to a database using SQLAlchemy

Definition of Flask routes to handle different API endpoints

home route

Shows all the routes and where they lead.

prediction route

Extracting input values from the request and used to make predictions with ML model.

The predicted result is rendered on a template.

model-scores route

Returns the evaluation scores of the breast cancer prediction model.

index route

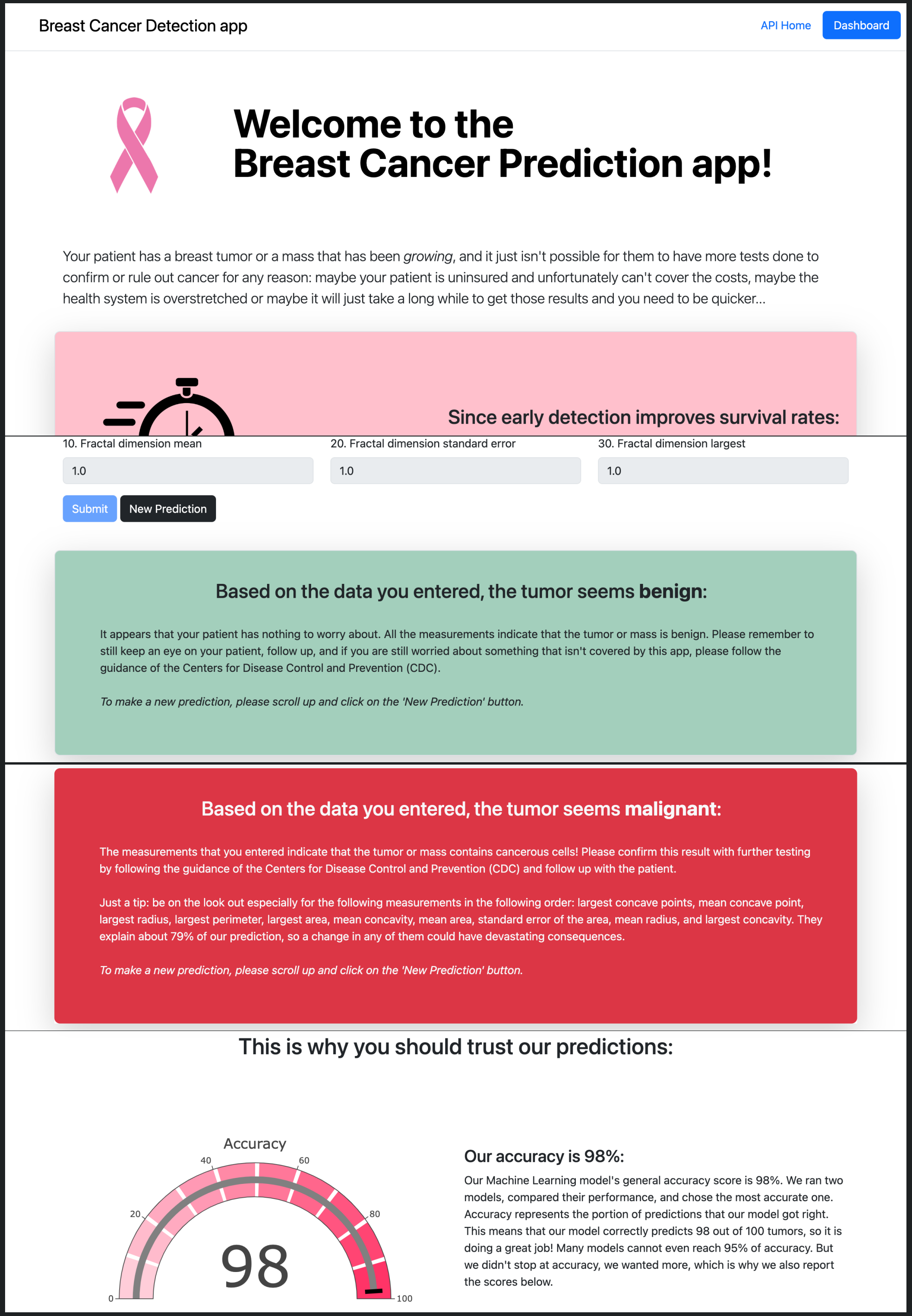
Returns an HTML file for visualization purposes.

VISUALIZATION / DASHBOARD

•**Friendly** and **simple**, empowered by a potent ML model.

•Allows users (doctors) to **make decisions** before recommending more expensive tests

- 30 inputs to fill out and compare the patient's tumor vs. prediction model
- Charts show how trustworthy our model is
- Prediction result and recommendation



Welcome to the
Breast Cancer Prediction app!

Your patient has a breast tumor or a mass that has been *growing*, and it just isn't possible for them to have more tests done to confirm or rule out cancer for any reason: maybe your patient is uninsured and unfortunately can't cover the costs, maybe the health system is overstretched or maybe it will just take a long while to get those results and you need to be quicker...

Since early detection improves survival rates:

10. Fractal dimension mean: 1.0

20. Fractal dimension standard error: 1.0

30. Fractal dimension largest: 1.0

Based on the data you entered, the tumor seems **benign**:

It appears that your patient has nothing to worry about. All the measurements indicate that the tumor or mass is benign. Please remember to still keep an eye on your patient, follow up, and if you are still worried about something that isn't covered by this app, please follow the guidance of the Centers for Disease Control and Prevention (CDC).

To make a new prediction, please scroll up and click on the 'New Prediction' button.

Based on the data you entered, the tumor seems **malignant**:

The measurements that you entered indicate that the tumor or mass contains cancerous cells! Please confirm this result with further testing by following the guidance of the Centers for Disease Control and Prevention (CDC) and follow up with the patient.

Just a tip: be on the look out especially for the following measurements in the following order: largest concave points, mean concave point, largest radius, largest perimeter, largest area, mean concavity, mean area, standard error of the area, mean radius, and largest concavity. They explain about 79% of our prediction, so a change in any of them could have devastating consequences.

To make a new prediction, please scroll up and click on the 'New Prediction' button.

This is why you should trust our predictions:

Our accuracy is 98%:

Our Machine Learning model's general accuracy score is 98%. We ran two models, compared their performance, and chose the most accurate one. Accuracy represents the portion of predictions that our model got right. This means that our model correctly predicts 98 out of 100 tumors, so it is doing a great job! Many models cannot even reach 95% of accuracy. But we didn't stop at accuracy, we wanted more, which is why we also report the scores below.

CONCLUSIONS / NEXT STEPS

- Our model was **successful** on learning with the provided dataset and developing a high level of prediction of breast cancer.

- The **random forest** model achieved the highest accuracy, precision, f1-scores and recall vs. logistic regression.

- ML has the potential to **reduce costs** while still being reliable for screening breast cancer.

NEXT STEPS

- Having **more recent data** could improve our precision and make our model be more in touch with today.

- Having **more feature variables** could improve the model scores or even work better with other ML models, such as a Neural Network.

ANY QUESTIONS?