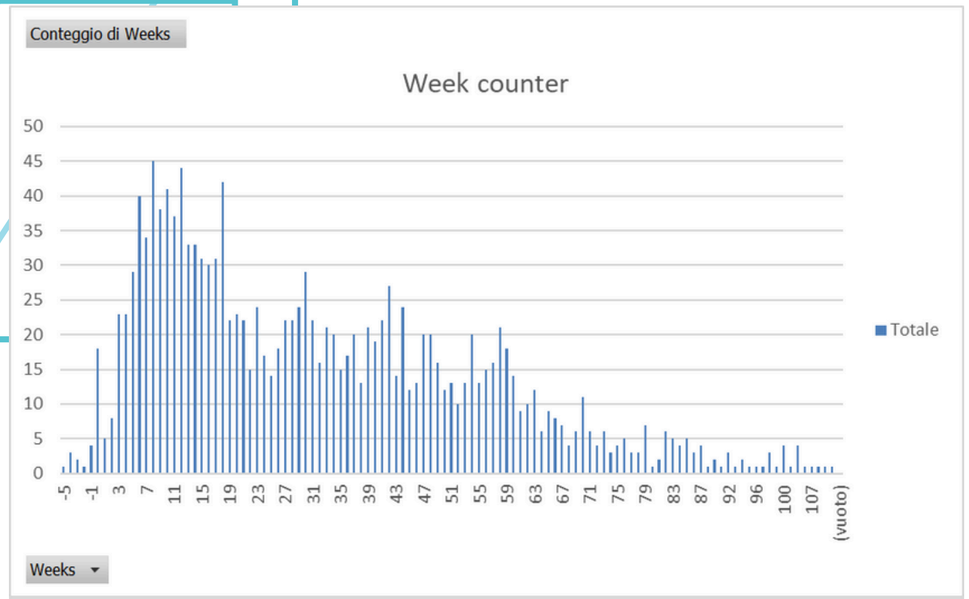
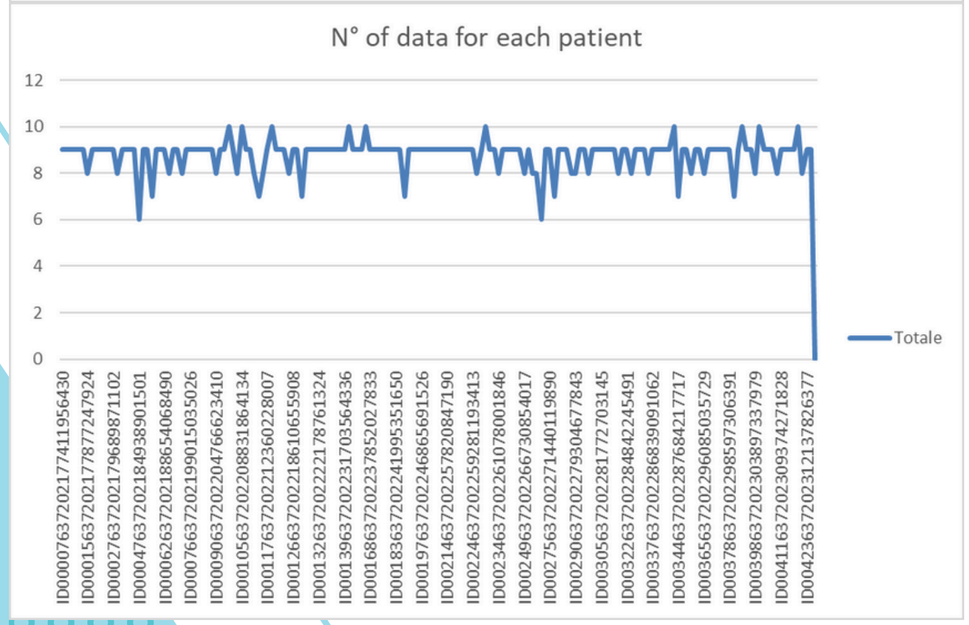
PROGRESS PREDICTION

3°WEEK

Data analysis

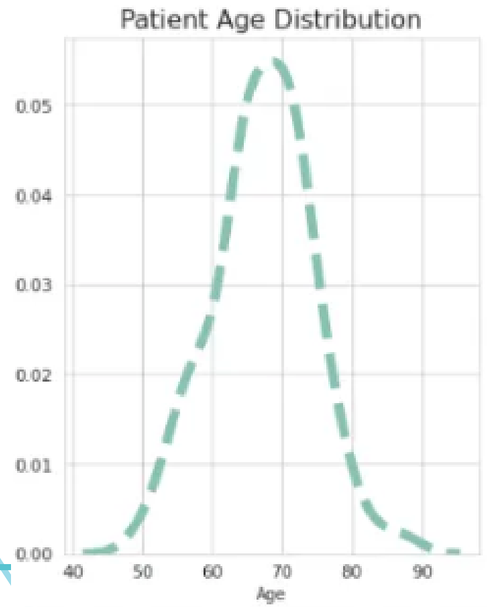


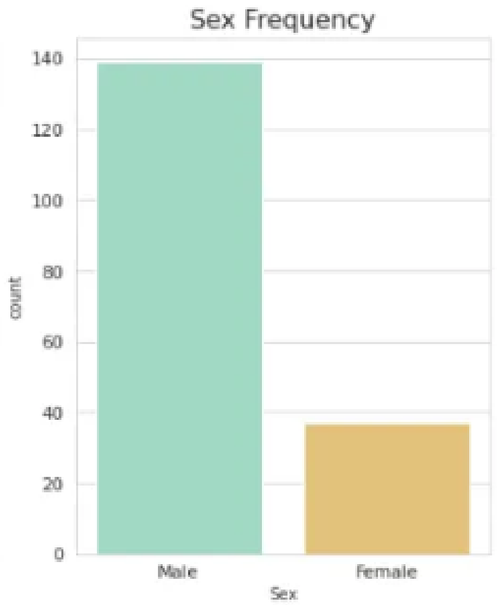


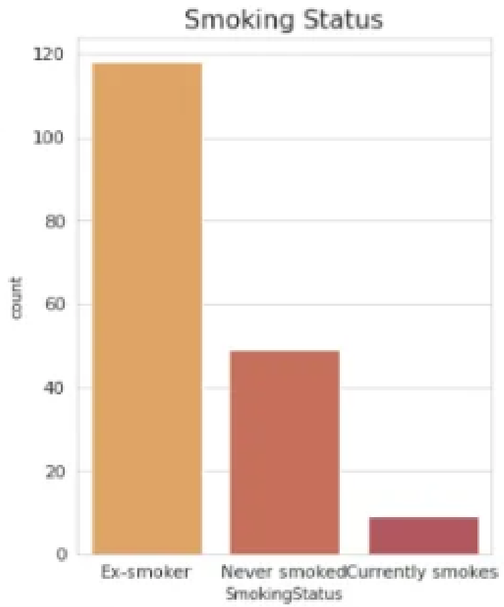
Week analysis:

- **18 patients have data at time 0**
- **Week 8 is the most present with 45 pairs**
- **Most patients have more or less same amount of data (176 distinct patients)**

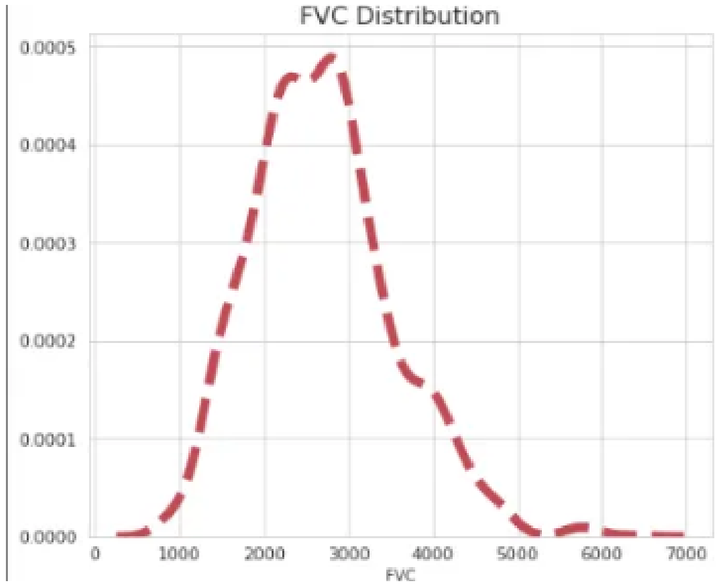
Data analysis

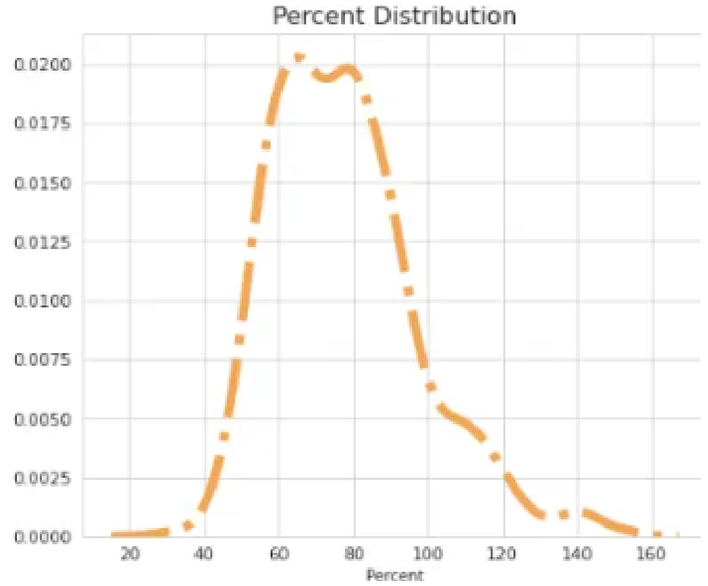




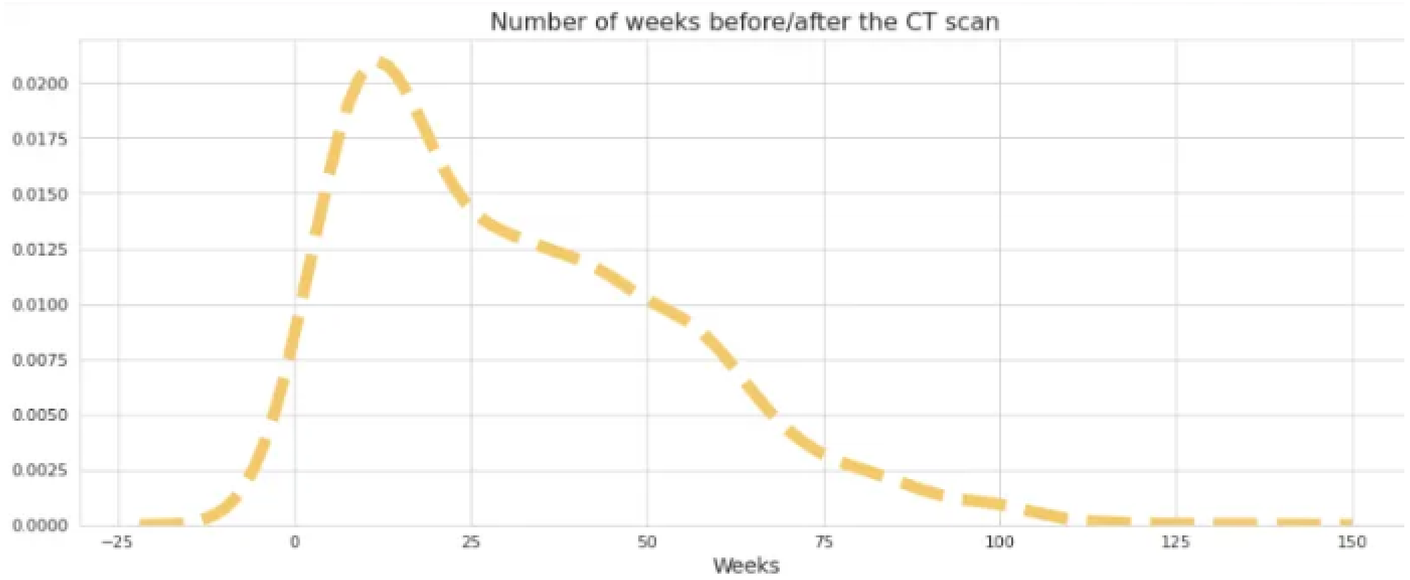


- **Patient age mean → 67 years**
- **Male most prevalent**
- **Ex- smoker most prevalent**

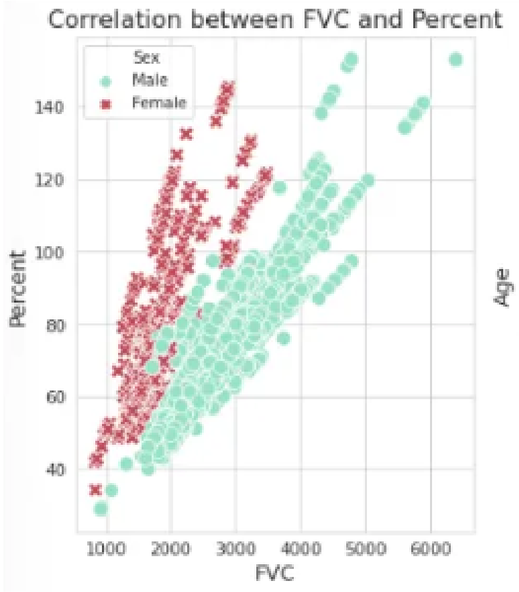


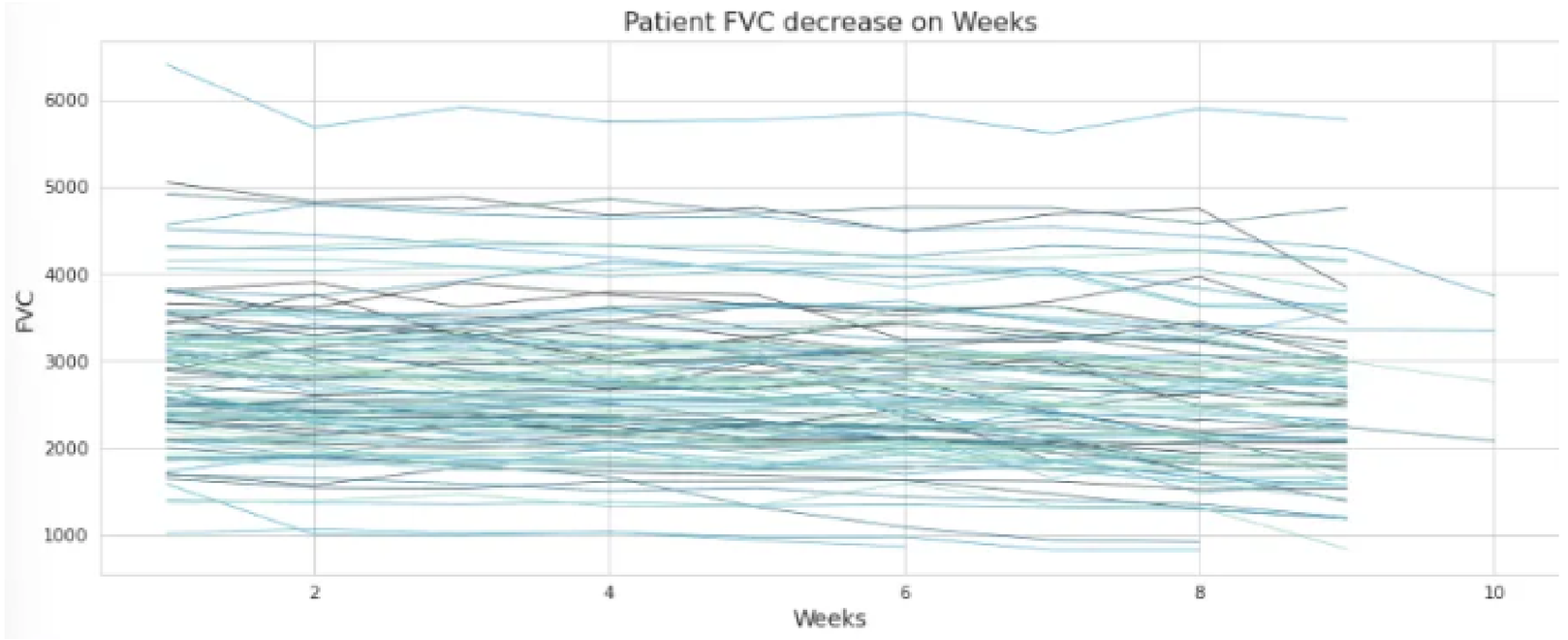


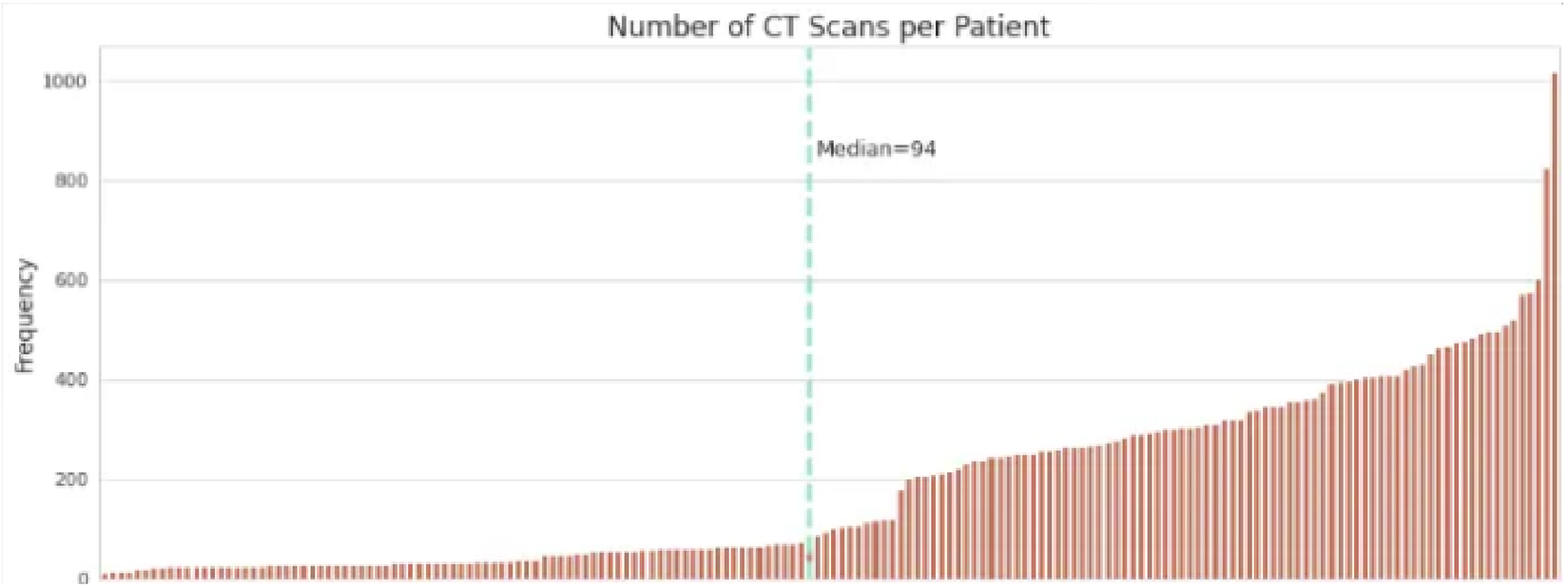
Data analysis



- **PEARSON CORR FVC X PERCENT: 0.672**
- **PEARSON CORR FVC X AGE: -0.09**
- **PEARSON CORR PERCENT X AGE: 0.096**




Patient FVC decrease on Weeks
FVC
6000
5000
4000
3000
2000
1000
2
4
6
8
10
Weeks


Number of CT Scans per Patient
Frequency
1000
800
600
400
200
0
Median=94

KAGGLE SOLUTION

11 images from the middle 30-60 percentile for each CT scan.
Using this the following features were extracted:

- **Lung volume**
- **Average of number of tissue pixels present across all selected images for a patient**
- **Average tissue area**
- **Average tissue volume**
- **Average of (number of tissue pixels)/(Total pixels on image)**
- **Average of (tissue area)/(lung area)**

Then merge image and metadata, clean outliers and normalize data.

KAGGLE BEST SOLUTION

Model ensemble

Linear Regression
CNN

+

Quantile Regression

KAGGLE BEST SOLUTION

1st model

- **Linear Regression → Take as input (Weeks,FVC) and determine a slope “a”**
- **Slope “a” → Indicates the progression through the weeks**

The slope is then used as a target for the pair (CT scans, data)

- **CNN (EfficientNet) → Take as input CT scan and data**

1st model

- K-fold cross-validation : 5 folds

Quantiles pre-defined for each fold as [0.8, 0.5, 0.1, 0.1, 0.1]

For each fold:
The model makes predictions for each patient, 20 CT slices are selected, for each slice a slope is determined and the quantile is used to aggregate the per-slice predictions.

The FVC at each week is predicted using the formula:

$$FVC = a^{*}w + B$$

(B is the intercept calculated based on the known FVC at week 0.

2nd model

- MLP

Takes as input only the available data + metadata.
Doesn't take in consideration features extracted from the CT scan.

Model ensemble

The two predictions are combined giving more importance to the 2nd model (0.6).

- 1st model : Gains complex information and patterns from the CT scans
- 2nd model: Stabilize the predictions

Final Output :

PatientID - Week - FVC - Confidence