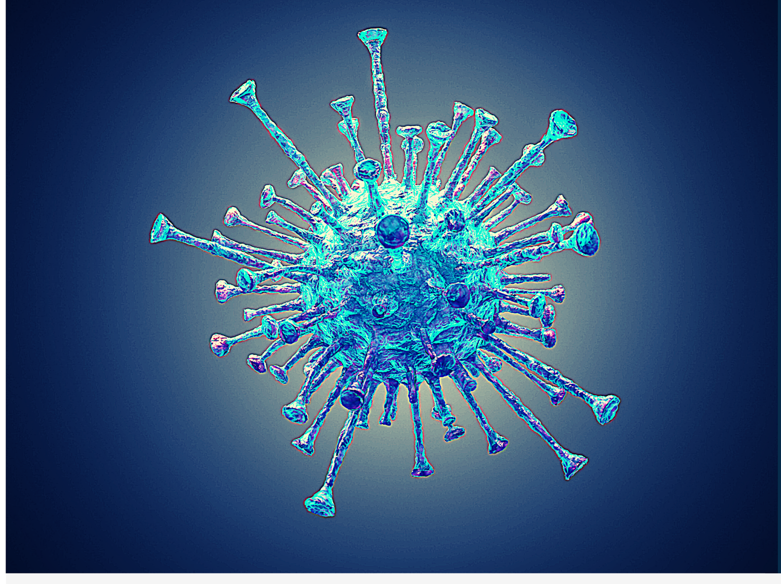


COVID-19 HOSPITALIZATION AND ICU PREDICTION

AN INTERACTIVE VISUALIZATION BASED ON PATIENTS' PRECONDITIONS

JAE PARK SONIYA GOYAL SIMON YEE
SALMAN AZMI SALIM KHAN YUNLIN QI

SUMMARY

COVID-19 infections have been on the rise since the beginning of this year, and there has been massive pressure on the **hospitals** and **medical facilities** to provide appropriate treatment to all **patients**. Since **Intensive Care Unit** (ICU) **beds** and **ventilators** aren't sufficient for all who require it, it has become extremely imperative to **predict** a **patient's need** for such medical facilities early on based on their **preconditions** and make them available to as many patients as possible. In this study, we will use some commonly known **models** and combine them with **visualization** of the data to understand how patients' preconditions can help us forecast the need for a hospital bed or ICU for them.

PROPOSED METHODS Intuition / Approach

PREPROCESSING / FEATURE SELECTION

Cleaning and **selecting** the appropriate **independent** variables ensuring models measure **real relationships** with highest **accuracy** possible, along with mitigating **bias risks**. Cleaning involved checking for instances where the **patient is male and is pregnant**, or for **expired dates** (not valid) and **data abnormally large** or coded as **not applicable**. Feature selection was achieved using **Stepwise Regression** and **Cross Validation** on the cleaned dataset.

MODEL

Careful selection of **two** machine learning algorithms for our predictive analysis: a **logistic regression model (LR)** employed as a baseline, and a **multilayer standard neural network model (SNN)** employed to better capture the nonlinear and complex relationship between various **patient's preconditions** (features) and outcomes (**ICU** or **hospitalization**). Extensive **model validations** across different analyzing tools (**TensorFlow, Scikit-learn, Weka, and Azure ML Studio**) ensuring **model fit**, **hyperparameter tuning** and metrics evaluation, **investigating** and **minimizing** the impact of an **imbalanced dataset** with **downsampling**, **oversampling** and **initial weights exploration**.

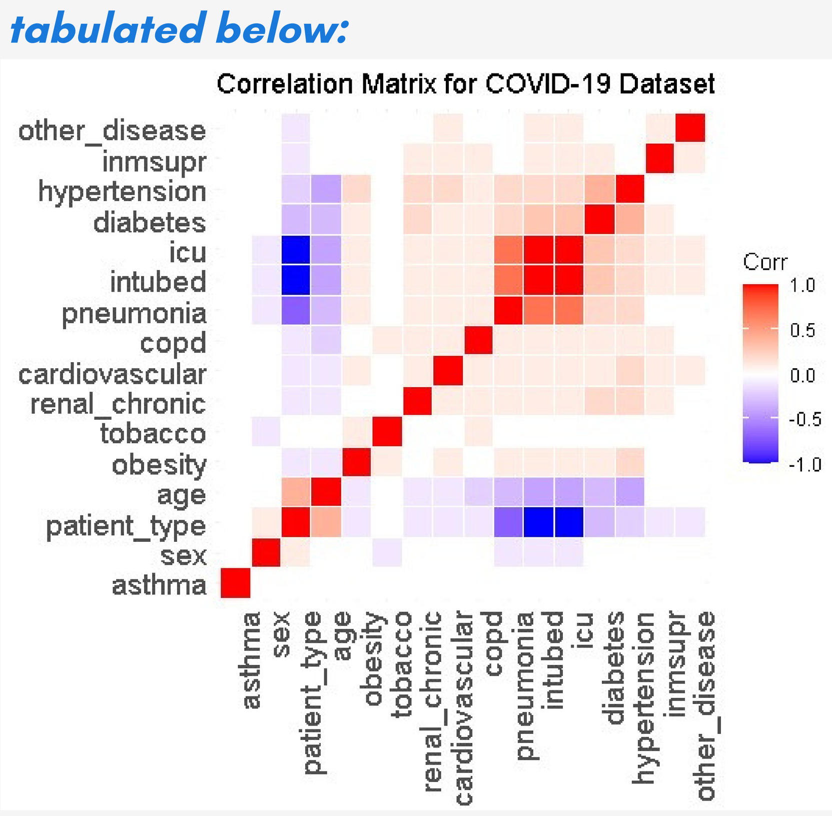
VISUALIZATION

Visualizing the **distribution** of **data** and results from our **model** on the train and test data, and relating them to the **preconditions** that played a **significant** role in determining the **patient's results**.

- **Interactive visualizations** enabling the user to view the distribution of the data across any feature or **combination of features**
- **ability** to **remove a feature** and see how the model behavior changes
- **visualizing the coefficients of the features** and understanding which user health condition **contributes to** determining if a **patient needs hospitalization or ICU**
- **visualizing**, given a **test set of data** and potentially from **user input**, **how many hospital and ICU beds** are needed for that dataset of patients

EXPERIMENTS AND RESULTS Feature Selection

Evaluation of variable selection methods through building two models. The **cross validation (CV)** model used all the predictors, and predictors having a **statistical significance of 0.05** were marked. We checked for **multicollinearity (variance inflation factor (VIF))** and compared the **variable performance** against a **step regression model** (SR). ***Results are tabulated below:***



	ICU - CV	ICU - SR	Hospitalization - CV	Hospitalization - SR
sex		✓	✓	✓
pneumonia	✓	✓	✓	✓
age	✓	✓	✓	✓
pregnancy				
diabetes	✓	✓	✓	✓
copd			✓	✓
asthma	✓		✓	✓
inmsupr			✓	✓
hypertension			✓	✓
other_disease			✓	✓
cardiovascular		✓		
obesity			✓	✓
renal_chronic	✓	✓	✓	✓
tobacco			✓	✓

Cross referencing our results with **medical research** tells us:

- **Immunocompromised (inmsupr)** and **obesity** not being selected in **ICU** model is probably the biggest surprise, given **CDC's serious warning**. Upon reviewing our data, this has to do with the skewness - only about **3.88% of individuals** in our data are in **ICU** for **COVID-19** and are **immunocompromised**
- **Cardiovascular** and **Hypertension** seem to be **correlated** and hence **cardiovascular** gets dropped from the **Hospitalization model**
- **Renal chronic** (kidney related issues) appears to only have an affect on individuals who have serious existing conditions already for the **ICU model**
- **Pregnancy** is likely affected by **skewness** in data and **collinearity** with **sex**. Hence was **removed** from the final feature selection

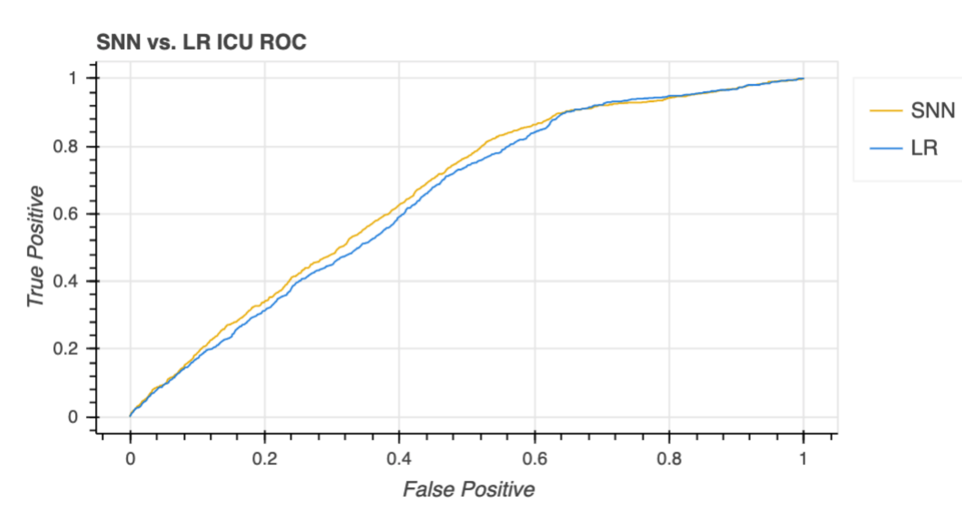
COVID-19 PATIENT PRECONDITION DATA

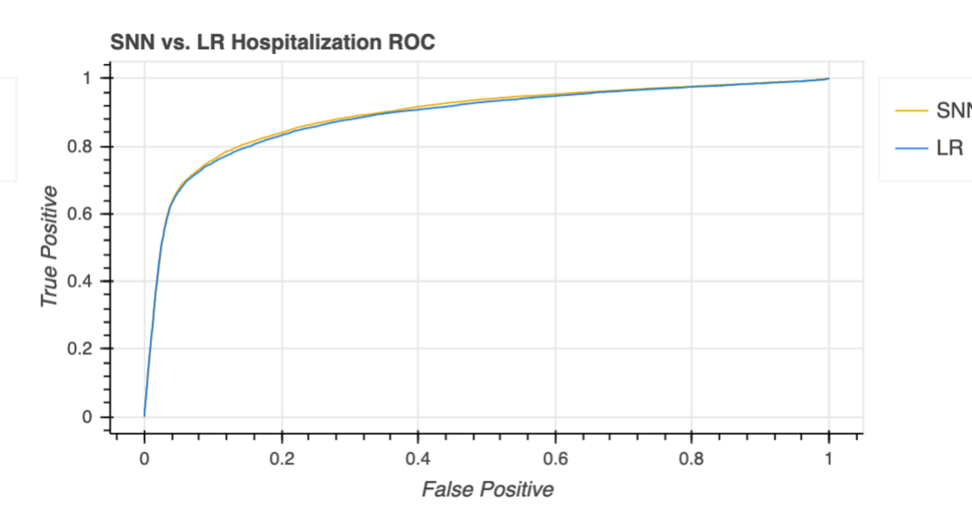
Multi-feature dataset provided by the **Mexican Government** and **downloaded** from **Kaggle**. The dataset does not contain any Protected Health Information (PHI) of the patient, and a unique random id is assigned to every row.

SIZE: 44.52 MB
rows : 563201
columns : 23
duration : Jan '20 - Jun '20

EXPERIMENTS AND RESULTS Predictive Model

Our primary models included **LR** and **SNN**. The data was split into **train (70%)**, **validation (15%)** and **test (15%)**. Both datasets (**ICU** & **Hospitalization**) were **imbalanced**. To address that, we first implemented **oversampling** & **undersampling** on train dataset in addition to **custom class weights** for positive labels, so that our model pays more attention to samples from an under-represented label. Secondly, our experiments also addressed impact of various **hyperparameters** (**batch size, epochs, learning rate** and **momentum**) on result metrics like **AUC**. Lastly, we also investigated how **classification threshold** affected the prediction results.

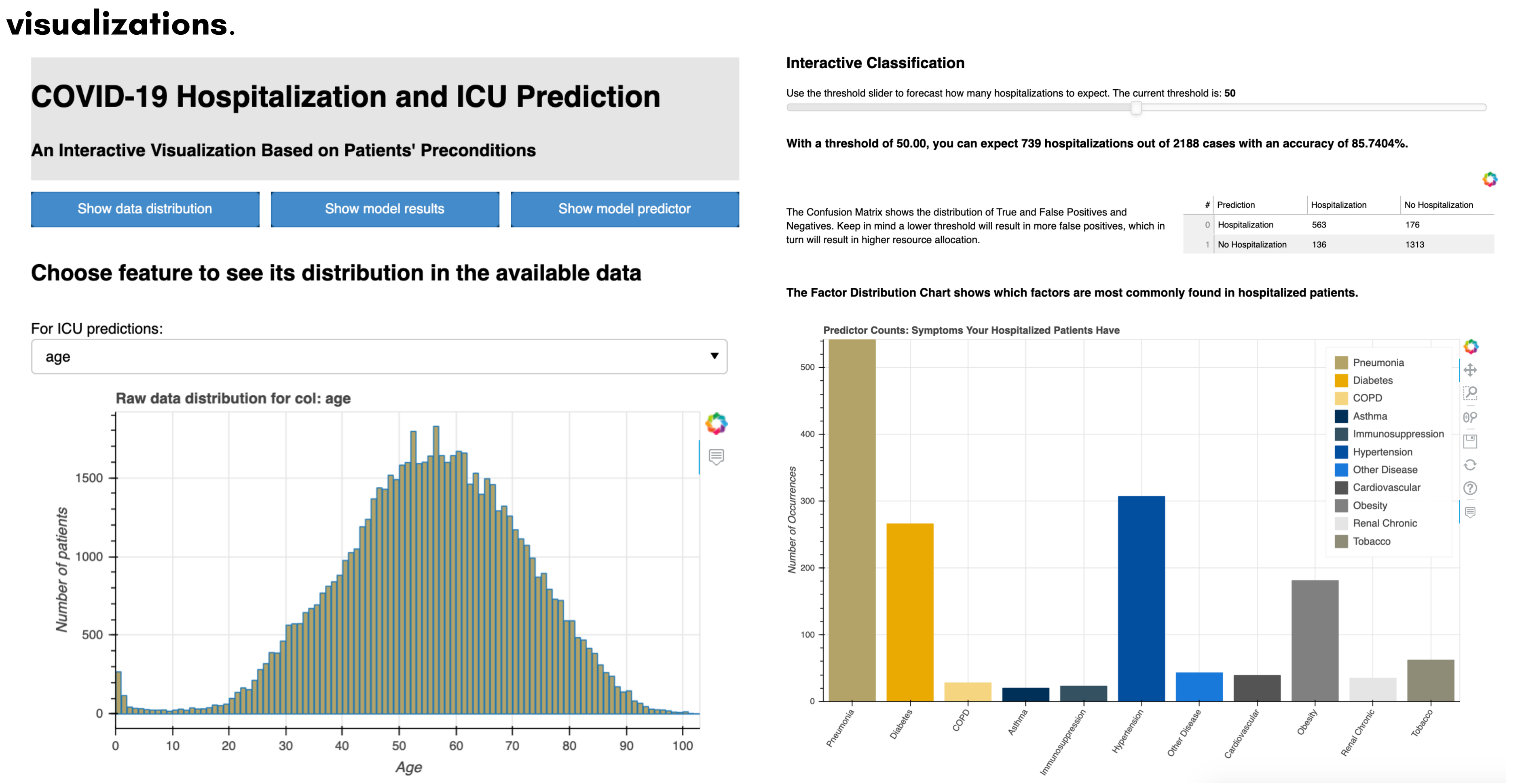




- For **ICU prediction**, **low accuracy** and around **0.65 AUC** performance were observed using strongest factors: **age, pneumonia, diabetes** and **renal chronic**. Additional pre-excluded features (except intubation) did not help with model improvement. Neither did the sampling strategies. For **hospitalization** prediction, on the other hand, **0.89 AUC** was achieved along with a **85% classification accuracy** on test data at **0.5 classification threshold**
- In **conclusion**, **no selected factors** appeared to perform strongly in predicting **ICU** need for a COVID-19 positive individual. In contrast, **strong** indicators (features) were identified for **hospitalization** prediction: in terms of **odds ratio** (**OR**, i.e., the odds that an outcome will occur with a particular feature compared to that without it), **pneumonia** predominated with **OR=29.93**, followed by **chronic renal (OR=2.34)**, **age (OR=1.9)**, **diabetes (OR=1.75)** and **immunosuppression (OR=1.74)**
- Compared to other methods in current studies, our approach addressed more on the existing **limitations** such as **lack of model validation**, **arbitrary data selection**, and **arbitrary thresholds**. As a result, our methods achieved **better prediction** results in terms of **AUC** and **accuracy**, and **less misconception** & **bias risks** in the outcome. Plus, our methods also exported prediction results for interactive visualization on web server

EXPERIMENTS AND RESULTS Visualizations

Experiments involved **evaluating tools** and **packages** for visualization based on requirements. We designed our component in **Bokeh** due to its **native support** for **interactive** and **customizable visualizations**.



Final website has **3 Interactive Tabs**. Dataset with a predefined structure can be loaded interactively.

1. allows for **interactive feature visualization** across the dataset along with a **correlation plot**
2. focuses on **model output analysis** with **AUC, precision, loss, recall**, and **ROC**
3. includes **prediction** data and **visualization** that focuses on **key features contributing** towards and **estimated counts** of **ICU** and **Hospitalizations**

WEBSITE @ *https://tinyurl.com/covid-visualisation*

Due to time constraints, we removed the patient interface for prediction component and opted to allow users to visualize needs using a test set of data. Adding and removing features from the visualization was also not achieved due to the same reason.