ANONYMIZED COVID PATIENTS' OUTCOME ANALYSIS & PREDICTION

Final Product: https://cse-6242-team12.herokuapp.com/

Team 012:

Ali Naji - Andy Senoaji - Muhammad Omer Farooq - Orhun Aydin - Smitkumar Contractor

Introduction, Motivation, & Goal

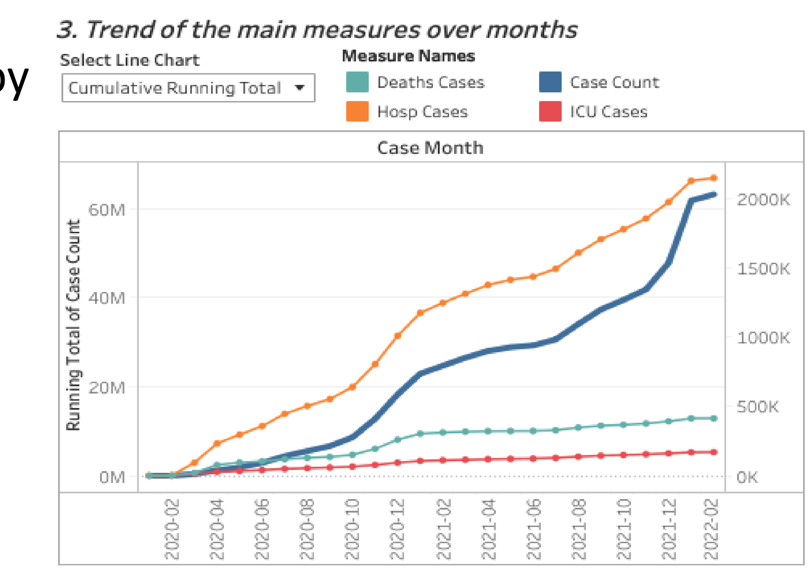
- Globally, the COVID-19 pandemic has infected 78.8M people in the US, claimed 946K fatalities, and changed virtually changed everyone's lives. The virus is indiscriminate, but it expose and affects differently among socio-demographics segments.
- Despite prior studies' focus on forecasting the epidemiological outcomes using empirical and data-driven models, a general statistical that incorporates societal variables is missing.
- These analytic techniques have also successfully helped determine the impact of factors like symptoms and demographics on the spread of the disease.
- We are motivated to explore the impact of powerful descriptive & predictive analytics on the well-being and safety of the population during the pandemic.
- The impact of this work is modeling and visualizing the impact of COVID-19 on disproportionately exposed groups in the US. The project answers "Given a patient's characteristics such as age, location, and gender, what are the expected health outcome given prior data?"

Data

Our main dataset is "COVID-19 Case Surveillance Public Use Data with Geography" published by CDC, updated monthly. We downloaded the data up to February '22. To enhance the main dataset, we use "COVID-19 Vaccinations in the United States" also published by CDC and downloaded from CDC.

Main dataset: **63M rows and 19 fields – 9 GB – 3 targets: deaths, ICU admits, Hospitalizations.**

Enhancement dataset: **1.6M rows and 66 fields – 690 MB**



Approaches & Challenges

1. Data Handling:

- The CDC datasets are very large.
- The two data sets have varying spatial scale and granularity. We aggregate the case data to gain same granularity and spatial scale.
- Three target outcomes instead of just 1 with missing values. We approach this with 2 rounds of preprocessing:
- Dropping rows with missing targets and features
- Aggregate time attribute and county attribute, use 1-hot encoding categorical features.

2. Modelling Approach

- With 3 target outcomes, we developed 2 models, that can be nested together:
- Single outcome model that predicts the most probable combination of outcomes of a patient
- Multi-outcome that predicts the probability of each outcome occurring to the patient.

3. Visualization Approach

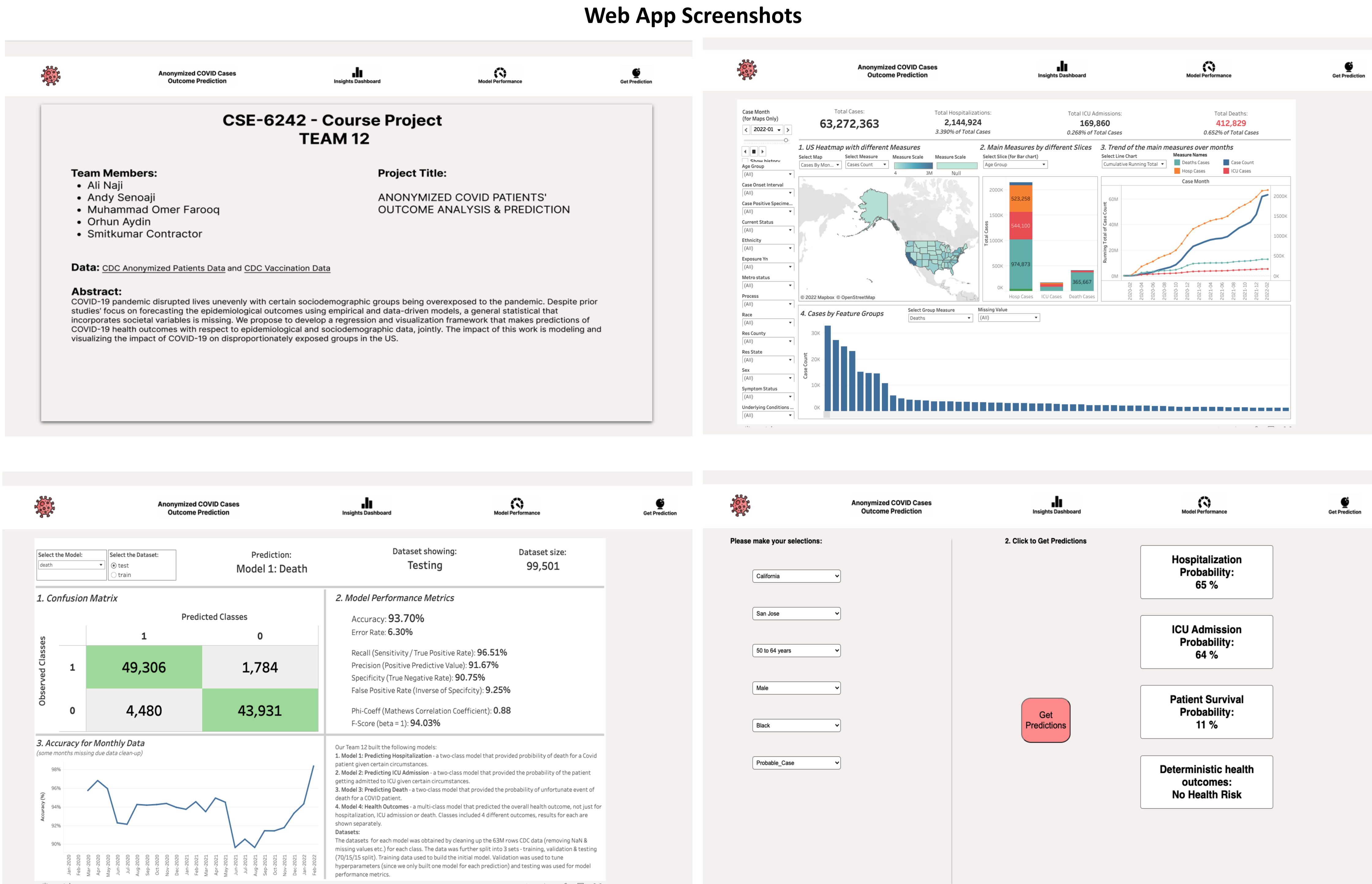
Leveraged Tableau & Public Server for publishing.

- Tableau for descriptive visualizations.
  Data is aggregated to 6.8M rows to meet Tableau Public limitation and keep query time manageable
- 2 dashboards – descriptive insights & ML Models performance.

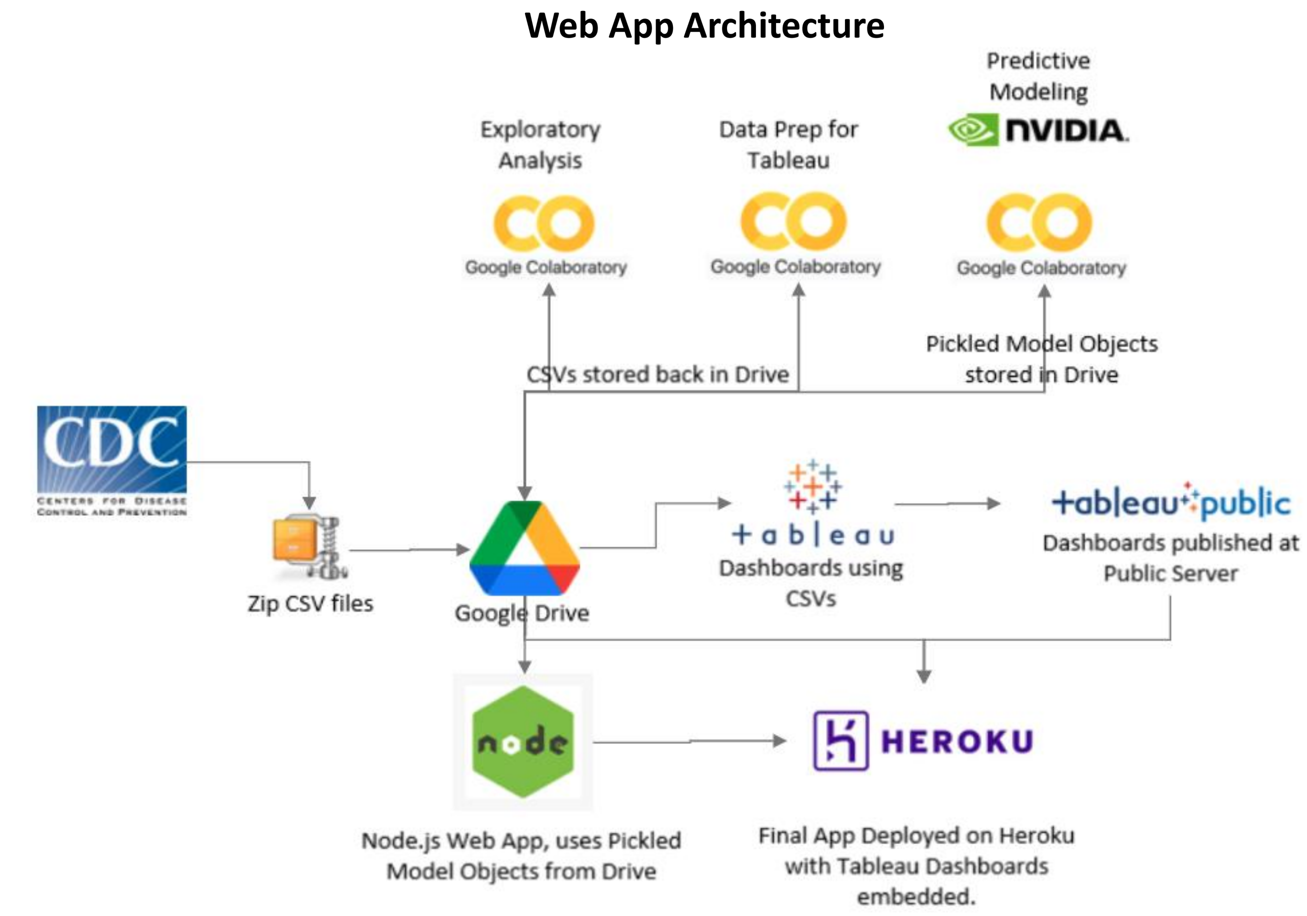
4. Dissemination

- Build an easy use node.js web application that was deployed on Heroku Server.
- An app that allows for easy packaging of Tableau dashboards (embedded from Tableau Public Server) and ML models for real-time single predictions.

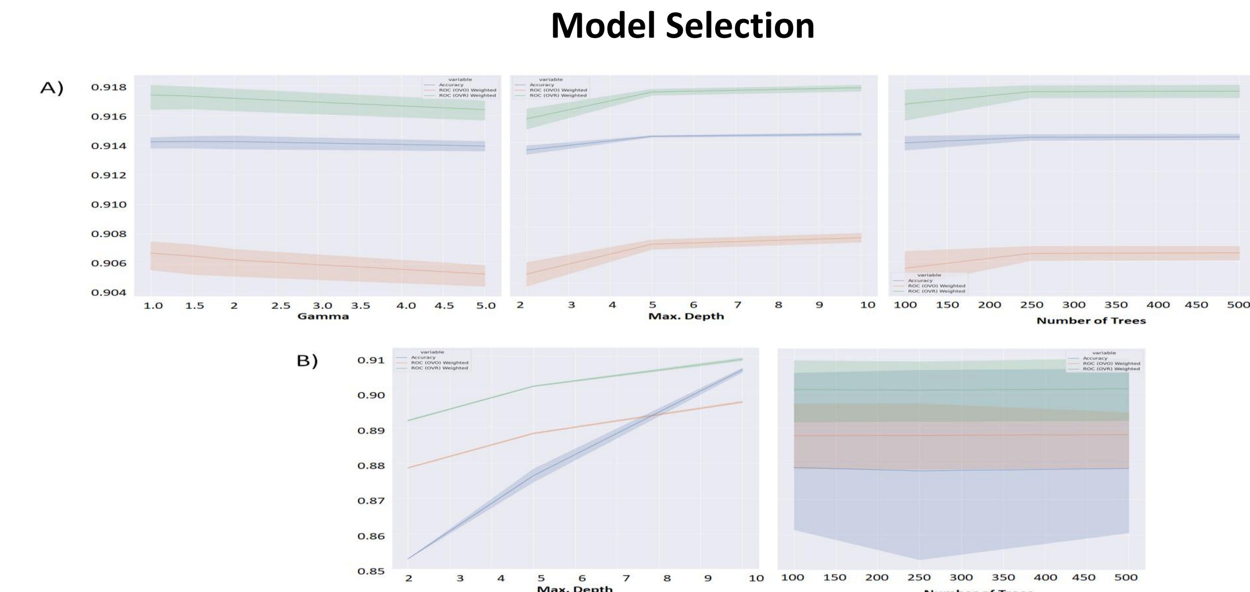
Web App Screenshots



Web App Architecture



Model Selection



Experiment and Results

1. Data Exploration

- Tools experiment: Used MySQL to perform EDA, but switched to Google Colab + (paid) for further analysis and modelling.

2. Data Aggregation

- Perform aggregation to reduce data size without information lost. Experimented on data clean-up, group-by case & case count aggregation, and check for data losses. Result is 6.8M rows (~89% reduction)

3. Tableau Dashboards Development

- Dashboard performance with 6.8M rows of data was a concern. We designed and built it slowly and monitored performance. The result is a dashboard with less than 3 clicks to get insight with reasonable user experience.

4. Visualization & Model Interaction

- Built a full working webapp prototype in Figma, experimenting webapp layout . By prototyping we tested multiple layouts for both descriptive visualization page and model interactions

5. Machine Learning Modeling

- Multi-tiered model for probabilistic and deterministic health outcomes
- Random forest found to perform best for seasonal model
- Boosted Decision Trees found to perform best for the deterministic model

Conclusion

- Decision-tree based methods struck the best balance between explain-ability and accuracy.
- Prediction metrics used allowed to fine tune models to capture rare health outcomes.
- With both 3-probability and distinct outcomes models, severe patient outcomes were highly predictable.
- Random forest was the best choice for the 3-probability model, whereas boosted decision trees provided highest accuracy for the distinct outcome model.
- The model performance dashboard helped the team understand model limitations, analyze metrics, and suggest new improvements.
- The final web application nicely packaged all components; dashboards and picked models to enable the user to truly explore the COVID data and get individual COVID case predictions.

ML Models Performance Metrics

Predicted Outcome	Accuracy	Recall	Precision	Specificity	Phi-Coef.	F-Score
Model 1 - Death	**93.7%**	96.5%	91.7%	90.8%	0.88	94.0%
Model 2 - ICU Admission	**82.8%**	85.5%	82.4%	79.7%	0.65	83.9%
Model 3 - Hospitalization	**77.5%**	76.4%	78.1%	78.6%	0.55	77.3%
Model 4 - No Health Risk	**92.4%**	99.6%	92.6%	14.3%	0.31	96.0%
Model 4 - Hosp No ICU	**94.0%**	12.5%	61.2%	99.5%	0.26	20.8%
Model 4 - Hosp w/ ICU	**99.6%**	0.01%	17.7%	100%	0.00	0.02%
Model 4 - Risk of Death	**98.1%**	6.1%	37.9%	99.8%	0.15	10.6%