Prediction of ICU Admission Among Hospitalized Patients with COVID-19

IT5006 Group 2 Project Presentation

Felicia, Gordon, Ansel

Agenda

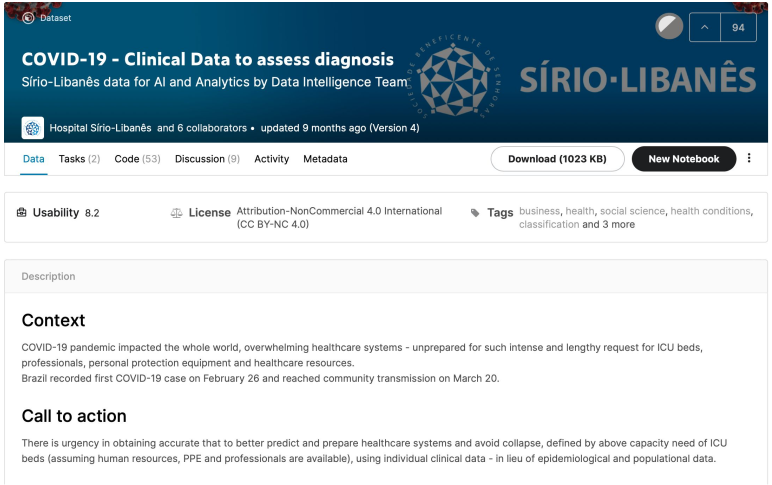
1. Overview
    - Motivation, dataset description and problem statement
2. Data cleaning
3. Exploratory data analysis (EDA)
    - PCA, Data visualization of features
4. Choice of models and analysis
    - Logistic regression
    - Decision Tree
    - SVM
    - Perceptron
5. Next Steps
    - SMOTE, Stacking

Motivation

- Huge burden of COVID 19 pandemic

- Differential clinical trajectory of COVID-19
  - Some patients require intensive care admission

- Predict ICU admission among hospitalized COVID-19 patients
  - Manpower allocation for sicker patients



Dataset description

- Anonymized Kaggle dataset - 385 hospitalized COVID-19 patients
- Each patient is observed over 5 time windows

What's Tracked in Each Time Window?

- Patient's level of care (ICU vs non-ICU)

- Total of 54 unique clinical, physiological, and laboratory parameters in each time window

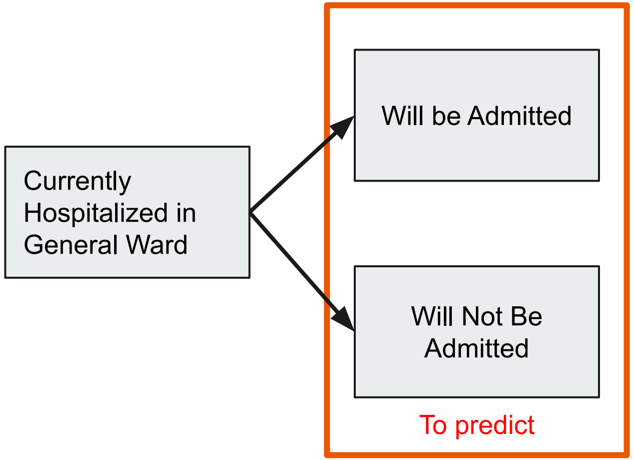
- 12 categorical feature variables - gender, age group, disease groups, hypertension, immunocompromised state

- 42 numeric feature variables - summary statistics for each feature variable.
  - Examples:
    - vital signs such as blood pressure and respiratory rate
    - laboratory parameters such as individual components of the full blood count
  - Numeric data preprocessed with Min Max Scaler for data privacy

Problem statement

If my patient has been admitted to hospital with COVID-19, what is the likelihood that he'll eventually be admitted to ICU?"

- Select Predictor Variables To Predict ICU Admission
- Implications for resource management and early intervention



Data cleaning

1. Engineered Target Variable

2. Removed patients who were already in ICU

3. Discarded time windows that occurred after ICU admission

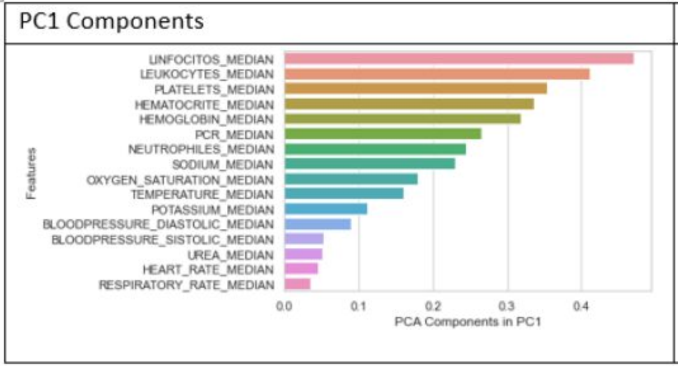
4. Outlier removal
   - Consistent Application of 1.5 IQR rule
   - Conditional: Did Not Apply 1.5 IQR rule for Variables With Very Narrow Distributions

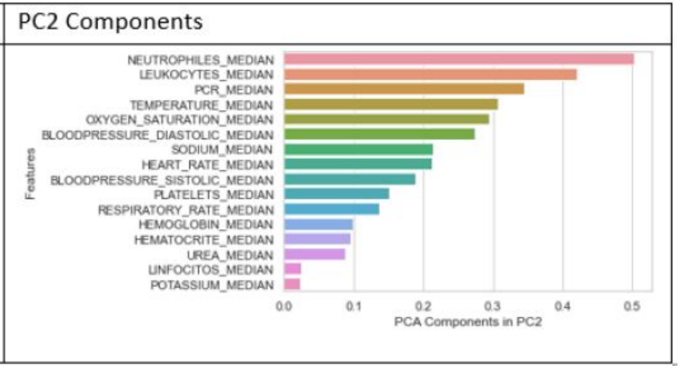
5. Missing values
   - Imputation using median

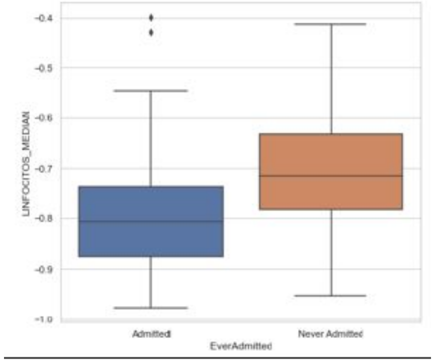
Principal Component Analysis (PCA)

Feature discovery for further exploratory data analysis

- PCA on 42 numeric feature variables
- Avoided multicollinearity - considered median summary statistic for each feature variable
- Attempted with and without outlier removal





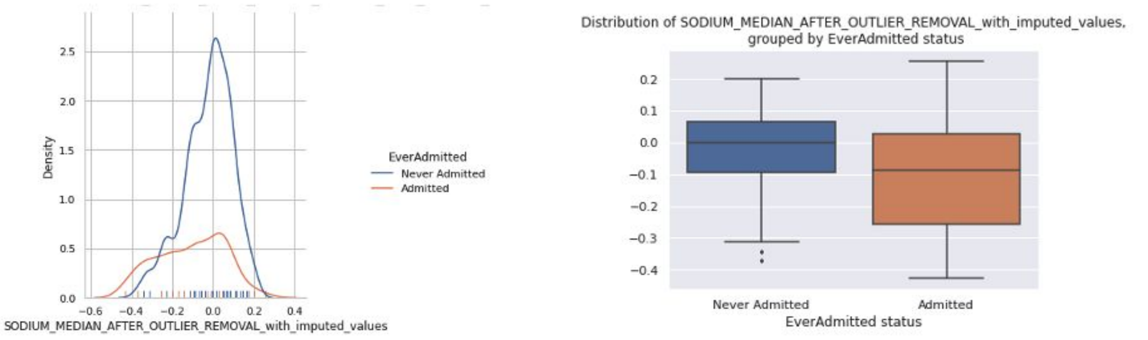


Data visualization identified additional features

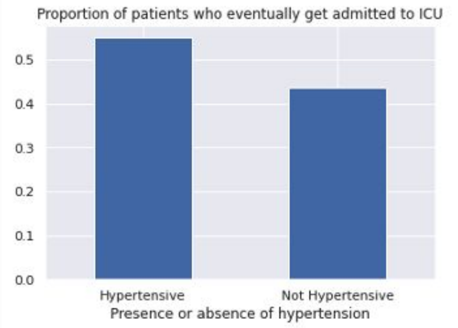
Boxplots, KDE plots and bar graphs

- Shortlisted a total of 18 potential numeric feature variables and 3 categorical feature variables

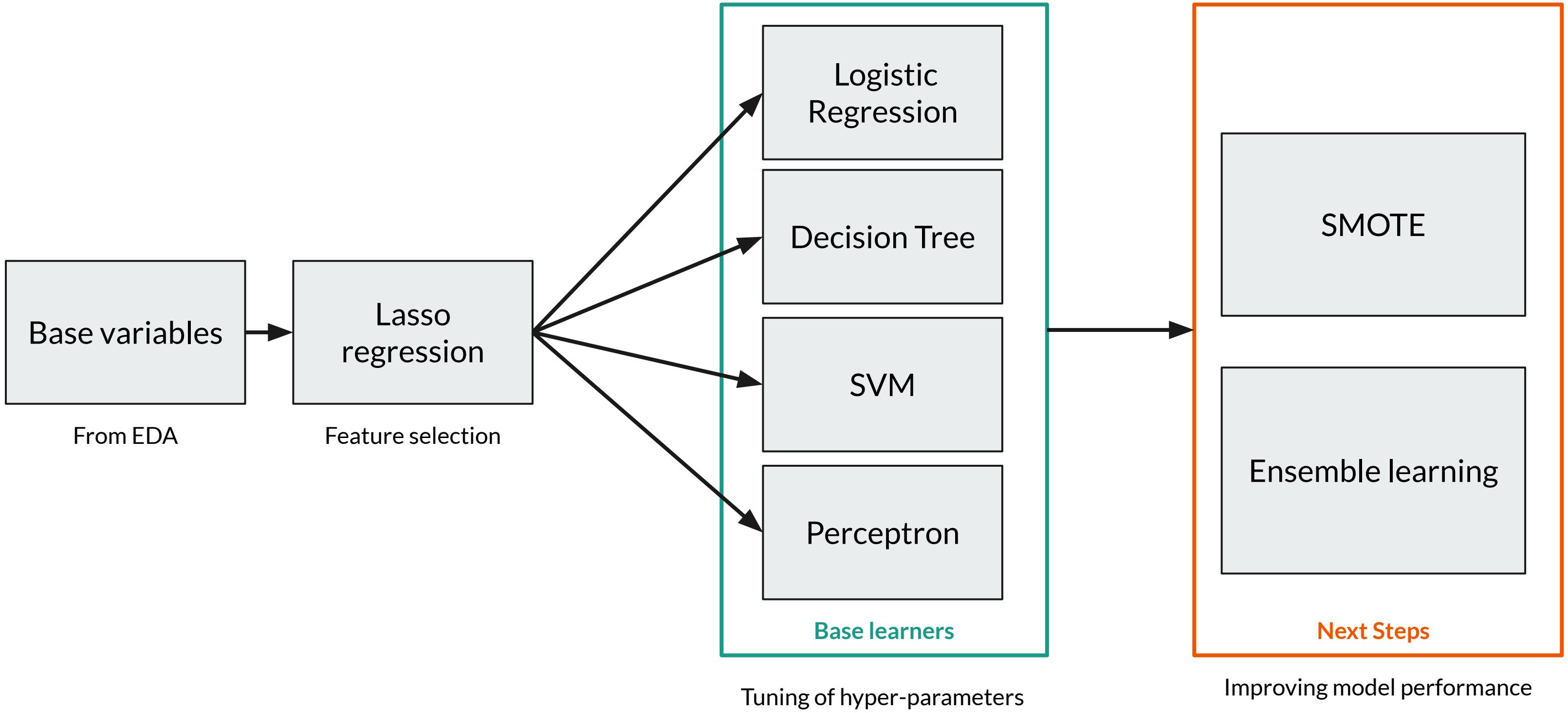
KDE and Boxplot of SODIUM_MEDIAN



Bar graph of HTN

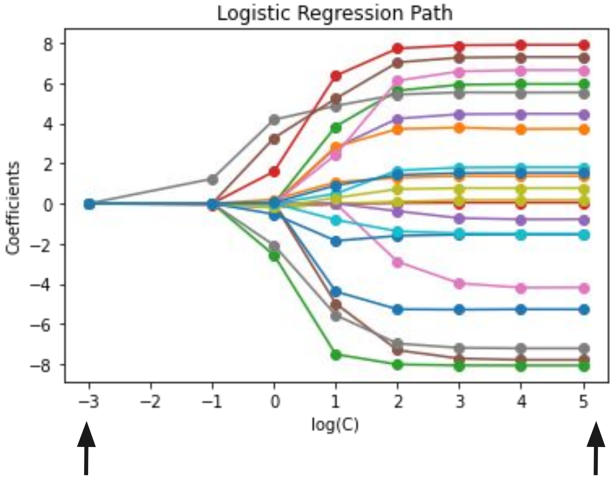


Choice of models - Our approach



Logistic Regression

- Logistic Regression provides a probabilistic interpretation of clinical risk.
- Tuned the regularization strength based on the best AUC score.

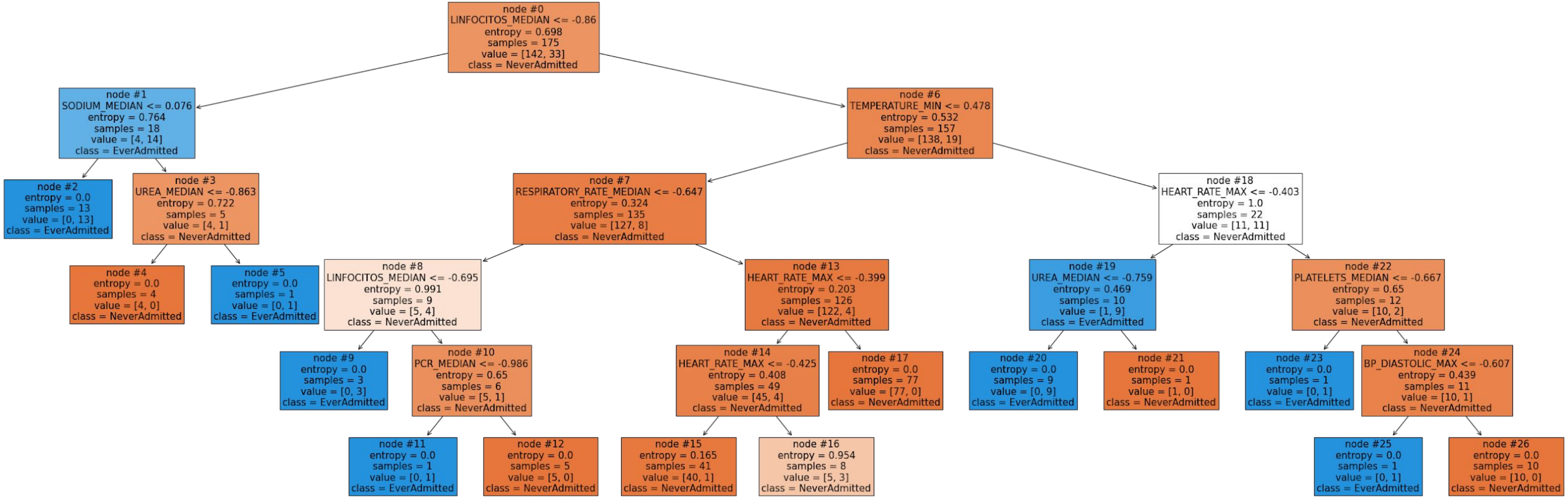


Over Regularized = Underfit

Under Regularized = Overfit

Decision Tree

- Provides an explainable, greedy model that physicians can use to predict ICU admission



Other base learners

- SVM
  - Works well for binary classification by finding a separating hyperplane between the two classes

- Perceptron
  - Biologically inspired, discriminative classifier that is iterative and minimizes misclassification error

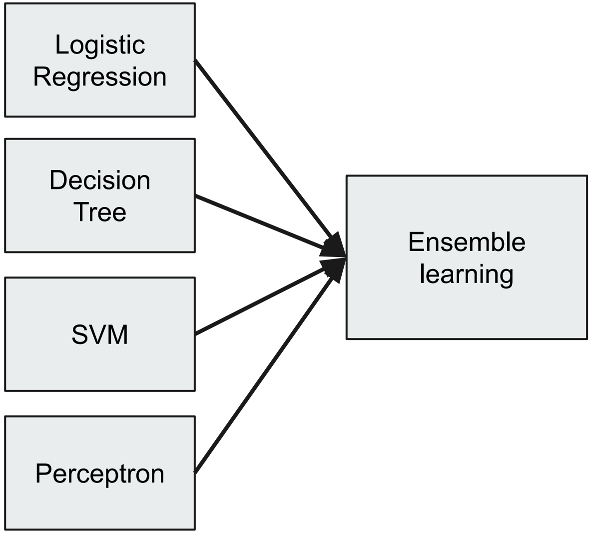
Model performance analysis

Models	AUC	F1	Accuracy	Precision	Recall	Specificity
Perceptron	0.786	0.545	0.659	0.375	1.000	0.571
SVM	0.776	0.632	0.841	0.600	0.667	0.886
Logistic Regression	0.721	0.556	0.818	0.556	0.556	0.886
Decision Tree	0.649	0.435	0.705	0.357	0.555	0.743

Next steps - Improving model performance

- SMOTE
  - Address problem of imbalanced classes of target variable (Ever Admitted and Never Admitted)

- Ensemble learning
  - Combine predictive power of different models to deliver a consensus prediction

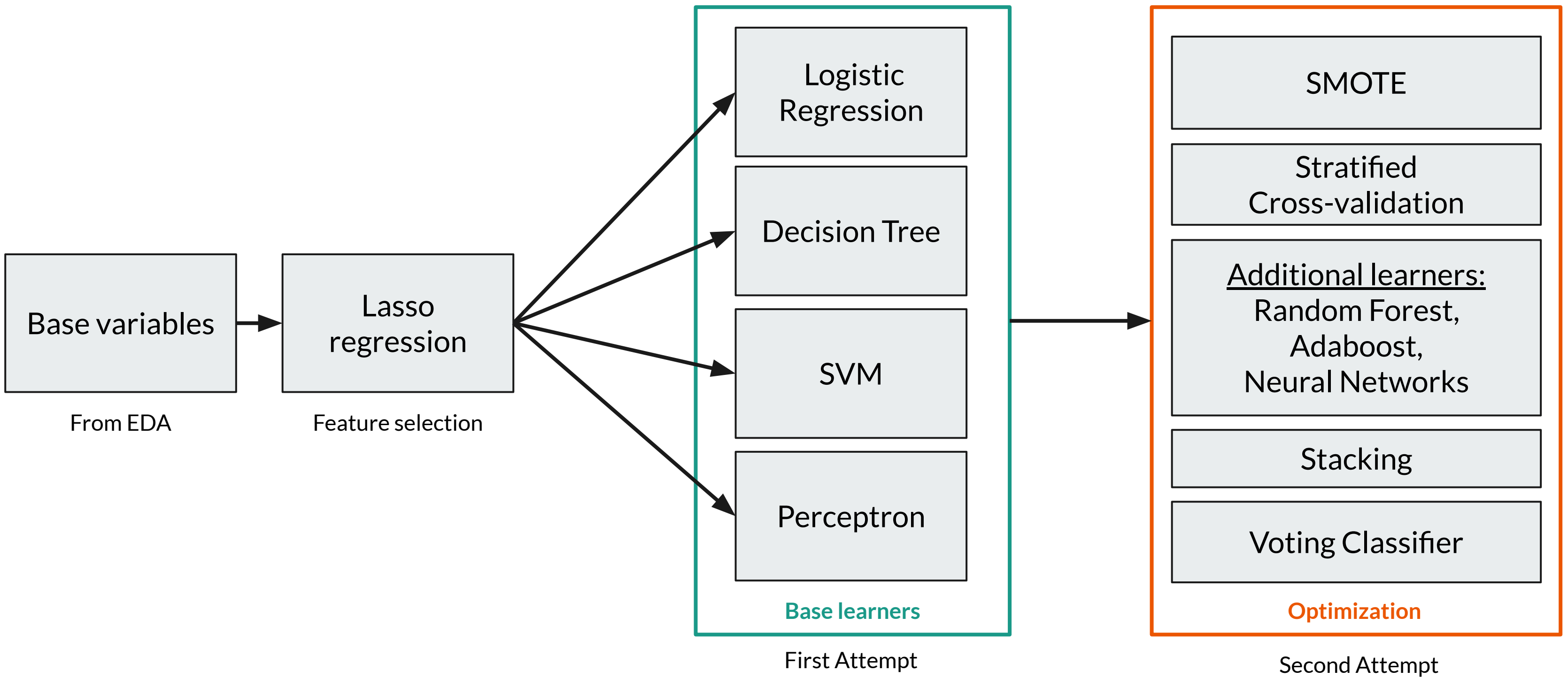


End

Charts for final report

Optimized Models (Post SMOTE & GridSearch CV)	
Perceptron	Perceptron(alpha=0.0003727593720314938, l1_ratio=0.7500000000000001, penalty='elasticnet', random_state=42)
SVM	SVC(C=10, degree=1, gamma=0.1, random_state=42)
Logistic Regression	LogisticRegression(C=0.1, l1_ratio=0.5, penalty='elasticnet', random_state=42, solver='saga')
Decision Tree	DecisionTreeClassifier(criterion='entropy', max_depth=5, min_samples_leaf=0.1, min_samples_split=0.1, random_state=42)
RandomForest	RandomForestClassifier(criterion='entropy', max_depth=4, min_samples_leaf=0.1, min_samples_split=0.1, n_estimators=50, random_state=42)
AdaBoost	AdaBoostClassifier(base_estimator=DecisionTreeClassifier(criterion='entropy', max_depth=17, min_samples_leaf=0.1, min_samples_split=0.1),

Choice of models - Our approach



Choice of models - Our approach