**Data Analytics Capstone Topic Approval Form**

**Student Name:** Julia Amanda Terzin

**Student ID:** 009968680

**Capstone Project Name:** XGBoost Classification Analysis of Comorbidities for Predicting Mortality from COVID-19

**Project Topic**: Predictive Model for COVID-19 Mortality

**X This project does not involve human subjects research and is exempt from WGU IRB review.**

**Research Question:** Can a predictive model be developed using the comorbidities of patients with COVID-19?

**Hypothesis**: H0: Comorbidities in patients with COVID-19 cannot be used to predict the risk of mortality with statistical significance.

H1: Comorbidities in patients with COVID-19 can be used to predict the risk of mortality with statistical significance.

**Context:** The contribution of this study to the field of Data Analytics and the MSDA program is to create a predictive model to identify patients at a higher risk of mortality from COVID. The COVID pandemic has caused excessive strain on hospital resources (French et al., 2021). The ability to identify mortality risk based on preexisting conditions will allow hospitals and healthcare organizations better manage resources through prevention, early intervention once diagnosed, and identification of patients who can be safely managed as an outpatient (Brüggemann et al.,2021). In this study, an eXtreme Gradient Boosting (XGBoost) classifier will be employed to predict the outcome of patients using their comorbidities as predictors. XGBoost is an ensemble learning algorithm that uses Classification and Regression trees (CART) to categorize an instance (a patient) based on its features (comorbidities)(Bex, 2021). Wang et al. (2022) created a successful model for predicting the prognosis of cardiac ICU patients using clinical data.

**Data:** The data set is publicly available synthetic data generated using The MITRE Corporation’s SyntheaTM, a Synthetic Patient Population Simulation. The data set was created to facilitate modeling COVID data without privacy and security risk to patients (Walonoski et al., 2020). It contains 19 csv files of which 4 will be used: patients, conditions, encounters, and observations. It contains data from 124,150 patients, 71,329 will be selected based on the criteria of adults with a positive COVID test.

[Downloads | Synthea (mitre.org)](https://synthea.mitre.org/downloads)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **csv** | **Rows** |  | **Columns** |  |
|  | **original** | **kept** | **original** | **kept** |
| patients | 124,150 | 71,329 | 25 | 3 |
| conditions | 1,143,900 | 114,349 | 6 | 3 |
| encounters | 1,881,954 | 21,929 | 15 | 6 |
| observations | 1,621,9969 | 132,648 | 8 | 4 |

Limitations: Actual patient data availability is restricted due to concerns for patient privacy and security. The synthetic data set is very useful for developing models to apply to actual patient data but does not contain immediately actionable insights to manage hospital resources (Walonoski et al., 2018).

Delimitations: The full data set has a large amount of data not relevant to this study, such as insurance information and billing, administrative codes, and location and demographic information. These fields will be removed from the final set. It also contains non-covid patients, which will also be removed. Only adult patients (>18) will be retained as children have a different susceptibility and mortality profile (Khera et al., 2021). Proper exclusion criteria are important to produce a consistent and reliable analysis (Nikolopoulou, 2022).

**Data Gathering:** After downloading the 4 csv files from miter.org, the data needed from each will be extracted. The patients file contains one row per patient with demographic information. The patient id, birth date, and gender will be extracted. The conditions, observations, and encounters files track patients using the patient id. There are multiple rows per patient by date of occurrence. The conditions file contains separate entries for each condition ascribed to a patient with dates assigned to each entry signifying when it was documented. The conditions of interest to this study dated before the diagnosis of COVID will be extracted for each patient. The observations file contains clinical data. Positive COVID tests with the date of diagnosis will be extracted as well as the cause of death with the date. Finally, the encounters file contains healthcare visit information, the inpatient visits will be used to extract intensive care and total hospital days. Once merged with the patients file, the final dataset will have one line per patient, 71,091 rows. All categorical variables will be one-hot encoded. Data sparsity is 0, so there will be no need to impute any missing values. ICU\_days and Total\_Hosp\_days will not be used in the model. They will be used to create a comparison of hospital utilization between the recovered and mortality groups to demonstrate the business application of the study. Age, the one continuous variable to be used in the model, contains outliers and will be treated. No scaling of the data will be done as the XGBoost classifier does not require it (Mendekar, 2021).

|  |  |  |  |
| --- | --- | --- | --- |
| **Field** | **Data Type** | **Variable Type** | **Use** |
| Id | Categorical | Independent | Index |
| ICU\_days | Continuous | Independent | Demonstration |
| Total\_Hosp\_days | Continuous | Independent | Demonstration |
| Age | Continuous | Independent | In model |
| Gender | Categorical | Independent | In model |
| Alzheimer`s\_disease | Categorical | Independent | In model |
| Asthma | Categorical | Independent | In model |
| Obesity | Categorical | Independent | In model |
| Chronic\_congestive\_heart\_failure | Categorical | Independent | In model |
| Chronic\_kidney\_disease | Categorical | Independent | In model |
| Chronic\_obstructive\_bronchitis | Categorical | Independent | In model |
| Coronary\_Heart\_Disease | Categorical | Independent | In model |
| Diabetes | Categorical | Independent | In model |
| History\_of\_myocardial\_infarction | Categorical | Independent | In model |
| Hyperlipidemia | Categorical | Independent | In model |
| Hypertension | Categorical | Independent | In model |
| Hypertriglyceridemia | Categorical | Independent | In model |
| Prediabetes | Continuous | Independent | In model |
| Pulmonary\_emphysema | Categorical | Independent | In model |
| Stroke | Categorical | Independent | In model |
| Current\_smoker | Categorical | Independent | In model |
| Former\_smoker | Categorical | Independent | In model |
| Mortality | Categorical | Dependent | In model |

**Data Analytics Tools and Techniques**: Design of the Study: Initial statistical evaluation of the data will include: 1) A chi-squared test of independence for the categorical variables to evaluate the relationship with the target variable (Bevans, 2022). 3) A correlation heatmap of the predictor variables to look at multicollinearity. XGBoost classification will be the model of choice due to its ability to handle the multicollinearity of the predictor variables and capture non-linear relationships as opposed to logistical regression (Mendekar, 2021). The data will be split into train (80%) and test (20%). Because the data is unbalanced with a 0.49% mortality rate, the split will be stratified to contain equal proportions of the target variable (Brownlee, 2020).

**Justification of Tools/Techniques:** Python will be used for the data preparation and for developing the model. Python was chosen for its versatility and numerous libraries available. While R is very strong for statistical computations, Python has a lot of additional functionality (Singh, 2022). Compared to SAS, Python is open source and more accessible as well as having much better customizable graphics (Dutta, 2021).

**Project Outcomes**: The expected project outcome is to create an XGBoost Classification model that can reliably predict mortality based on comorbidities. Support for the alternative hypothesis is found in Wang et al. (2022) using XGBoost classification to predict mortality.

**Projected Project End Date**: 11/19/2022

**Sources**:

Bevans, R. (2022). Choosing the Right Statistical Test | Types and Examples. Retrieved November 08, 2022, from [Choosing the Right Statistical Test | Types & Examples (scribbr.com)](https://www.scribbr.com/statistics/statistical-tests/)

Bex T. (2021). Beginner’s Guide to XGBoost for Classification Problems. Retrieved November 08, 2022, from [Beginner’s Guide to XGBoost for Classification Problems | Towards Data Science](https://towardsdatascience.com/beginners-guide-to-xgboost-for-classification-problems-50f75aac5390)

Brownlee, J. (2020). Train-Test Split for Evaluating Machine Learning Algorithms. Retrieved November 08, 2022, from [Train-Test Split for Evaluating Machine Learning Algorithms (machinelearningmastery.com)](https://machinelearningmastery.com/train-test-split-for-evaluating-machine-learning-algorithms/)

Brüggemann, S., Chan, T.C., Wardi, G., Mandel, J., Fontanesi, J., & Bitmead, R.E. (2021). Decision support tool for hospital resource allocation during the COVID-19 pandemic. *Informatics in Medicine Unlocked, 24,* 100618 - 100618. [Decision support tool for hospital resource allocation during the COVID-19 pandemic - ScienceDirect](https://www.sciencedirect.com/science/article/pii/S2352914821001088)

Dutta, B. (2021). 8 Differences Between SAS and Python. Retrieved November 08, 2022, from [8 Differences Between SAS and Python | SAS vs Python | Analytics Steps](https://www.analyticssteps.com/blogs/8-differences-between-sas-and-python)

French G, Hulse M, Nguyen D, et al. (2021). Impact of Hospital Strain on Excess Deaths During the COVID-19 Pandemic — United States, July 2020–July 2021. *MMWR Morb Mortal Wkly Rep*, *70,* 1613–1616. [Impact of Hospital Strain on Excess Deaths During the COVID-19 Pandemic — United States, July 2020–July 2021 | MMWR (cdc.gov)](https://www.cdc.gov/mmwr/volumes/70/wr/mm7046a5.htm#suggestedcitation)

Khera, N., Santesmasses, D., Kerepesi, C., & Gladyshev, V. N. (2021). COVID-19 mortality rate in children is U-shaped. *Aging, 13(16)*, 19954–19962. [COVID-19 mortality rate in children is U-shaped - PubMed (nih.gov)](https://pubmed.ncbi.nlm.nih.gov/34411000/)

Leung, K. (2021). Assumptions of Logistic Regression, Clearly Explained. Retrieved November 08, 2022, from [Assumptions of Logistic Regression, Clearly Explained | by Kenneth Leung | Towards Data Science](https://towardsdatascience.com/assumptions-of-logistic-regression-clearly-explained-44d85a22b290)

Mendekar, V. (2021). Machine Learning – it’s all about assumptions. Retrieved November 08, 2022, from [It’s All About Assumptions, Pros & Cons | by Vishalmendekarhere | The Startup | Medium](https://medium.com/swlh/its-all-about-assumptions-pros-cons-497783cfed2d)

Nikolopoulou, K.(2022). Inclusion and Exclusion Criteria | Examples & Definition. Retrieved November 11, 2022, from [Inclusion and Exclusion Criteria | Examples & Definition (scribbr.com)](https://www.scribbr.com/methodology/inclusion-exclusion-criteria/)

Singh, M. (2022). Python Vs R in 2022! Retrieved November 11, 2022, from [Python Vs R in 2022! Welcome back! R and Python are very… | by Manpreet Singh | Medium](https://preettheman.medium.com/python-vs-r-in-2022-a26ad7ffe6a7)

Wang, X., Zhu, T., Xia, M., Liu, Y., Wang, Y., Wang, X., Zhuang, L., Zhong, D., Zhu, J., He, H., Weng, S., Zhu, J., & Lai, D. (2022). Predicting the Prognosis of Patients in the Coronary Care Unit: A Novel Multi-Category Machine Learning Model Using XGBoost. *Frontiers in cardiovascular medicine, 9,* 764629. [Predicting the Prognosis of Patients in the Coronary Care Unit: A Novel Multi-Category Machine Learning Model Using XGBoost - PubMed (nih.gov)](https://pubmed.ncbi.nlm.nih.gov/35647052/)

Walonoski, J., Klaus, S., Granger, E., Hall, D., Gregorowicz, A., Neyarapally, G., Watson, A., & Eastman, J. (2020). Synthea™ Novel coronavirus (COVID-19) model and synthetic data set. *Intelligence-based medicine, 1*, 100007. [Synthea™ Novel coronavirus (COVID-19) model and synthetic data set - ScienceDirect](https://www.sciencedirect.com/science/article/pii/S2666521220300077?via%3Dihub)

Walonoski, J., Kramer, M., Nichols, J., Quina, A., Moesel, C., Hall, D., Duffett, C., Dube, K., Gallagher, T., & McLachlan, S. (2018). Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. *Journal of the American Medical Informatics Association: JAMIA, 25(3)*, 230–238. [Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record - PubMed (nih.gov)](https://pubmed.ncbi.nlm.nih.gov/29025144/)

**Course Instructor Signature/Date:**

**Institutional Review Board Quiz and Approval**

Have you read and understood the “Human Subjects FAQ” page and completed the “Human Subjects FAQ Quiz” at the WGU Institutional Review Board (IRB) website? (<https://irb.wgu.edu/info/Pages/Home.aspx>)

Yes, I have read and understood the “Human Subjects FAQ” and have provided email proof of my completed quiz in appendix A. (<https://irb.wgu.edu/info/Pages/Human-Subjects-FAQ-Quiz.aspx>)

No, I have not completed the Human Subjects FAQ quiz.

Assess whether your capstone proposal complies with WGU’s IRB standards for exemption status. Explain why you believe the proposed project complies with the standards for exemption status. If it does not, make arrangements with a course mentor and the IRB for approval.

The research complies with WGU’s IRB exemption status because:

* Research involving the collection or study of freely available de-identified existing data
* Research that does not employ methodology on human subjects.

The research requires approval from WGU’s IRB because:

Yes, I would like to schedule a conference to discuss my project.

To be filled out by a course mentor:

The research is exempt from an IRB Review.

An IRB approval is in place (provide proof in appendix B).

Course Mentor’s Approval Status: Approved

Date: Click here to enter a date.

Reviewed by:

Comments: Click here to enter text.

**#http://scikit-learn.org/stable/auto\_examples/model\_selection/plot\_roc.html#sphx-glr-auto-examples-model-selection-plot-roc-py**

**fpr, tpr\_recall, \_ = sklearn.metrics.roc\_curve(self.y\_true,self.y\_score,pos\_label = 1)**

**roc\_curve, = self.ax[0].plot(fpr, tpr\_recall, color=self.roc\_color, lw=self.lw, linestyle = self.main\_linestyle)**

**self.ax[0].fill\_between(fpr, tpr\_recall, step='post', alpha=0.2, color=self.roc\_color)**

**self.ax[0].plot([0, 1], [0, 1], color=self.neutral\_color, lw=self.lw, linestyle=self.neutral\_linestyle) #diagonal line**

**self.ax[0].set\_xlim([0.0, 1.0])**

**self.ax[0].set\_ylim([0.0, 1.05])**

**self.ax[0].set\_xlabel('False Positive Rate')**

**self.ax[0].set\_ylabel('True Positive Rate (Recall)')**

**self.ax[0].set\_title('AUROC=%0.2f' % sklearn.metrics.auc(fpr, tpr\_recall))**