DASHBOARD FOR MACHINE LEARNING MODELS IN HEALTH CARE

by

Wejdan H. Bagais
A Thesis
Submitted to the
Graduate Faculty
of
George Mason University
in Partial Fulfillment of
The Requirements for the Degree
of
Master of Science
Health Informatics

Committee:

Janusz Wojtusiak, Ph.D., Thesis Director

Janusz Wojtusiak, Ph.D., Program Director

Pj Maddox

Peggy Maddox, Ph.D., Department Chair

Germaine Louis

Germaine M. Louis, Ph.D., Dean and Professor, College of Health and Human Services

Date: 7/29/2021

Summer Semester 2021
George Mason University
Fairfax, VA

Dashboard for Machine Learning Models in Health Care

A Thesis submitted in partial fulfillment of the requirements for the degree of Master of Science at George Mason University

by

Wejdan H Bagais
Bachelor of Science
King Abdulaziz University, 2015

Director: Janusz Wojtusiak, Associate Professor of Health Informatics and Director of the Machine Learning and Inference Laboratory
Department of Health Administration and Policy

Summer Semester 2021
George Mason University
Fairfax, VA

DEDICATION

This is dedicated to my family for their support and believe on my abilities. My supervisor who was the guiding light every step of the way as I researched for this dissertation

ACKNOWLEDGEMENTS

I would like to thank my family, friends, and supporters who have made this happen. Dr. Janusz Wojtusiak, and the other members of my committee were of invaluable help. Finally, thanks go out to the Fenwick Library for providing a clean, quiet, and well-equipped repository in which to work.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Figure Page

LIST OF ABBREVIATIONS

Machine Learning	ML
Area Under the Curve	AUC
Receiver operation characteristic	ROC
False Positive Rate	FPR
True Positive Rate	TPR

ABSTRACT

DASHBOARD FOR MACHINE LEARNING MODELS IN HEALTH CARE

Wejdan H Bagais, M.S.

George Mason University, 2021

Thesis Director: Dr. Janusz Wojtusiak

Presentation of machine learning (ML) model results plays an important role in decision makers' trust and use. Yet, there has been little agreement on how information should be visualized to present models' evaluations. The purpose of this thesis is to formulate an approach to visualize the results of classification model's evaluation to increase decision makers' trust. This work proposes a dashboard that visualizes supervised ML model performance in a dashboard which is split into three main sections: statistical measures, feature importance, sensitivity analysis. Three sample dashboards were generated and evaluated using a survey by ten faculty members and students from George Mason University most of which said that the dashboard provides useful information and gives a better understanding of the model behavior than other methods they have experienced.

Model evaluation strategies differ based on the prediction problem considered. However, a consistent representation of evaluation results may increase decision makers' trust in the models. The next step of this project is to visualize the difference between multiple models.

INTRODUCTION

The use of machine learning (ML) has grown massively over the last decade. Most current research on ML stops at reporting the statistical measures and fail to report details of models which affect the decision-makers’ trust and adoption. For example, Krause et al. (2016) explain the experience of a stakeholder who struggled whether not to employ a model that predicted diabetic risk. The model had high accuracy, but the analysts could not explain how the features impacted the prediction. In healthcare, understanding the effect of the predictors is crucial to trusting the model (Apley & Zhu, 2020). Visualization methods are among the most useful tools for understanding a ML model (Alsallakh et al., 2014). Tonekaboni et al. (2019) emphasize that carefully designed visualizations increase the clinicians’ understanding. However, there is little agreement on what should be visualized to evaluate a model.

The type of evaluation visuals change based on the type of ML classifier and the prediction problem. This thesis claims that some of the common model evaluation elements should be visualized in all models. Therefore, this work proposes a dashboard that takes the model, the training data, the testing data, and the list of attributes and then displays the most common evaluation visuals in in a user-friendly dashboard. The dashboard has three main sections: statistical measures, feature importance, and sensitivity analysis. The purpose of the dashboard is to help decision makers understand the strength and weaknesses of the model and uncover the relationship between features and predictions, which lead to an increase in the decision makers' trust.

Related Work

While a considerable amount of literature has been published on explaining the performance of ML models, most studies focus on one measure, a specific ML method, or interactive presentation of ML results. Papers under interactive ML are the closest work to this thesis. In interactive machine learning, "the model gets updated immediately in response to user input" (Amershi et al., 2014, p. 106). Most of the model explanation systems that use interactive machine learning ask the user to input a hypothetical scenario and display the model performance for that scenario. However, this thesis focuses on the global explanation for the model and the effect of the features rather than the local explanation (per patient scenario).

There have been a number of related visualization and model explanation systems developed over the past years, such as:

The what-if tool (WIT) is "an open-source application that allows practitioners to probe, visualize, and analyze ML systems, with minimal coding" (Wexler et al., 2020, p. 56). WIT has four main functions. First, exploring data which shows summary statistics and charts of distributions of all features in the loaded dataset. Second, investigating What-If hypotheses, shows model performance based on user test hypotheses by finding counterfactuals and observing partial dependence plots. Third, evaluating performance and fairness, analyze and compare model performance based on slices of data. Forth, comparing two models, which compares all supported measures and partial dependence plots between the two models (Wexler et al., 2020). As the name suggests, the WIT is an interactive system that shows the model behavior based on user input scenarios. In comparison, this thesis focuses on displaying the final model behavior without diving into the local sensitivity analysis.

Manifold, is “a generic environment for comparing and debugging a broad range of machine learning models” (Zhang et al., 2019, p. 9). Manifold compares ML models using two main visuals: a summary statistics at feature level and a comparison of model pairs (Zhang et al., 2019). Both Manifold and this thesis display the features’ distribution per classification category to explain the relationship between the prediction and the features.

Prospector provides interactive partial dependence diagnostics to understand the effect of features on prediction. Prospector visualizes: patient selection (a list of patients based on prediction and ground truth), Patient inspection (the change of prediction based on the change of feature values for the selected patient), and partial dependence plots (which demonstrate the effect of a feature on the prediction) (Krause et al., 2016). Both Prospector and this thesis include the visualization of partial dependence plots. However, Prospector focuses more on patient-level analysis while this work focuses on the overall feature effect.

Similarly, there are several systems that focus on prediction explanation as part of decision support. The more notable of the systems are:

LIME is “a modular and extensible approach to faithfully explain the predictions of any model in an interpretable manner” (Ribeiro et al., 2016, p. 114). LIME explains the predictors for a specific case, while this paper focuses on the global explanation for the model and its features.

SHAP stand for “Shapely Additive explains Explanations”. SHAP explains the output of any ML model using a game theory approach. SHAP also focuses on local explanation (Lundberg & Lee, 2017).

Some other papers focus on a specific type of data or measures. For example, FeatureInsight which focuses on defining dictionary features for classification models (Brooks et al., 2015), Samek et al. (2017) paper focuses on visualizing deep neural network DNN, Adams & Hand (1999) proposed LC index as an alternative for the ROC curve, and Raymaekers et al. (2020) advised using mosaic plot instead of confusion matrix.

BACKGROUND

There are different definitions of machine learning (ML) consistent with how the field has evolved over time. Machine learning (ML) is ‟a computer observes some data, builds a model based on the data, and uses the model as both a hypothesis about the world and a piece of software that can solve problems" (Russell & Norvig, 2021, p. 1224). More broadly, machine learning (ML) is about building computer systems that learn how to perform given tasks, instead of being explicitly programmed to do so.

In healthcare, ML is increasingly used to "describe automatized, highly flexible, and computationally intense approaches to identifying patterns in complex data structures (Jiang et al., 2020.)" There are three main types of learning: supervised, unsupervised and reinforcement. In supervised learning, the machine learns a function that map given input to the output, or in other words learn from provided examples. The model uses labeled data to find the patterns. In unsupervised learning, the model identifies the patterns/clusters without predefined outcome (output attributes). In reinforcement learning, the model learns from sequences of rewards and punishments (Russell & Norvig, 2021, p. 1226 - 1227), that is explores possible solutions some of which being good and others bad.

This thesis focuses on supervised learning, but the presented methods can be extended to other types of learning. While supervised learning is increasingly adapted in health research, there is little agreement on what information should be visualized to present models' evaluations. Visualizing model results helps both the analyst who constructs the model and the clinicians who are end users better understand its performance and consequently trust and use the model. However, most researchers and data analysts who construct ML-based models stop at providing basic statistical measures to evaluate the model. Even though, there are a vast number of literatures about model explanation, most of which is insightful only for algorithms experts (Tonekaboni et al., 2019). The purpose of this thesis is to formulate a consistent approach to visualize the performance of ML classification models to help data analysts understand properties of the constructed models and at the same time increase decision makers' trust. The following sections in this chapter introduces the most important concepts when developing model visualizations.

Supervised Learning

Supervised learning is the ML task that aims to creating model M that maps the input attributes X to predict the output attributes Y. The model M is learned from examples (labeled data). There are two types of outputs: classification and regression. In classification, the predicted output Y is categorical (qualitative) while in regression the predicted output Y is continuous (quantitative). Consequently, metrics for evaluating the supervised learning model performance differ for classification and regression problems. This work focuses on visualizing the result of supervised classification learning models to improve the understanding of the model behavior.

The model used a training set to find the best function that maps the inputs to the output. To evaluate the model, testing set is used to test the model performance on data the model has not seen before. The model "generalizes well if it accurately predicts the outputs of the test set.” (Russell & Norvig, 2021, p. 1229). There are many nuances of how the test sets are created that are out of scope of the presented work.

Bias and Variance

The model bias is the tendency to simplify the model while the variance is the tendency to fit the training data. Strong bias may lead to underfitting while strong variance on the other hand may results of overfitting the data. An example of strong bias is the linear function. If there is a trend that does not fit in the overall slope of the line, the model will not be able to capture it. “Often there is a bias–variance tradeoff: a choice between more complex, low-bias hypotheses that fit the training data well and simpler, low-variance hypotheses that may generalize better” (Russell & Norvig, 2021, p. 1231). The model overfits the data when it performs well on the training data but poorly on the testing data. Figure 1 visualizes underfit, overfit, and balanced model.

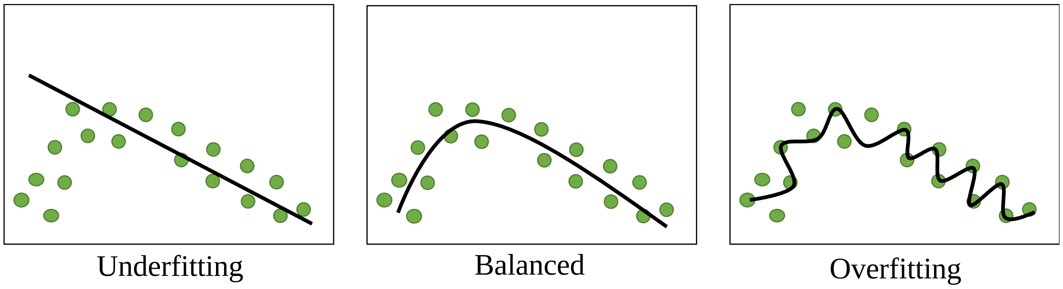


Figure 1 Model Fit: Underfitting vs. Overfitting

Evaluation Types

The first step in understanding the model performance is by applying some of numerous statistical metrics. The statistical results show the overall model performance by comparing the correctly classified cases with misclassified cases using multiple measures (see next section). Nonetheless, these statistical measures alone are not enough to understand model properties and consequently gain the trust and adoption of the model especially in healthcare settings.

Feature/attribute importance refers to the strength of relationship between that attribute and the model output. Additionally, the uncertainty analysis explains how the features (inputs) affect the output. The study of Tonekaboni et al. 2019 shows that feature importance and uncertainty analysis improve the model explanation for decision makers. While information about individual cases in the model gives a lot of information about how a model makes decision, this work is focusing on global model performance only. The goal of this study is to generate a dashboard that visualizes the model performance using statistical measures, features importance, and sensitivity analysis.

Statistical Measures

There is a wide range of statistical measures of model accuracy. The most frequently used are confusion matrix, accuracy, area under the ROC, precision, recall, and f-score. The confusion matrix table shows the number of cases that classified correctly and cases that misclassified by the model. The confusion matrix is a table that shows the number of true positive, true negative, false positive, and false negative. True means the model correctly classified the cases, while false means the model misclassified. Positive means belong to the targeted class, while negative means do not belong to the targeted class. Often, the confusion matrix is also visualized using a heat map.

Accuracy is the number of all cases that are correctly classified divided by the total number of cases. The Accuracy has two drawback points: do not count the different between types of errors and reliant on the class distribution (Jasmina Dj. Novaković et al., 2017). Accuracy is not a good measure when the data are skewed/imbalanced. For example, in a case of predicting disease and 90% of the data were healthy patients, the model can achieve 90% accuracy if the model predicts that no one has the disease. Therefore, many researchers prefer using Area under the Receiver Operating Characteristic curve (AUC) when dealing with binary classification problems. AUC allows the user to select the is a tradeoff point between false positive and false negative (Russell & Norvig, 2021, p. 1337 & 1338). AUC measures the area under the Receiver Operating Characteristic (ROC) curve. Swets (1988) introduced the Receiver Operating Characteristic (ROC) and call it the measurement accuracy based on the surface. The ROC curve x-axis is False Positive Rate (FPR), and the y-axis is True Positive Rate (TPR).

$$TPR = \frac{TP}{TP+FN}, FPR = \frac{FP}{TN+FP}, AUC = \int_{FPR=0}^{1} TPR(FPR)$$

Precision is the percentage of correctly classified cases from all positive classified cases. In other words, out of all the cases that the model is classified as positive, how many is correctly classified (Wojtusiak, 2021). The precision formula is $\frac{TP}{TP+FP}$. On the other hand, the recall (known as sensitivity in biomedical literature) is the percentage of the correctly classified cases from all positive cases. In other words, out of all positive cases, how many correctly predicted (Wojtusiak, 2021). The recall formula is $\frac{TP}{P}$. Usually, there is a trade-off between precision and recall. When the recall value is high, the model predicts most of the positive cases, which may include a high number of false positives. When the precision is high, the number of false positives cases is low, which may result in misclassifying some positive cases. F-score is a measure of accuracy that balances recall and precision. F-score formula is equal to $\frac{2 \times precision \times recall}{precision + recall}$ (Wojtusiak, 2021).

Features Metrics

Statistical measures show the general model performance; the next step is to dive more into the specific attributes that are used by the model. Some of the feature measures are features importance and feature sensitivity. Also, in the case of using tree algorithms, presenting the model tree can show the model logic of predicting. However, if there are many attributes, printing the full tree may not be helpful.

Features Selection

Understanding the relationship between the features and the output helps in improving the prediction. Having a large number of features may result in overfitting (Tang et al., 2014) when data are small. The feature selection goal is to reduce the number of features to have better learning performance, lower computational cost, and better model interpretability (Tang et al., 2014). In the presented research, the purpose of using feature selection methods is to visualize features importance to understand their effect on the model output. The three categories for supervised feature selection methods are filter models, wrapper models, and embedded models.

Filter model "relies on measures of the general characteristics of the training data" (Tang et al., 2014). Filter models do not influence by the selected ML model. Some examples for filter models are Relief, Fisher score, and Information Gain based methods. The wrapper model “uses the predictive accuracy of a predetermined learning algorithm to determine the quality of selected features.” (Tang et al., 2014). Wrapper models are expensive to run when having large number of features. Wrapper methods examples are recursive feature elimination, sequential feature selection algorithms, and genetic algorithms. Embedded models are a combination of filter and wrapper models. “The embedded model usually achieves both comparable accuracy to the wrapper and comparable efficiency to the filter model” (Tang et al., 2014). Some method examples are LASSO and random forest feature importance. In the presented work visualizations are created for the feature importance using correlation, LASSO, random forest feature importance, and permutation.

Feature Sensitivity Analysis

The sensitivity analysis is “the study of how the uncertainty in the output of a model (numerical or otherwise) can be apportioned to different sources of uncertainty in the model input” (Saltelli et al., 2004, p. 45). There are local and global sensitivity analysis. Local sensitivity analysis focuses on explaining the model decision for a specific case. While this analysis if very informative, it is out of the thesis scope. The thesis focuses on explaining the overall model behavior. Therefore, global sensitivity analysis used.

Global sensitivity analysis is the effect of a single attribute on the model. After selecting an attribute X, the dashboard shows X distribution, predictions’ means per X value for training, testing, and random data, and predictions’ means using fixed values for

X to check the effect of X alone. Finally, two scatter plots are displayed to compare the original predictions and predictions when X values change slightly.

Visualization Techniques

Data visualization is defined as the representation of the information in a graphical form. Visualization is used because graphs are typically faster to interpret compared to plain text and contain more information. In machine learning (ML), analysts usually visualize ML models to understand and compare models' performance and to communicate the model performance with the decision makers so they can trust and use the model. Visualization is the key to understanding the model performance for both analysts and decision makers. Tonekaboni et al. study emphasizes that careful design visualization plays an important role in clinicians understanding (Tonekaboni et al., 2019). Therefore, having clear graphs that do not present misleading information is important. This section describes the guidelines that this paper will follow in order to have clear graphs.

Nussbaumer Knaflic in her book *Storytelling with Data* notes that data visualization is a way of communication. Therefore, before starting to build the visuals, identifying the audience and the purpose is needed. Identifying the audience and the message at the beginning saves time and ensures that the communication meets its purpose (Nussbaumer Knaflic, 2015, p 21-23). The targeted audience are the analyst and the decision makers in healthcare setting. The paper message is to represent the models' performance in two dashboards one to clarify a single ML model performance and the other to compare different models' performance.

There are some general practices that help in having a clear visualization. Starting with the two graphical practices devised by Edward Tufte's book *The Visual Display of Quantitative Information* (2001): The first practice is graphical integrity, which is telling the truth about the data. The goal is to "show data variation, not design variation" (Tufte, 2001, p 61). To check if the variation is misrepresented, Tufte calculates the Lie Factor that is the size of the effect shown in the graph divided by the size of the effect in the data. A lie factor greater than 1.05 or less than 0.95 indicates considerable misrepresentation. Additionally, the data can be misleading if the graph does not start with zero or when using line charts if the data does not have a continuous meaning (Healy, 2015, p 84 - 86).

The second practice, as discussed by Tufte, is graphical excellence, in which excellent "consists of complex ideas communicated with clarity, precision, and efficiency" (Tufte, 2001, p 13). The graphics should focus on the data rather than the design. Tufte also writes that "Graphical excellence is that which gives to the viewer the greatest number of ideas in the shortest time with the least ink in the smallest space" (Tufte, 2001, p 51). One of Tufte's theories is the data-ink, "a large share of ink on a graphic should present data-information" (Tufte, 2001, p 93). The data-ink ratio equation is data-ink divided by total ink used to print the graphic. To improve the statistical graphics, Tufte discussed three fundamental principles based on the data-ink theory:

- Above all else show the data
- Maximize the data-ink ratio, within reason
- Follow the erasing principles

- Erase non-data-ink, within reason
- Erase redundant data-ink, within reason (Tufte, 2001, P 92, 96, 100)

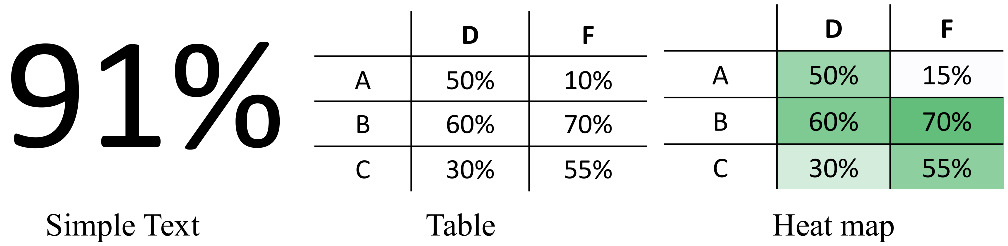
Commenting on the data-ink theory, Kieran Healy argues that sometimes simple is not better. In some cases, breaking the rules makes the visualization more unexpected and therefore memorable. Therefore, following ink minimalist theory is not always the best (2019, p. 9); balance is needed.

A balanced graph means that all the graph elements add some meaning or have a purpose. Therefore, the way to achieve balanced graphical excellence is by avoiding cluttering, the visual element that did not increase understanding. For example, one must avoid grids, duplicated information like duplicated labels, shadows, and 3D graphs when it is not necessary (Healy, 2019, p 11-14) (Nussbaumer Knaflic, 2015, p 73 - 84). Additionally, one should never add text to fill up the white space. Nussbaumer Knaflic compared the white space with the pauses in public speaking; if the speaker talks without pauses, the audience will lose their attention; likewise, if the graph has no white space, the audience will not get what the graph is about.

Choosing Visuals

There is a difference in the meaning between charts and graphs. Charts are the graphical representations of the data, while graphs are the graphical representation of the mathematical relationships between data. In other words, graphs are one type of chart. However, since most of the charts that will be used in this study are graphs, the words graph, and chart will be used interchangeably. Although there is a variety of charts that can be used, prioritize the audience's level of familiarity is important. Using popular charts mean less learning for the audience. Yet, if using a new chart is needed, an extra explanation is needed. This study uses a wide range of the most used charts and some other charts that might be new to the audience. The next section will show the usages of the most used charts and graphs: simple text, table, heatmaps, points, lines, bars, area, and pie chart.

Nussbaumer Knaflic highlights the usage of the most popular charts as follows: Simple text, used when we have one or two numbers only. If the audience members are interested in different parts of the data, tables represent the data in rows. Accordingly, each member looks at what they are interested in. To rank the table's data, heatmaps are used to add colors to the table to demonstrate the ranking meaning (2015, p. 35 - 43). Figure 2 shows an example of these charts.



91%

Simple Text

	D	F
A	50%	10%
B	60%	70%
C	30%	55%

Table

	D	F
A	50%	15%
B	60%	70%
C	30%	55%

Heat map

Figure 2 Most popular charts (simple text, table, and heat map)

Nussbaumer Knaflic also explained the usage of the most popular graphs: scatterplots, lines, bars, area, and pie. Scatterplots show the relationship between two things and are commonly used in scientific fields. For line graphs, Nussbaumer Knafli reviewed two types of line graphs: line and slope graphs. Line graphs are usually used for continuous data like time. One important practice in using line graphs is always to start with zero. If starting with zero is not possible, highlight that you are not starting from zero. Slope graphs are used when having two periods or points. For example, comparing this year patients' satisfaction per hospital with last year patients’ satisfaction, using the slope graphs help identify the increases and decreases easily (2015, p. 43 - 49). Figure 3 shows an example of these charts.

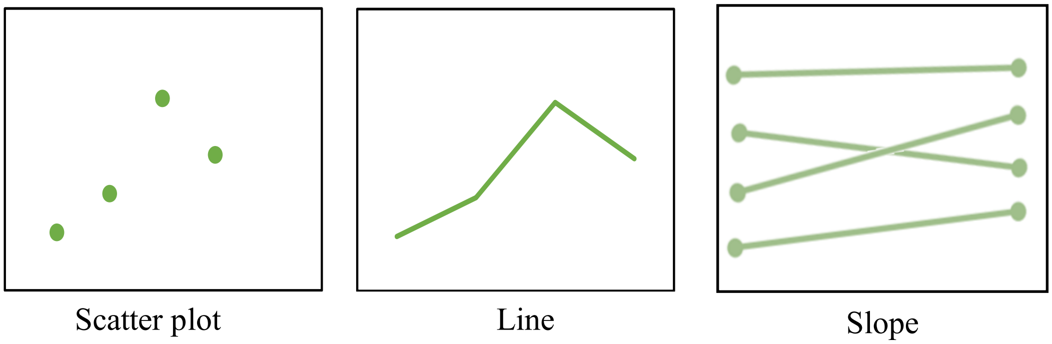
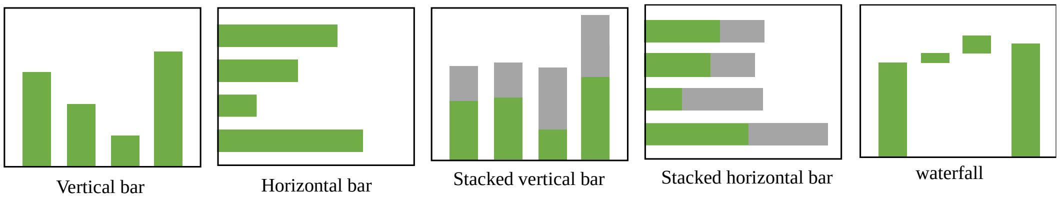


Figure 3 Most popular charts (scatter plot, line, and slope)

Bars graphs are another common and straightforward type of graphs. Some rules for bar graphs are you must have zero baseline, and the width of the bars should be wider than the white space between the bars. There are five types of bar charts: first, the vertical bar chart, which is the most popular. Second, the horizontal bar chart is a flip of the vertical bar chart to make it easier to read. Third, the stacked vertical bar chart compares the totals across different categories. Keep in mind that when the number of the categories increase, bars become harder to compare. Fourth, the stacked horizontal bar chart is similar to stacked vertical but adds the subcomponent parts' sense of consciousness. Fifth, the waterfall chart shows a starting point, increasing, decreasing, and ending point (Nussbaumer Knaflic, 2015, p. 50 - 59). Figure 4 shows examples of bar chars.



Vertical bar

Horizontal bar

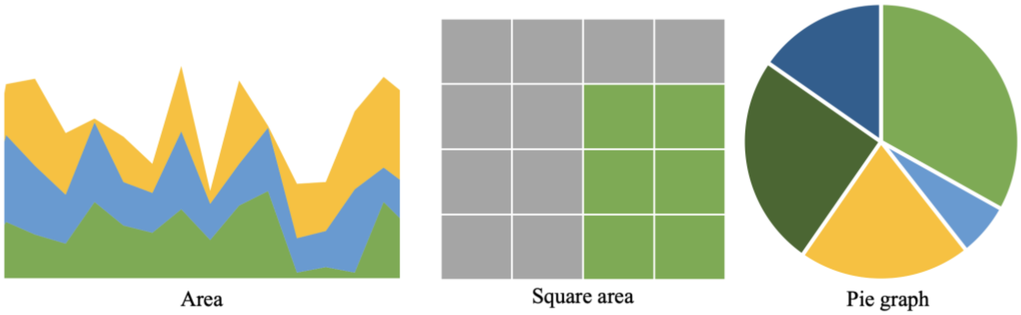
Stacked vertical bar

Stacked horizontal bar

waterfall

Figure 4 Bar charts

Area and pie graphs are less common compare to the other types of graphs. Nussbaumer Knaflic suggests avoiding them if possible. Area graphs are hard to read. However, area graphs can be useful to compare a widely different number of magnitudes (Nussbaumer Knaflic, 2015, p. 59 - 61). Figure 5 shows example for area and pie graphs.



Area

Square area

Pie graph

Figure 5 Area and pie graphs

METHODS

Presentation of machine learning (ML) models and their results plays an important role in analysts' and decision makers' understanding and consequently trust of the models. Visualization methods are among the best ways to explain the model performance as noted by Tonekaboni et al. (2019) who emphasizes that careful design visualizations increase the clinicians' understanding. While a considerable amount of literature has been published on explaining the performance of ML models, most studies focus on one measure or a specific ML method. This work summarizes the most important factors for evaluating any classification supervised ML model in one place using a dashboard which is represented in a website built using Flask. The website's inputs are the model, the attributes for both testing and training sets, and the columns' names. The output is the dashboard which contains the following parts: statistical measures, features important, and features sensitivity.

To demonstrate the dashboard, a random forest model was built to predict if the patient has Heart Disease using UCI Machine Learning Repository (1988) data set. The output attribute is the status of having heart disease: one if the patient had heart disease and 0 if the patient did not have heart disease. The input attributes are age in years, sex (1 for male and 0 for female) , chest pain type (1 for typical angina, 2 for atypical angina, 3 for non-anginal pain, and 4 for asymptomatic), resting blood pressure (in mm Hg on admission to the hospital), serum cholesterol (in mg/dl), fasting blood sugar (1 if > 120 mg/dl otherwise 0), resting electrocardiographic results (0 for normal,1 for having ST-T wave abnormality, and 2 for probable or definite left ventricular hypertrophy according to Estes' criteria), maximum heart rate achieved, exercise-induced angina, ST depression induced by exercise relative to rest, the slope of the peak exercise ST segment (1 for upsloping, 2 for flat, and 3 for down-sloping), number of major vessels (0-3) colored by fluoroscopy, thallium stress test (thal) (3 for normal, 6 for fixed defect, 7 for reversible defect) (UCI, 1988).

Statistical Measures

This section starts with a summary that describes the purpose of the prediction and represents the data source followed by four visuals: overall model performance, ROC curve, prediction distribution, and confusion matrix.

Overall Model Performance

Accuracy, precision, recall, f-score, and AUC measures are most frequently used to evaluate ML model performance. This section compares these measures for both training and testing data in a heatmap to show the strength and weaknesses of the model. The heatmap in figure 6 shows that the heart disease predicting model has very good performance in all measures. AUC and accuracy are the same, but F1-score is higher. Additionally, the model has a slightly higher precision value than recall.

The heatmap colors scale is between 0.5 and 1 for two reasons: first, the scale starts from 0.5 because a model with less than 0.5 accuracy is considered random. Second, the color scale based on the range of the data could misrepresent the differences. For instance, in the heart disease prediction model, all statistical measure values are high. In heatmaps, the darker the color the better the model performance. For instance, when using scale based on the range of the data as shown in figure 7, the accuracy color is white (the lowest color scale), which may give a perception that the model has a low accuracy value; however, the accuracy value is 0.8 which is considered a high value.

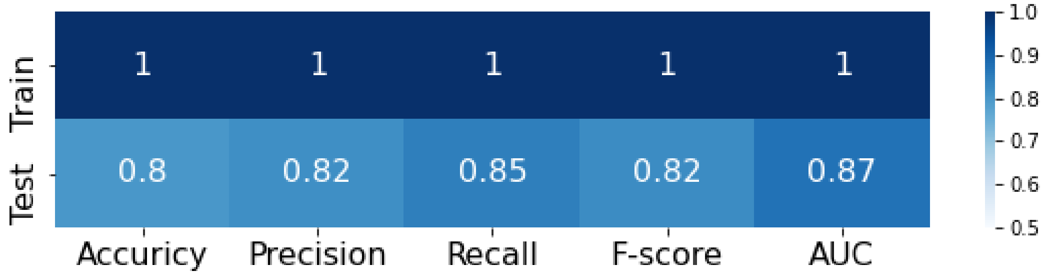


Figure 6 Example of heat map for the statistical measures using scale from 0.5 to 1

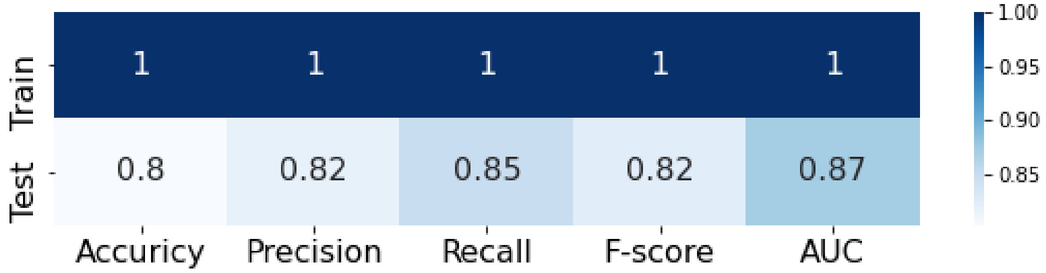


Figure 7 Example of heat map for the statistical measures using scale based on the data range

Receiver-Operator Curve (ROC)

The default classification threshold is 0.5 which means that if the model prediction score is greater than or equal to 0.5, the model predicts that the patient has heart disease and when the predicted percentage is less than 0.5 the model predicts that the patient does not have heart disease. However, a threshold of 0.5 is not always the best. The ROC shows all possible values of true positive rate (recall) and false positive rate as classification threshold varies. Figure 8 shows the curve for heart disease model, in the curve the red points represent the best selected threshold which is 0.55 using testing data and 0.66 using training data.

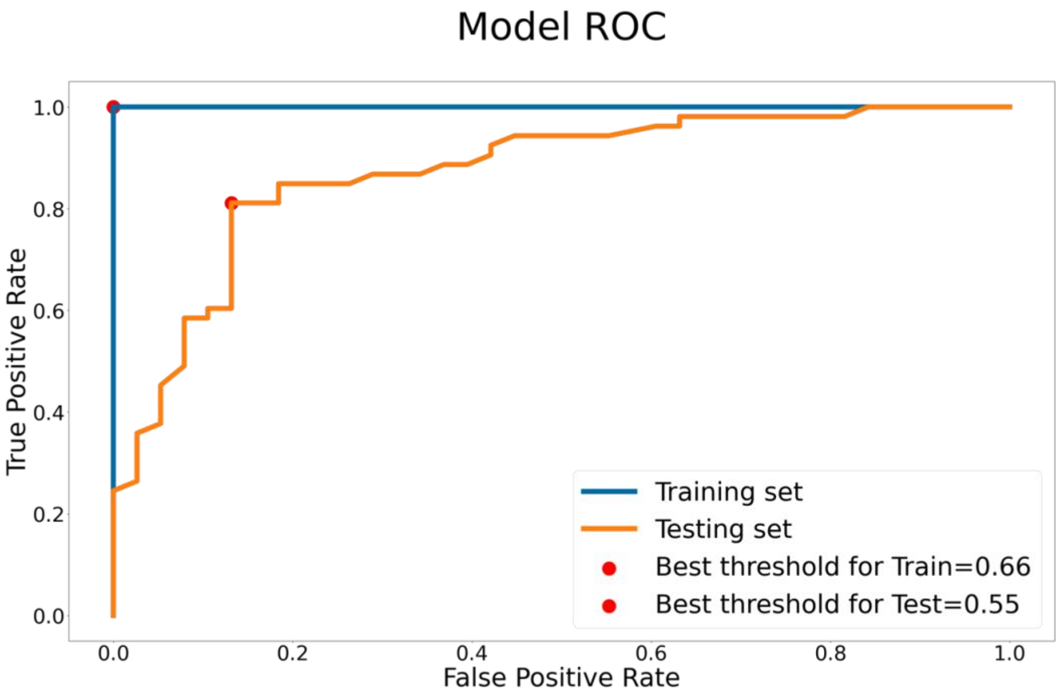


Figure 8 Example of ROC curve

Confusion Matrix

After identifying the best threshold, the confusion matrix is visualized to show classification performance. Usually the confusion matrix is visualized using a heatmap, yet Raymaekers et al. (2020) suggested using stacked mosaic plot that adds the area perspective to show the proportion of the number of objects in each class. This additional information indicates if the data is skewed or not. The mosaic plot shows the actual classes on the horizontal axis and the predicted classes on the vertical axis. Figure 9 shows an example of a stacked mosaic plot for the confusion matrix with two classes. As seen below, the number of misclassified cases in each class are the same but the data set has higher number of patients with heart disease compared to the number of patients without heart disease. For the training data the accuracy is 100%, which indicates that the model over fitted the training data.

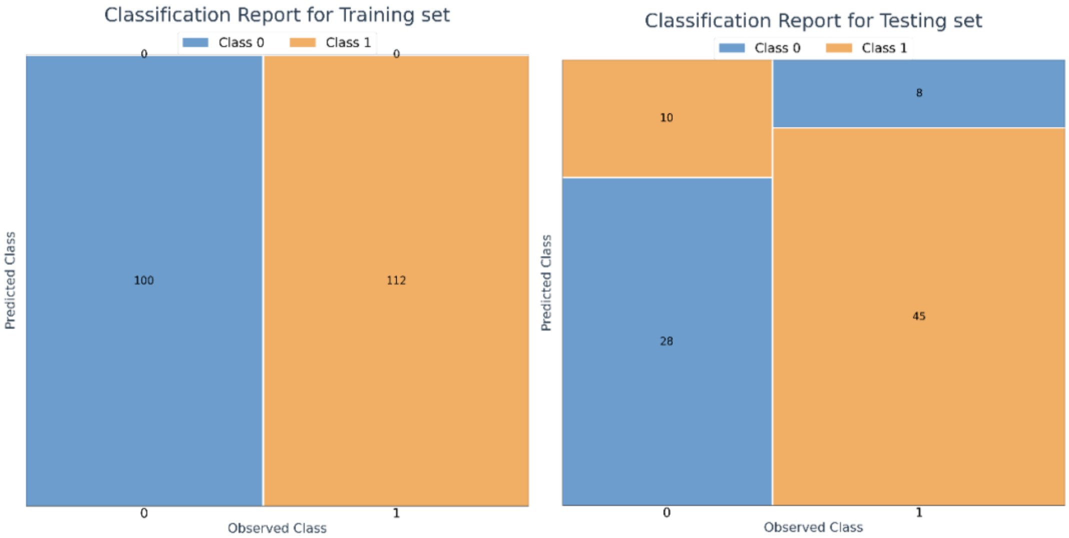


Figure 9 Example of stacked mosaic plot for confusion matrix result

Prediction Distribution

The model level of confidence is shown using the prediction distribution using bar chart with color representing the actual classes. A good model will have more cases near the 0 and 1, and less cases in the middle near the threshold. The larger number of cases near the threshold means that the model is not very confident about the decision. Figure

10 shows the prediction distribution for the heart disease prediction model. The training plot shows that the classes are clearly split at 0.5. The prediction percentages for patients with heart disease are between 0.6 and on; while for patients without heart disease, the prediction percentages are between 0 and 0.4. However, in the testing set there are some overlaps between 0.4 and 0.6 prediction percentages. Additionally, most of the patients with heart disease were predicted correctly as the number of cases between 0.8 and 1 is high.

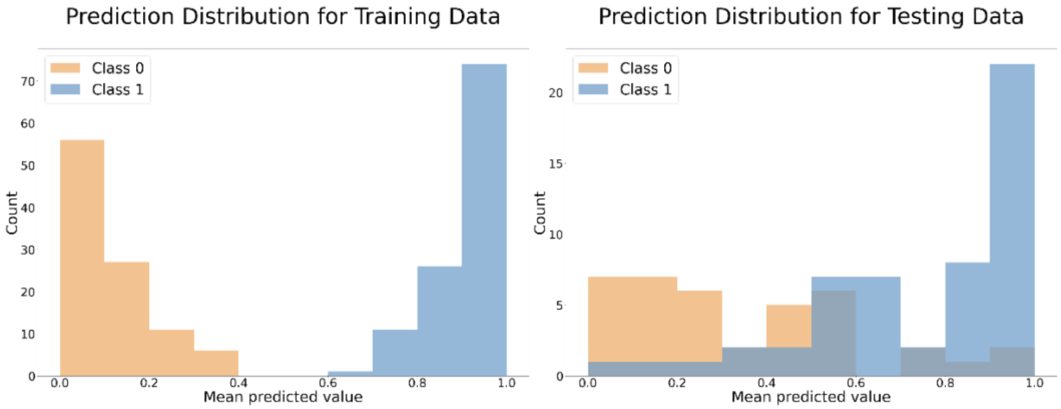


Figure 10 Example of prediction distribution for training and testing datasets

Features' Importance

Understanding the relationship between the attributes and the output is important for decision makers to understand the model performance because it gives some explanation of the model decision. This section visualizes the features' importance by the following visuals: correlation heatmap, LASSO, random forest, and premutation bar chart, learning curve based on number of cases using line chart, and learning curve based on number of features using line chart. When the number of attributes is large it is hard to display them all; therefore, the number of displayed attributes is limited to the top 20 to avoid cluttering. The top 20 attributes were selected based on the average of lasso, random forest, premutation scores after normalizing them between 0 and 1.

Correlation Plot

The first step is to represent the correlation between the features to show how they are related to each other and to the dependent attribute. Figure 11 shows an example of the correlation graph using heatmap. The first column is larger than the others because the relationship between all independent attributes and the output attribute is more important than the relationship between all attributes.

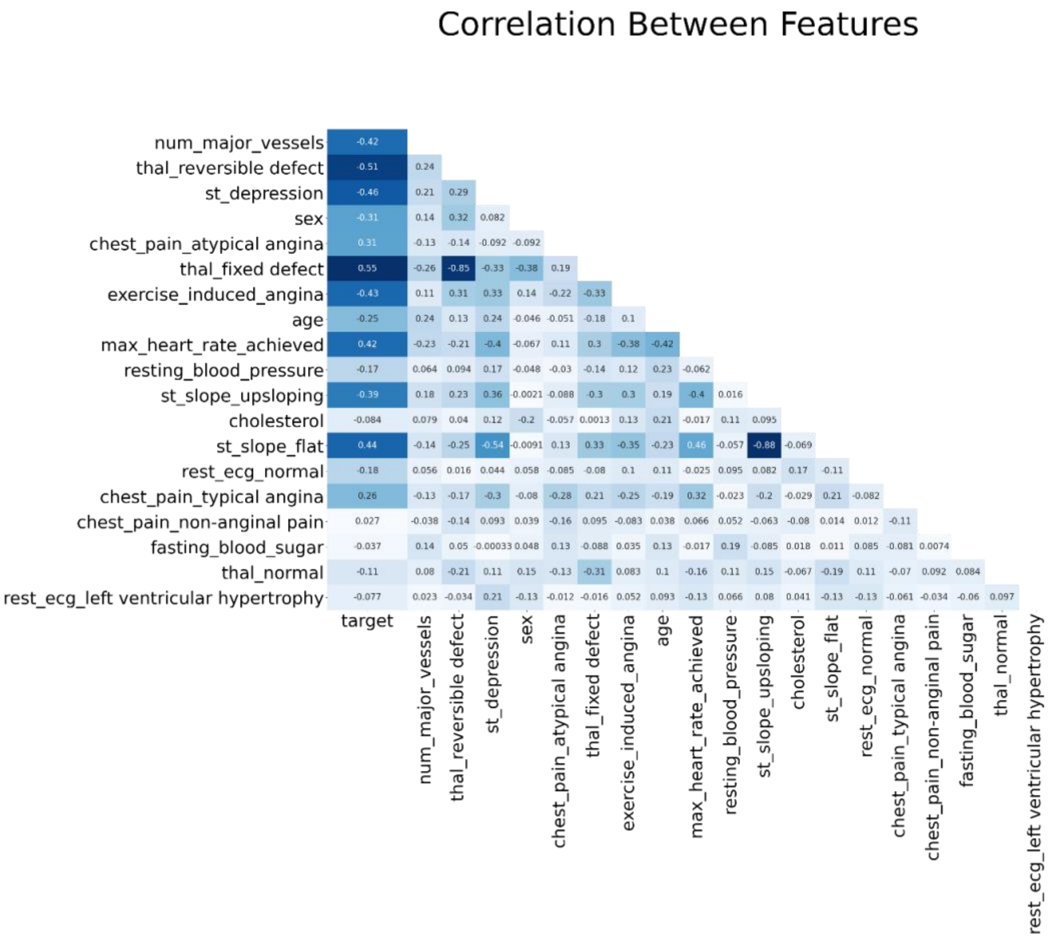


Figure 11 Example of correlation between attributes

Lasso, Random Forest, and Premutation

Correlation is based on linear relationship and did not take the model into consideration; therefore, features selection techniques are plotted to explain the feature importance. Usually, the features selection techniques are used to reduce the number of features; however, since the model is already built at this stage and the features are already chosen, the purpose is to understand the importance of the features. The selected supervised feature selection methods are Lasso, random forest (embedded methods), and permutation (wrapper method). These scores were displayed using a vertical bar chart to show the difference between each method judgment. Figure 12 shows that all methods agree that number of major vessels is the most important feature. However, random forest gives a high score for age compared to the other methods.

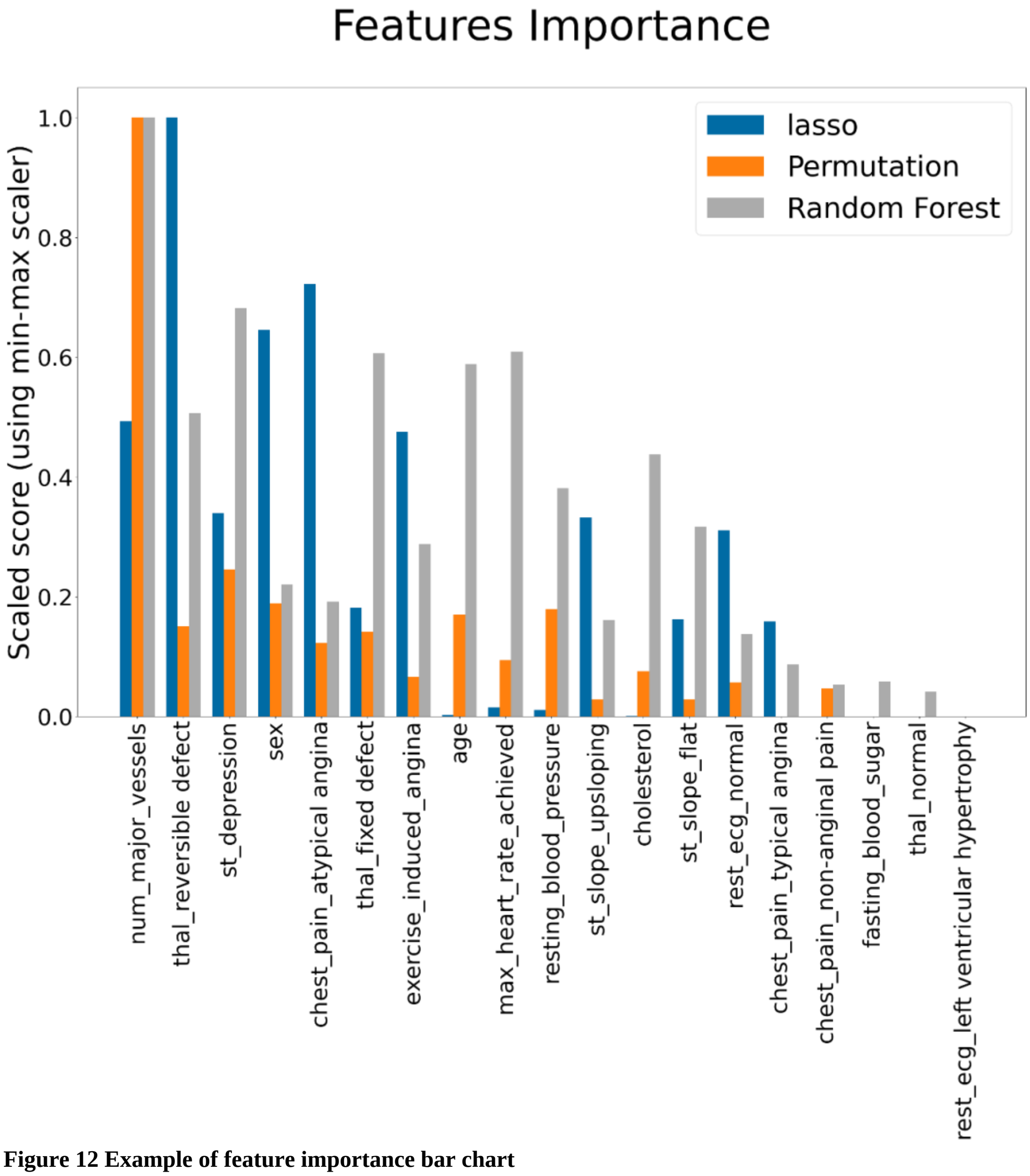


Figure 12 Example of feature importance bar chart

Learning Curve

The two types of learning curves used here represent the relationship between number of cases with the model AUC and number of attributes with the model AUC. Figure 13 shows the first learning curve for number of heart disease model cases. The testing score line stop increasing after 90. For the number attributes curve in figure 14, the AUC slightly improved after 13 attributes. In deep learning community, the term learning curve is also used to visualize convergence of learning as a neural network is learned. This meaning of the term is not used here.

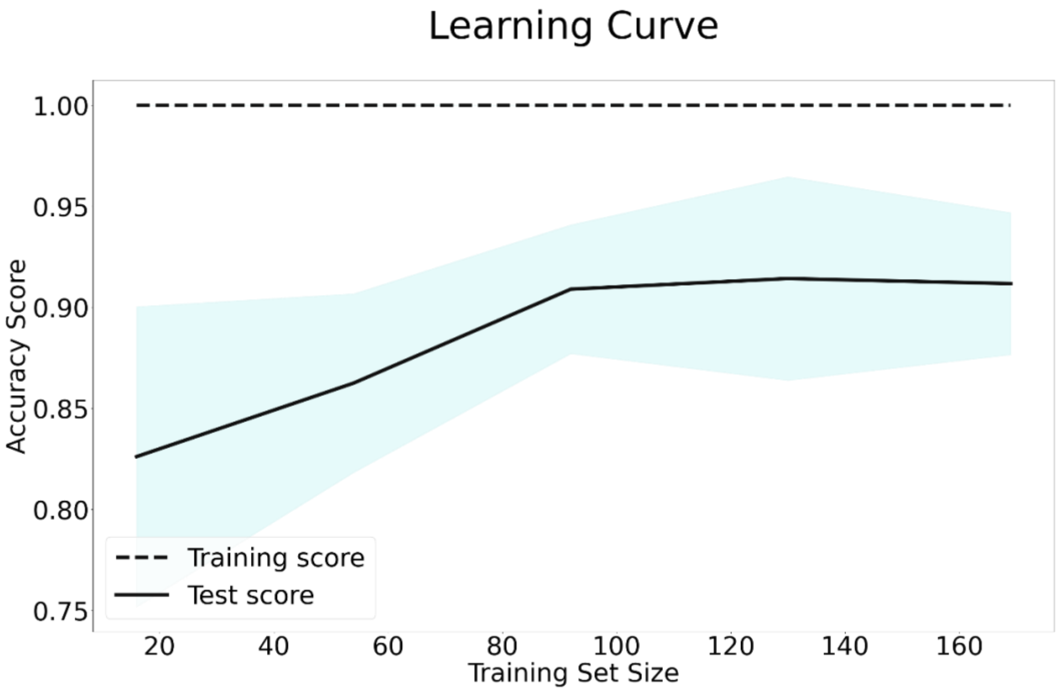


Figure 13 Example of learning curve for number of cases

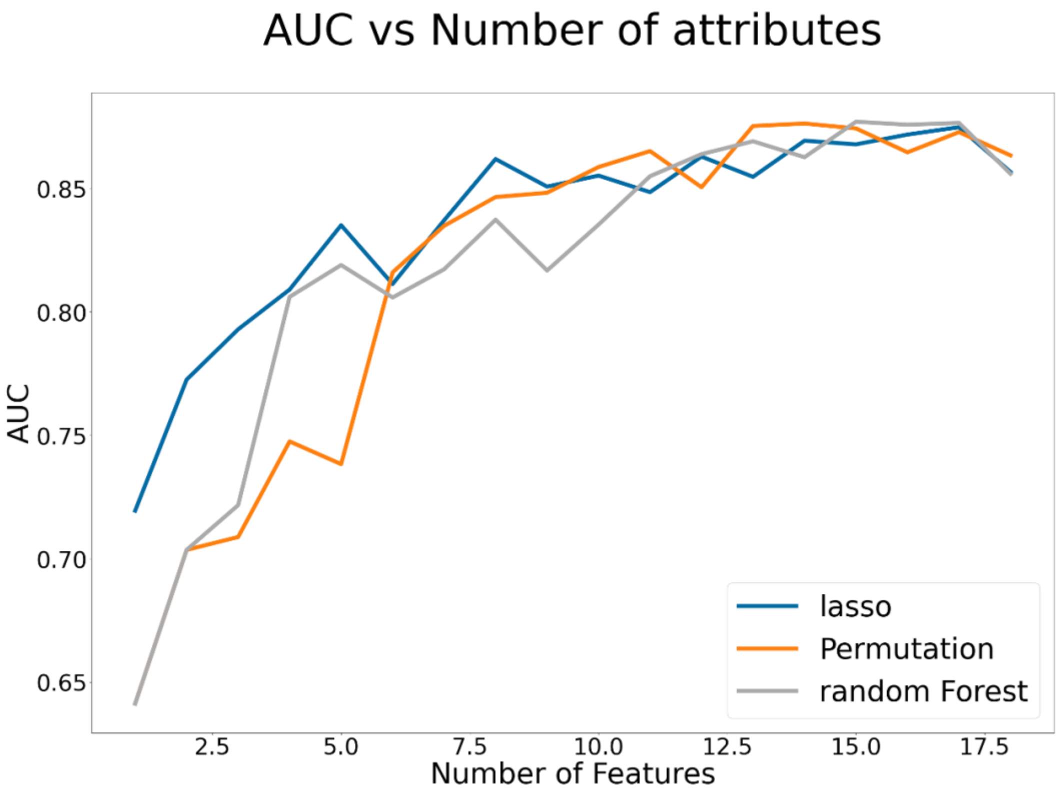
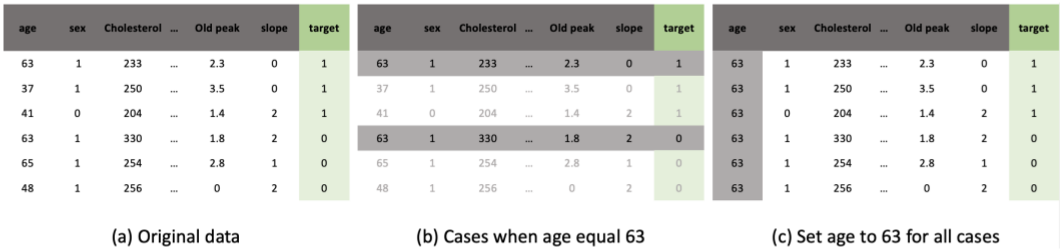


Figure 14 Example of learning curve for number of attributes

Sensitivity Analysis

The purpose of this section is to identify the relationship between an attribute and the model prediction. This analysis is done for the top 20 predictors only. Using a selection button, the dashboard visualizes the impact of a single input attribute into the output attribute using sensitivity measures. These types of plots are sometime referred to as partial dependency plots. The type of plots depends on the type of the data; therefore, the first step is to identify the categorical and numeric attributes using a default threshold of 10. If the number of unique values for an attribute is less than 10 then the attribute is identified as categorical otherwise the attribute is identified a numeric. After selecting the attribute, four visuals are displayed: the distribution; two partial dependence plots: the mean prediction based on the selected attribute, and the mean prediction when the attribute value is fixed; and the difference between the original AUC and the AUC when the selected attribute changes slightly. For some of the visuals, a random dataset is needed to check the attribute behavior regardless of the correlation with other attributes. For numeric attributes, the random data has the same minimum, maximum, mean, and standard deviation as the original data, while for the values for the categorical attributes have the same probabilities as the original data.

Partial Dependence Plots (PDP) show the effect of the selected attribute on the prediction. (Jerome H. Friedman, 2001). The Prospector system uses this concept to examine the impact of an attribute by fixing the value of the selected attribute while keeping all other attributes as they were (Krause et al., 2016). However, this approach ignores the effect of the interaction between other attributes. Wojtusiak et al. paper added the results using randomly generated data to show the interaction between the selected attribute and the predictions only (Wojtusiak et al., 2018). The dashboard shows the PDP in two plots. First, for each unique value i in the selected attribute X: the first plot selects the cases with the selected value (where X=i). In the dashboard this plot is referred to as "Mean Prediction for X". In the second plot, all values in the selected attribute (column) X is set to i. In the dashboard this plot is referred to as "Mean prediction based on fixed values for X". Figure 15 shows an example when X is age and i is 63. Figure 15.a shows the original data, figure 15.b shows the selected cases for Mean prediction for age, and figure 15.c shows the cases for Mean prediction based on fixed values for X.



age	sex	Cholesterol	...	Old peak	slope	target
63	1	233	...	2.3	0	1
37	1	250	...	3.5	0	1
41	0	204	...	1.4	2	1
63	1	330	...	1.8	2	0
65	1	254	...	2.8	1	0
48	1	256	...	0	2	0

(a) Original data

age	sex	Cholesterol	...	Old peak	slope	target
63	1	233	...	2.3	0	1
37	1	250	...	3.5	0	1
41	0	204	...	1.4	2	1
63	1	330	...	1.8	2	0
65	1	254	...	2.8	1	0
48	1	256	...	0	2	0

(b) Cases when age equal 63

age	sex	Cholesterol	...	Old peak	slope	target
63	1	233	...	2.3	0	1
63	1	250	...	3.5	0	1
63	0	204	...	1.4	2	1
63	1	330	...	1.8	2	0
63	1	254	...	2.8	1	0
63	1	256	...	0	2	0

(c) Set age to 63 for all cases

Figure 15 An illustration of how partial dependence is computed for age 63

Distribution Plot

The distribution plot provides a general idea about the attribute trend for testing, training, and random data. For numerical attributes the distribution is shown using a line plot and colored by the data type. Figure 16 shows the distribution of age attribute for the heart disease data set. Since the data set is small, the testing data did not follow the training data trend. In the training and random data, the peak of number on patients is in the late 50s. For categorical attributes, the distribution is shown using bar chart. Figure 17 shows the destruction of the number of major vessels. Most patients had value of 0 and very small number of patients had value of 4.

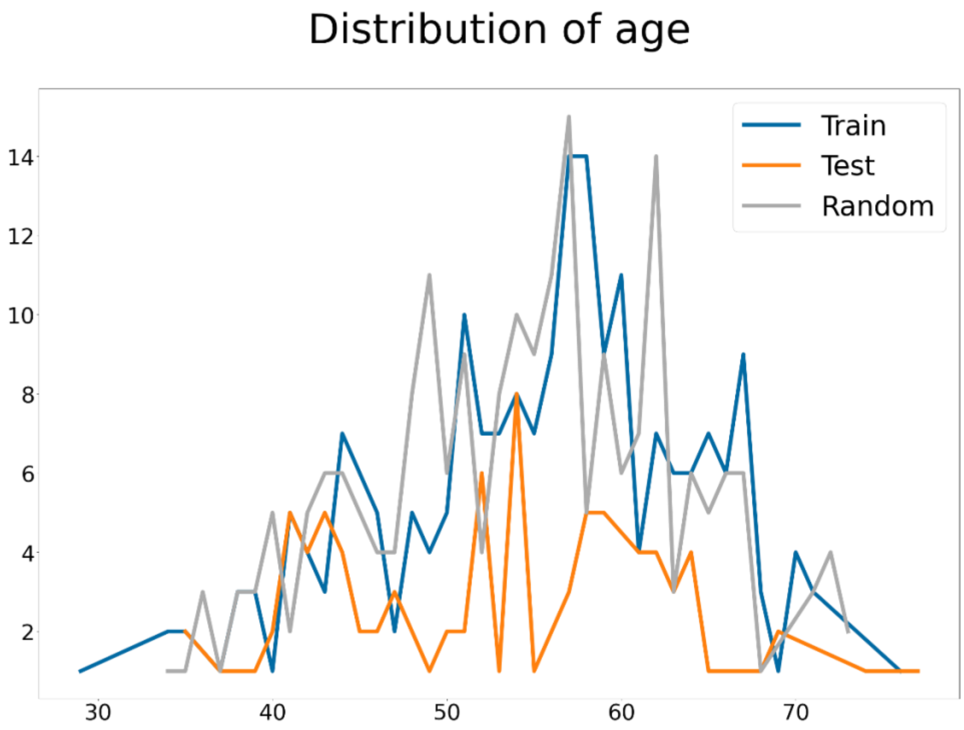


Figure 16 Example of distribution plot for age

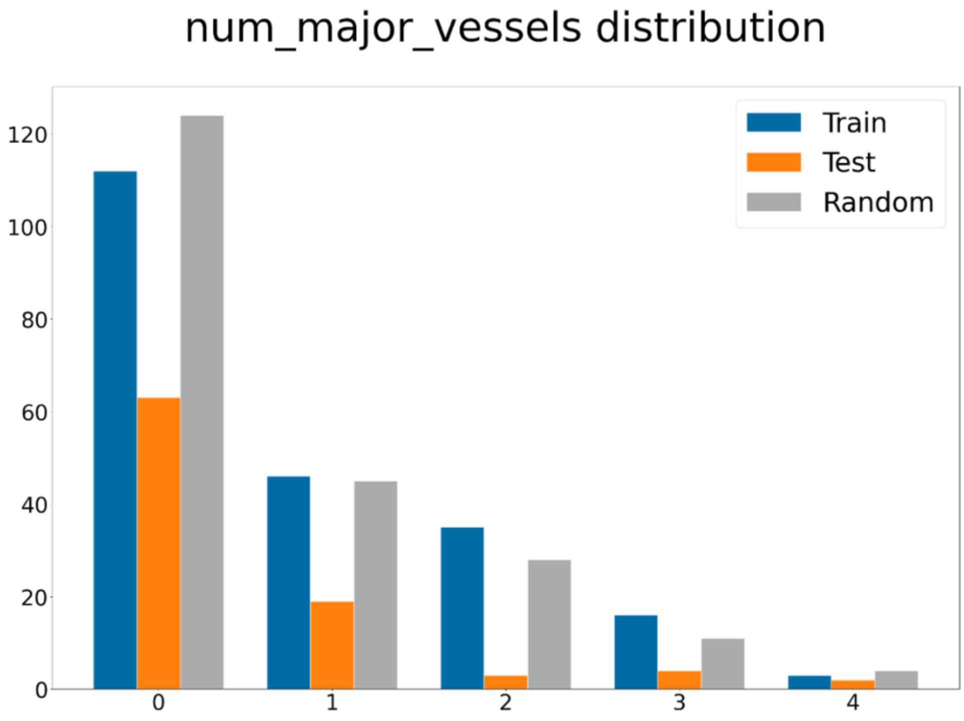


Figure 17 Example of destruction of the number of major vessels

Mean Prediction based on the selected attribute values

For numerical attribute, the plot shows the predictions' means per each value of the selected attribute using training and testing data. The horizontal axis represents the attributes values, and the vertical axis represents the predictions' means. The training and testing trend shows the model behavior for each value in the selected attribute. Figure 18 shows the predictions' means based on age using the heart disease data. There is no clear trend between age and having heart disease. The training and testing data trend shows that there is a drop of the AUC percentage around age 60.

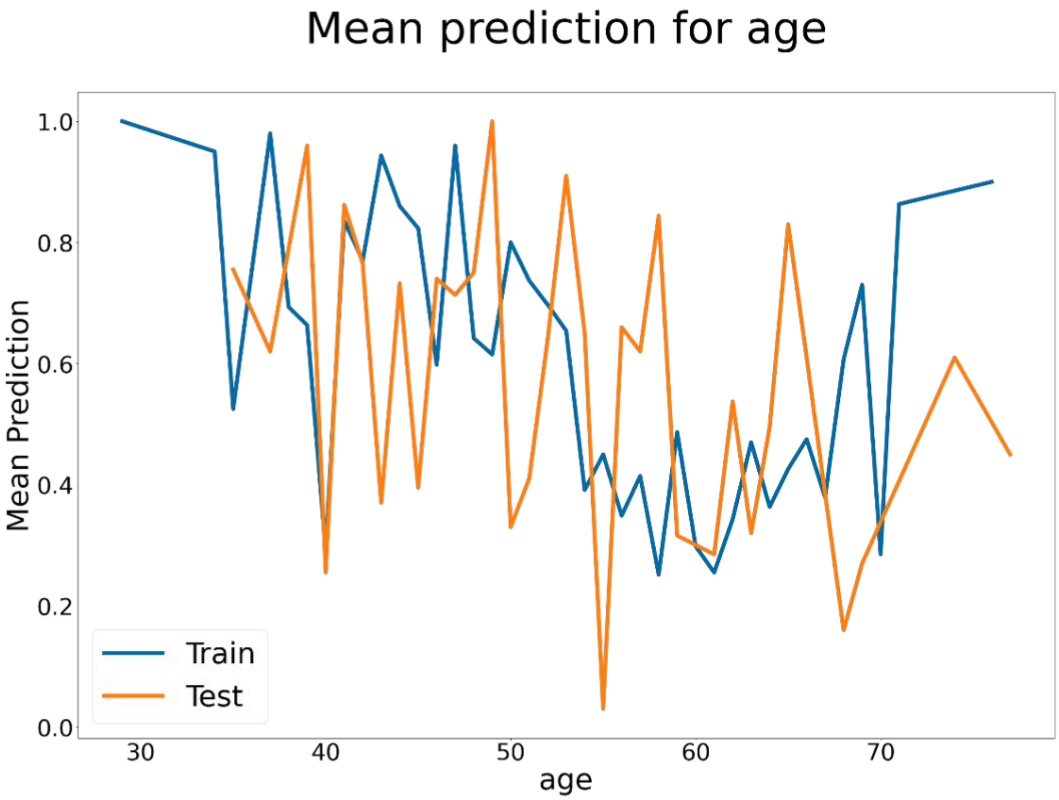


Figure 18 Example of mean prediction based on age

For categorical variables, the prediction distribution is visualized for each value of the selected attribute. Figure 19 shows the prediction distribution for the number of major vessels. From the training data, the number of vessels is positively correlated with having heart disease when it's value equal to 0 and negatively correlated with heart disease when its value equal to 1, 2, or 3.

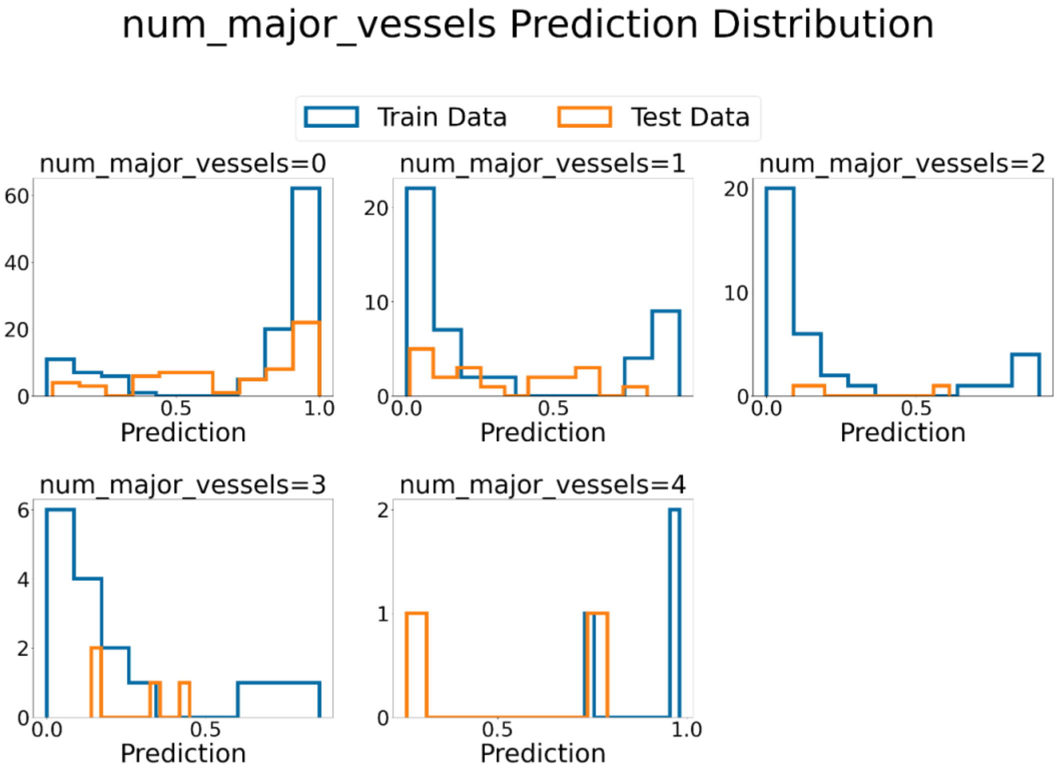


Figure 19 Example of prediction distribution for number of major vessels

Mean Prediction Based on Fixed Values

To check the effect of an attribute ignoring the interaction with other attributes, this work uses the method introduced by Wojtusiak et al. (2018) when examining models for predicting 30-day post hospitalization mortality. For numeric attributes, the selected attribute values are set to a fixed value, then the mean AUC is calculated. This calculation is done for all unique values of the selected attribute as a fixed value. The result of the random dataset shows the effect of that attribute regardless of all other attributes changes (Wojtusiak et al., 2018). Figure 20 shows the mean prediction when age is fixed for all cases. Training, testing, and random data has the same trend. The plot shows that there is a correlation between age and having heart disease. Patients at age 60, have the lowest AUC probability of having heart disease. While this drop needs more investigation, explaining the trend is out of the dashboard scope.

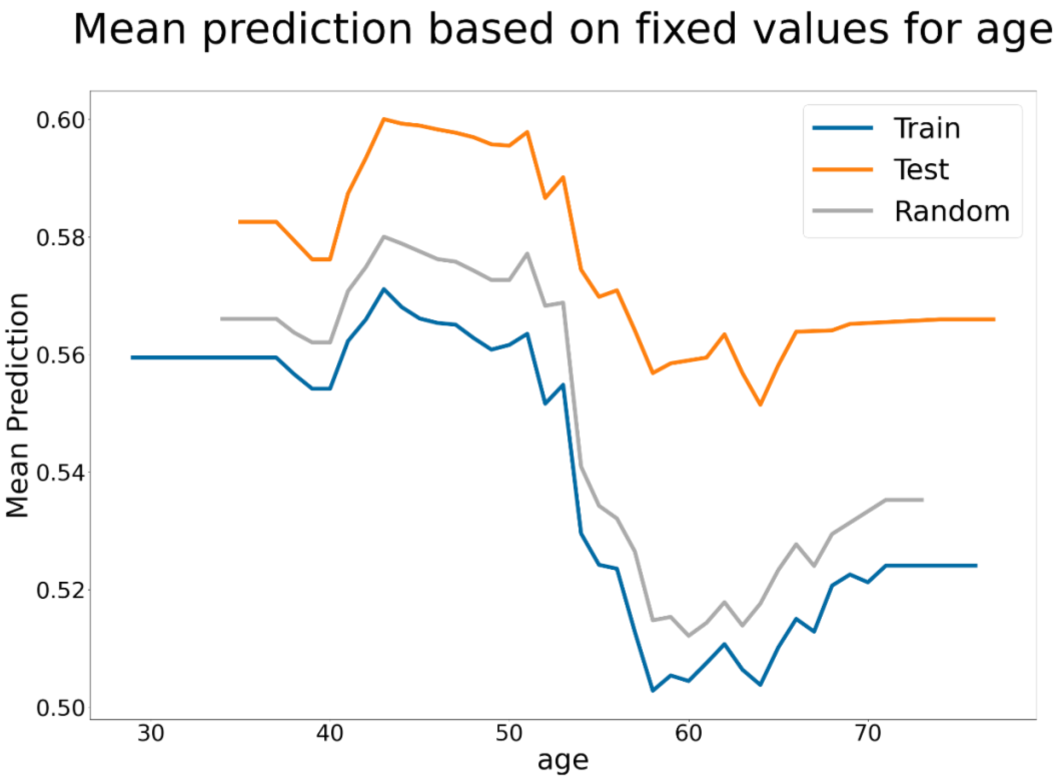


Figure 20 Mean prediction when age is fixed for all cases

For categorical attributes, all data for each value for the selected attribute, all data is set to that value and the prediction distribution is visualized using histogram plot. Figure 21 shows the prediction distribution for number of major vessels. When the number of major vessels is set to 0 for all patients, the data is skewed to the lift. For the other types the training data were skewed but the random data were symmetric. Therefore, trend might be cause by the correlation between number of major vessels and other attributes.

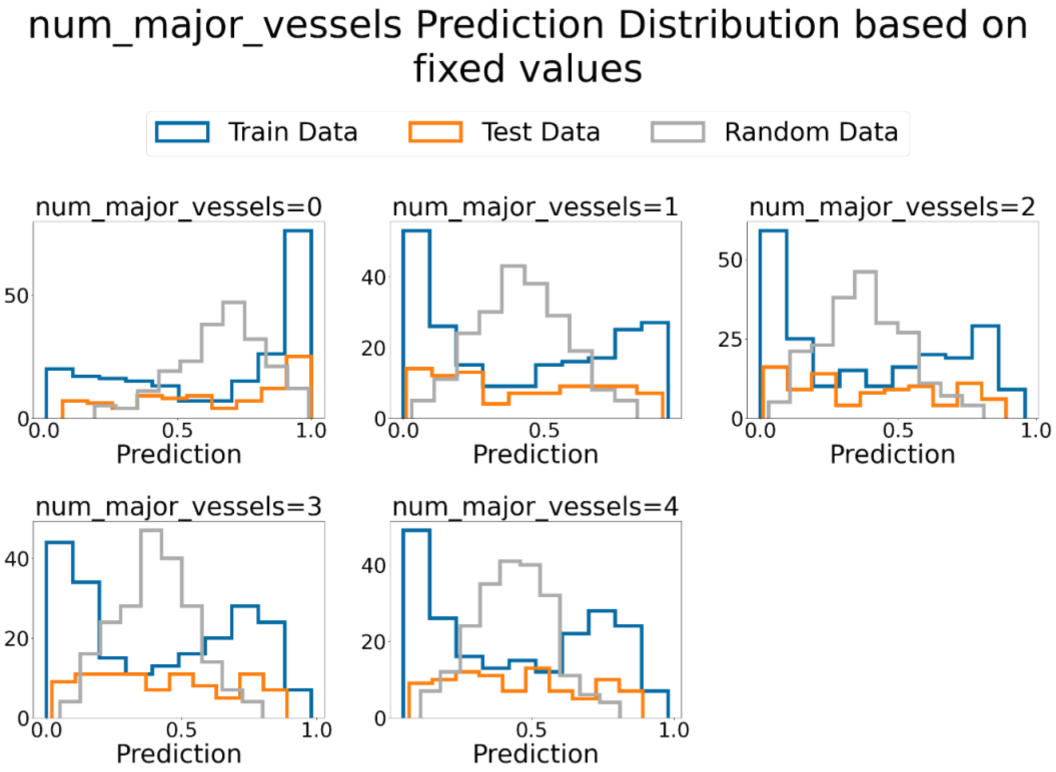


Figure 21 Chest pain prediction distribution when number of major vessels values fixed

Original AUC vs. AUC when the selected attribute change slightly

The prediction should not change dramatically when the attribute value changes slightly. For example, in the prediction of heart disease model, if the patient age increase or decrease by two years, the change percentage of getting heart disease should not change significantly. To ensure that the model is stable, the prediction comparison is visualized for numeric attributes only. This displayed only for the numerical attributes.

For numeric attributes, the data changed by adding or subtracting the standard deviation. The closer the data to the diagonal line, the less sensitive the model is to the small change. Figure 22 shows the age AUC vs. Age minus/plus standard deviation using test data. Most data are around the diagonal line; therefore, the model is not sensitive to small changes to age.

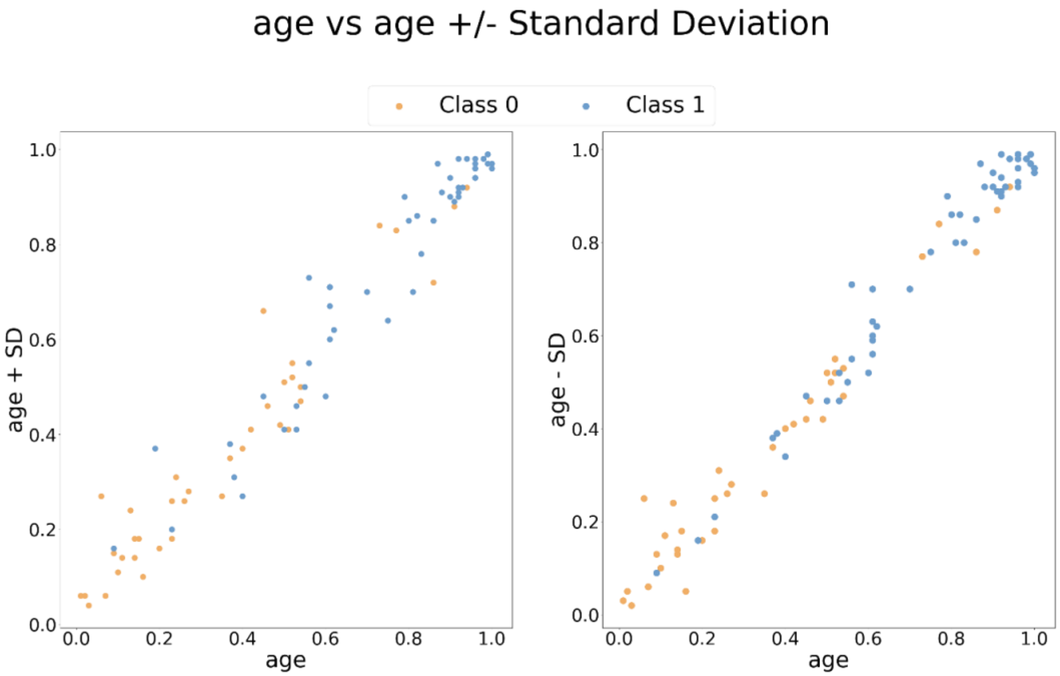
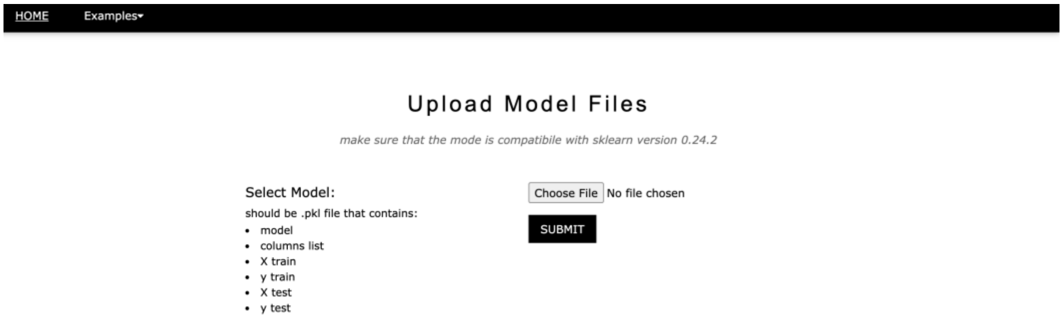


Figure 22 Age AUC vs. Age minus/plus standard deviation

CASE STUDIES

The purpose of the dashboard is to visualize performance of machine learning (ML) models. A website was built to allow users to upload their models along with data for visualization. The home page of the website takes the models files: the model, input data (X train and X test), output data (Y train and Y test), and the columns list. Then the website will display the dashboard. Figure 23 shows the screenshot of the home page. To test the dashboard, three models were generated using heart disease, claims, and covid-19 datasets.



Upload Model Files

make sure that the mode is compatibile with sklearn version 0.24.2

Select Model:

should be .pkl file that contains:

- model
- columns list
- X train
- y train
- X test
- y test

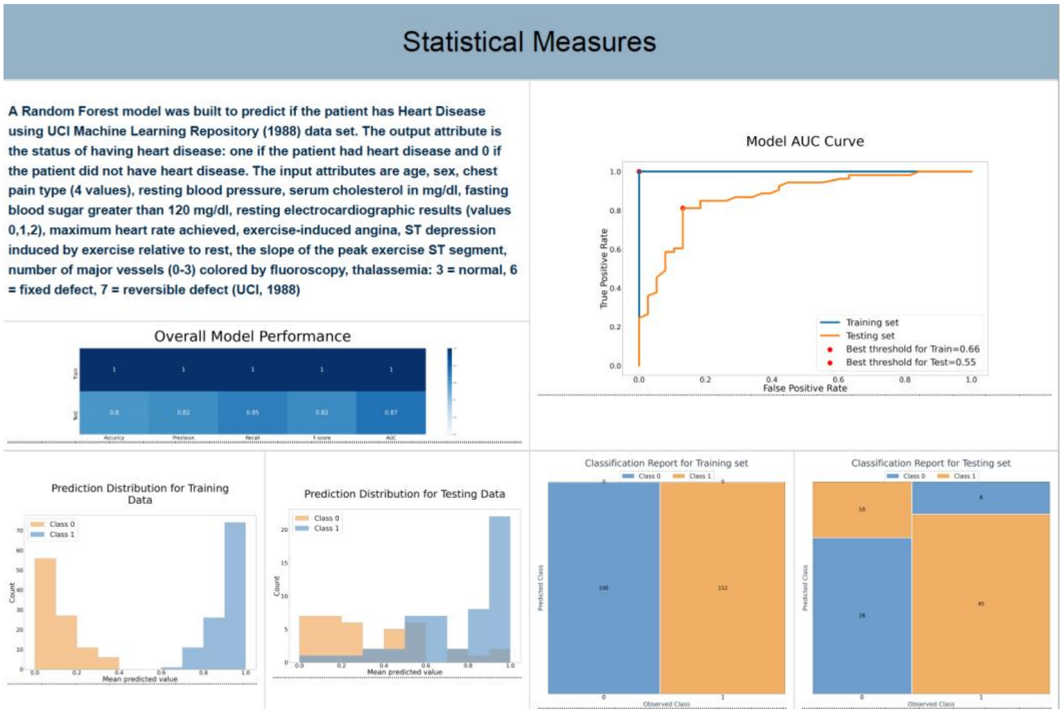
Figure 23 Home page of the model visualization tool

Heart Disease Data

The first model used to test the constructed system was a random forest model used to predict if the patient has heart disease using UCI machine learning repository. Some of the results of this model were used in the method section to explain the dashboard visuals.

Statistical Measures Section

Figure 24 shows the statistical measures section for heart disease dataset. The model has high values for accuracy, precision, recall, f-score, and AUC. From the ROC curve, the best classification threshold is 0.55. For the prediction distribution, most of the patients with heart disease had prediction value near the one. Also, the classification report shows that the model overfit the training data. For the testing data, the model classification for patients with heart disease is better than the classification for patients without heart disease. Also, the data has more cases for patients with heart disease than patients without heart disease.



Statistical Measures

A Random Forest model was built to predict if the patient has Heart Disease using UCI Machine Learning Repository (1988) data set. The output attribute is the status of having heart disease: one if the patient had heart disease and 0 if the patient did not have heart disease. The input attributes are age, sex, chest pain type (4 values), resting blood pressure, serum cholesterol in mg/dl, fasting blood sugar greater than 120 mg/dl, resting electrocardiographic results (values 0,1,2), maximum heart rate achieved, exercise-induced angina, ST depression induced by exercise relative to rest, the slope of the peak exercise ST segment, number of major vessels (0-3) colored by fluoroscopy, thalassemia: 3 = normal, 6 = fixed defect, 7 = reversible defect (UCI, 1988)

Overall Model Performance

	Accuracy	Precision	Recall	F-score	AUC
Train	1	1	1	1	1
Test	0.8	0.82	0.85	0.82	0.87

Model AUC Curve
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Training set
Testing set
Best threshold for Train=0.66
Best threshold for Test=0.55
Prediction Distribution for Training Data
Class 0
Class 1
Count
Mean predicted value
Prediction Distribution for Testing Data
Classification Report for Training set
Class 0
Class 1
Predicted Class
Observed Class
0
100
112
0
1
Classification Report for Testing set
10
8
28
45

Figure 24 Statistical measures section for heart disease dataset

Features Importance Section

Figure 25 shows the features' importance section for heart disease dataset. Since the total number of attributes is less than 20, all attributes are displayed.

Fixed defect on thallium stress test (thal) and reversible defect on thal are the two features highly correlated with having heart disease. Additionally, fixed defect on thal is also highly correlated with the number of major vessels and ST slop upsloping is highly correlated with flat on ST slope. From the feature’s importance plot, the following features are the top predictors based on the average score of all methods scores: number of major vessels, reversible defect on thal, ST depression, and sex.

From the learning curves, the AUC stopped improving after 85 cases. While for the number of features the AUC was increasing till using all number of features.

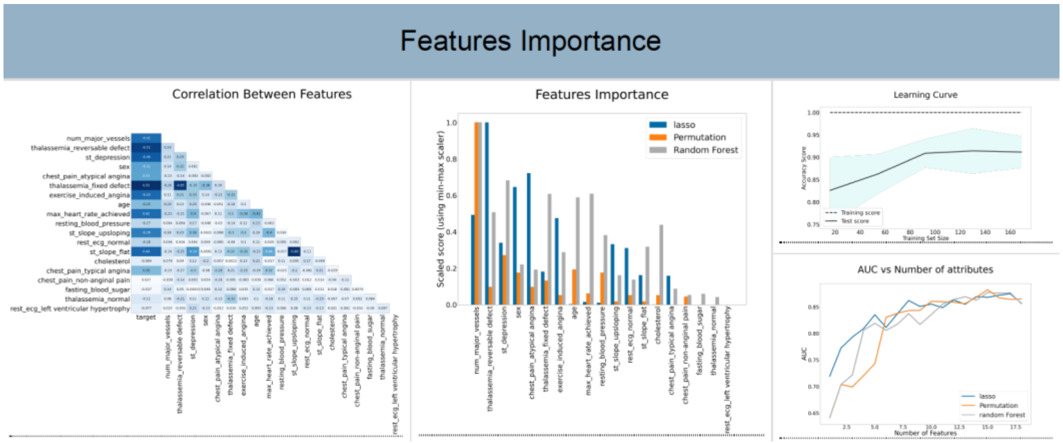


Figure 25 Features' importance for heart disease dataset

Sensitivity Analysis Section

For each of the categorical attributes the sensitivity analysis results are:

Number of major vessels. Figure 26 shows the sensitivity analysis for the number of major vessels in the heart disease dataset. The distribution shows that most cases have the value of 0 and less than 10 cases have the value of 4. For the prediction distribution, most of the cases with value 0 correlated with having heart disease and most of the cases with a value of 1, 2, and 3 correlated with not having heart disease. However, from the fixed value prediction distribution, using random data results in normal distribution for all values except for 0. The correlation shown using the training and testing data could be caused by the interaction with other variables.

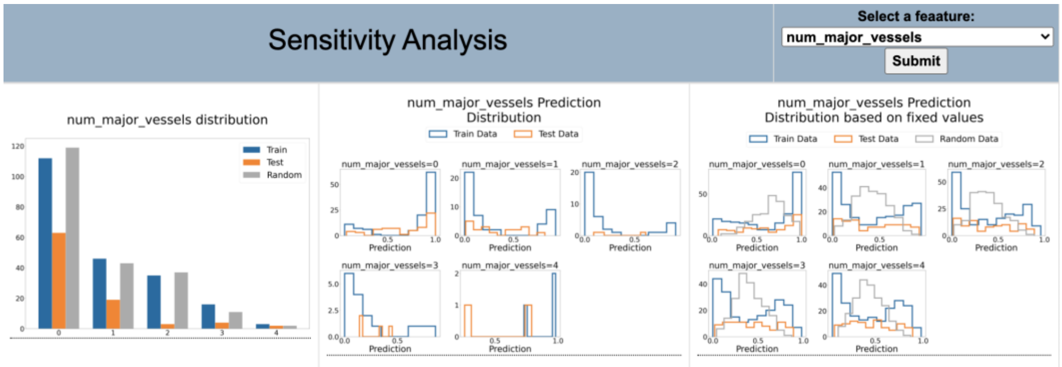
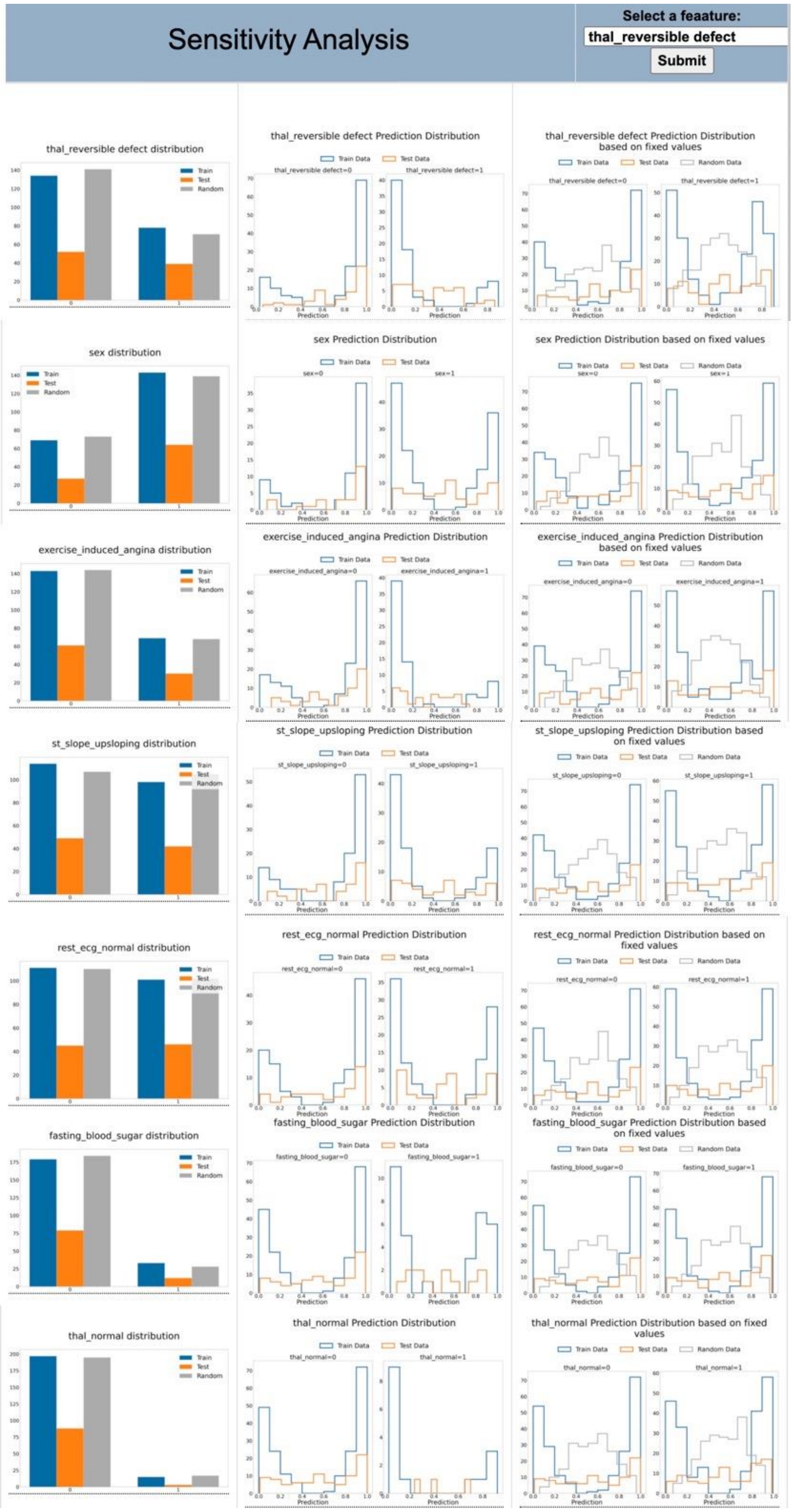


Figure 26 Sensitivity analysis for the number of major vessels in the heart disease dataset

For all the categorical attributes the random data prediction distribution is normal distribution, which means that the attributes by themselves does not have a prediction trend. However, when the attribute is not independent from other attributes the following trends found.

Figure 27 shows the sensitivity analysis for the attributes that have negative correlation with prediction probability. In other words, the peak prediction probability when the attribute present is at 0. This trend is shown in the following attributes: reversible defect thal, sex, exercise induced angina, ST slope upsloping, rest electrocardiogram (ECG) normal, fasting blood and sugar, normal thal



Sensitivity Analysis

Select a feaature:

thal_reversible defect

Submit

thal_reversible defect distribution
Train
Test
Random
thal_reversible defect Prediction Distribution
Train Data
Test Data
thal_reversible defect=0
thal_reversible defect=1
Prediction
thal_reversible defect Prediction Distribution based on fixed values
Random Data
sex distribution
sex Prediction Distribution
sex=0
sex=1
sex Prediction Distribution based on fixed values
exercise_induced_angina distribution
exercise_induced_angina Prediction Distribution
exercise_induced_angina=0
exercise_induced_angina=1
exercise_induced_angina Prediction Distribution based on fixed values
st_slope_upsloping distribution
st_slope_upsloping Prediction Distribution
st_slope_upsloping=0
st_slope_upsloping=1
st_slope_upsloping Prediction Distribution based on fixed values
rest_ecg_normal distribution
rest_ecg_normal Prediction Distribution
rest_ecg_normal=0
rest_ecg_normal=1
rest_ecg_normal Prediction Distribution based on fixed values
fasting_blood_sugar distribution
fasting_blood_sugar Prediction Distribution
fasting_blood_sugar=0
fasting_blood_sugar=1
fasting_blood_sugar Prediction Distribution based on fixed values
thal_normal distribution
thal_normal Prediction Distribution
thal_normal=0
thal_normal=1
thal_normal Prediction Distribution based on fixed values

Figure 27 Sensitivity analysis for heart disease attributes with negative prediction correlation.

Figure 28 shows the sensitivity analysis for the attributes that have positive correlation with prediction probability. In other words, the peak prediction probability when the attribute present is at 1. This trend is shown in the following attributes: fixed thal, ST slope flat, chest pain atypical angina, and chest pain typical angina.

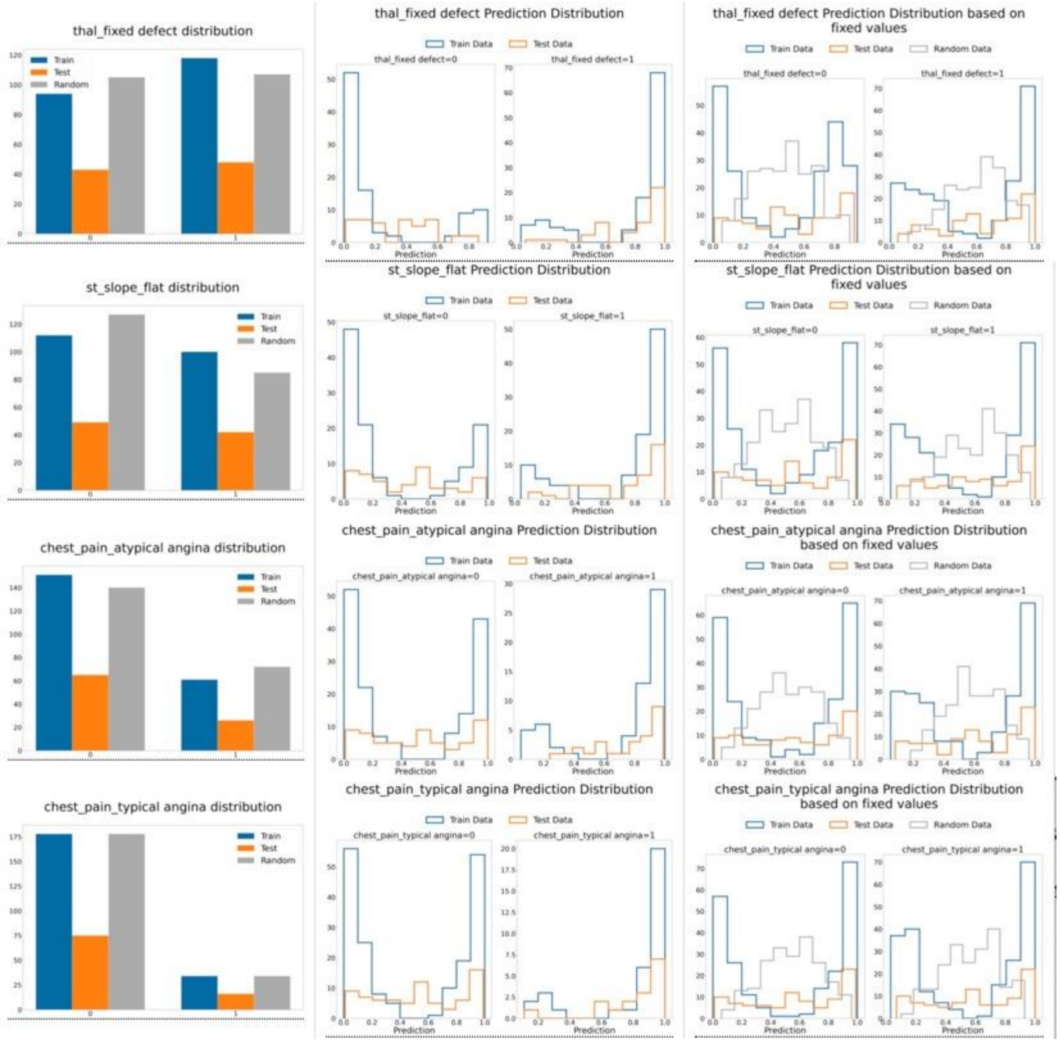


Figure 28 sensitivity analysis for heart disease attributes with positive prediction correlation.

Figure 29 shows that there is no clear trend for chest pain non anginal pain attribute and the number of cases when the attribute present is very small.

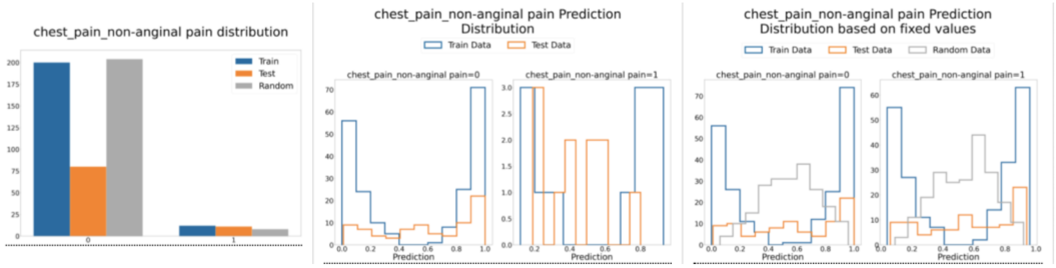
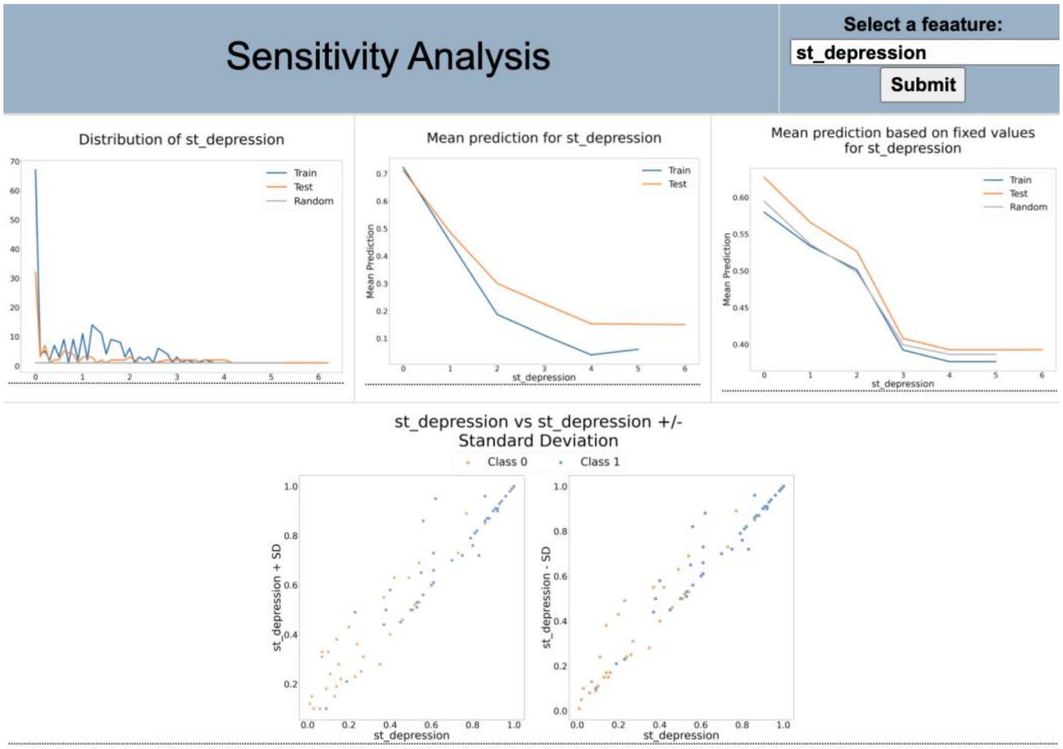


Figure 29 Sensitivity analysis for chest pain non anginal pain

For the rest ECG left ventricular hypertrophy, all cases have value of 0; therefore, there is no need to present any sensitivity analysis. For each of the numerical attributes the sensitivity analysis results are:

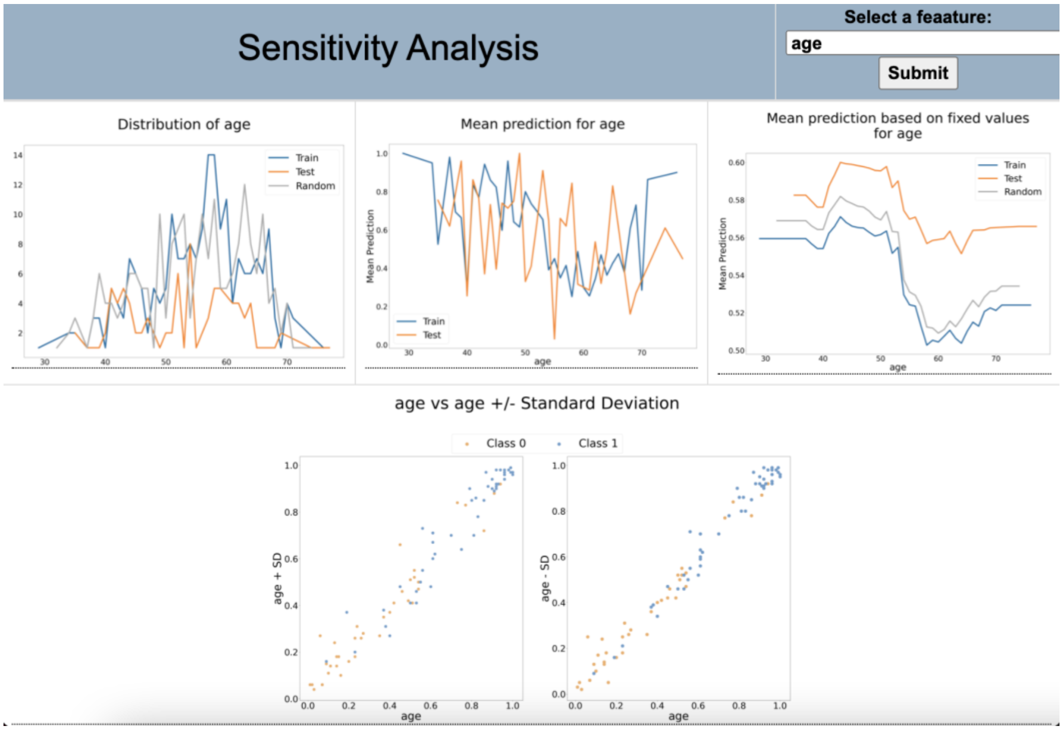
ST Depression. Most of the data has a value of 0. The mean prediction distribution for both actual data and fixed data decreased when the ST depression value increased. Also, the model is not sensitive towards a small changing in the ST depression value. Figure 30 shows sensitivity analysis for ST depression in the heart disease dataset.



Select a feaature: st_depression

Figure 30 sensitivity analysis for ST depression in the heart disease dataset

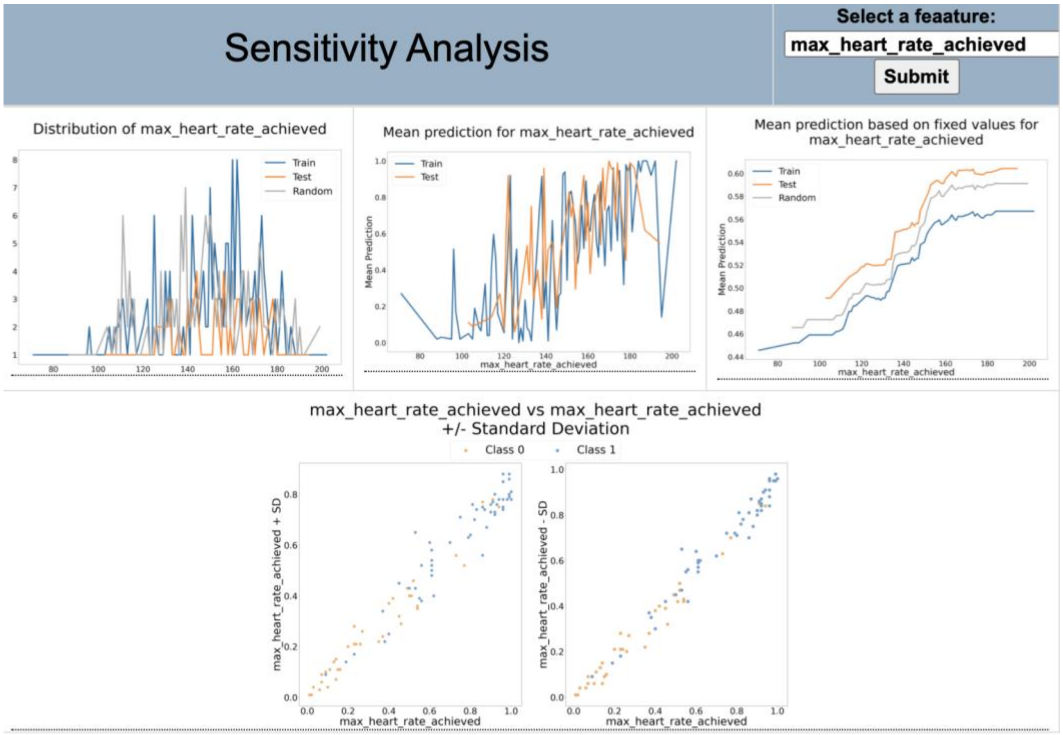
Age. The highest number of cases is around age 60. There is no clear trend in the mean prediction plot. However, using fixed values shows a decrease in the prediction around age 60 for all datasets (train, test, and random). Finally, the model is not sensitive for small changes in age. Figure 31 shows the sensitivity analysis for age in the heart disease dataset.


Sensitivity Analysis

Select a feaature: age

Figure 31 The sensitivity analysis for age in the heart disease dataset

Maximum Heart Rate Achieved. Figure 32 shows the sensitivity analysis for maximum heart rate in the heart disease dataset. The mean prediction using actual and fixed values shows the same trend. From both mean prediction and mean prediction based on fixed values, the mean prediction increased when the maximum heart rate value increased. Lastly, the model is not sensitive for small changes in maximum heart rate value since the model predictions did not change significantly after adding or subtracting the standard deviation.



Sensitivity Analysis

Select a feaature: max_heart_rate_achieved

Submit
Distribution of max_heart_rate_achieved
Train
Test
Random
1
2
3
4
5
6
7
8
80
100
120
140
160
180
200
Mean prediction for max_heart_rate_achieved
Train
Test
Mean Prediction
0.0
0.2
0.4
0.6
0.8
1.0
max_heart_rate_achieved
Mean prediction based on fixed values for max_heart_rate_achieved
Train
Test
Random
Mean Prediction
0.44
0.46
0.48
0.50
0.52
0.54
0.56
0.58
0.60
max_heart_rate_achieved
max_heart_rate_achieved vs max_heart_rate_achieved +/- Standard Deviation
Class 0
Class 1
max_heart_rate_achieved + SD
max_heart_rate_achieved - SD
0.0
0.2
0.4
0.6
0.8
1.0
max_heart_rate_achieved

Figure 32 The sensitivity analysis for maximum heart rate in the heart disease dataset

For both resting blood pressure and cholesterol, the predictions' means decreased steeply after certain value. Figure 33 shows the sensitivity analysis section for the mean prediction based on fixed values for resting blood pressure and cholesterol.

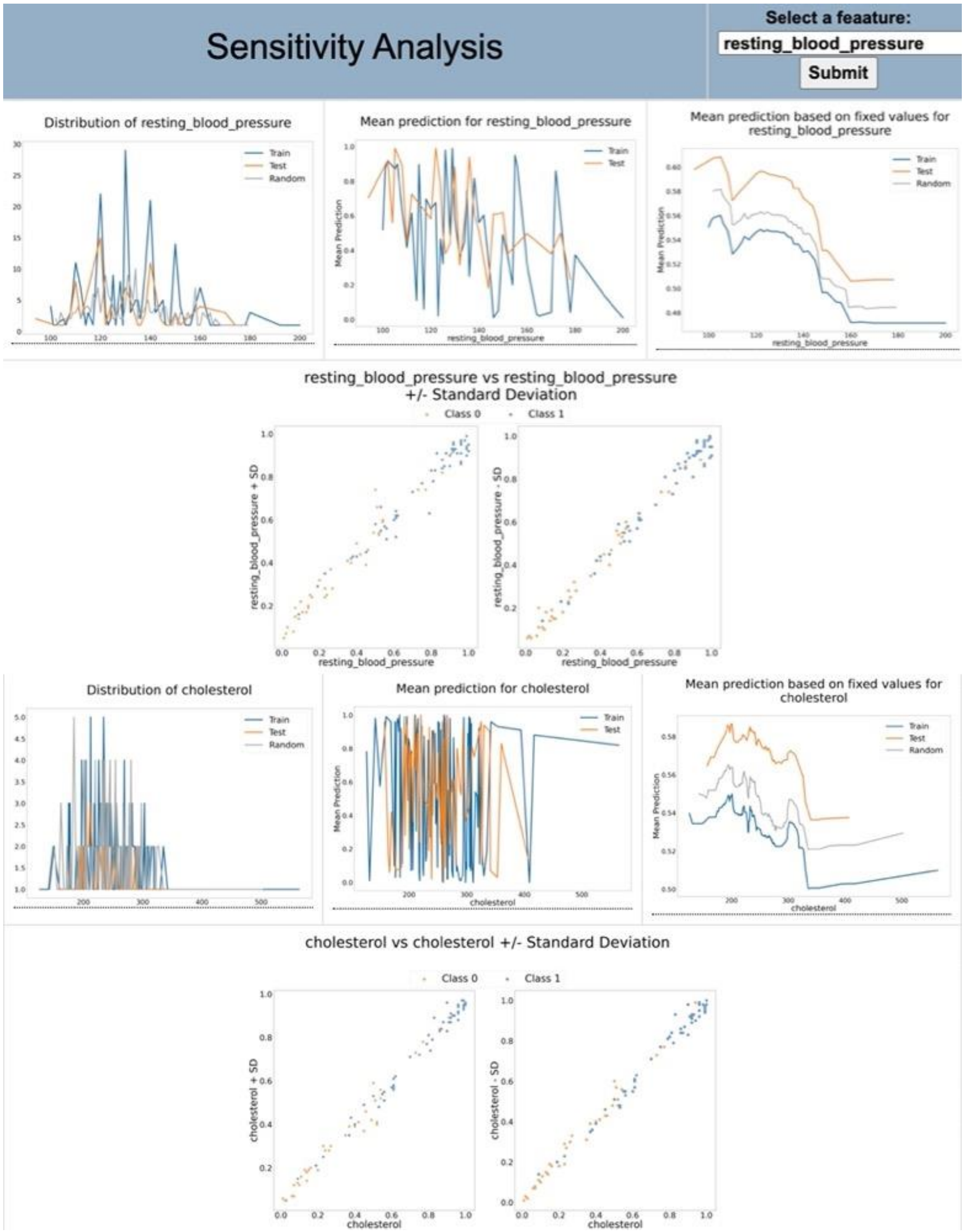

Sensitivity Analysis
Select a feaature:
resting_blood_pressure
Submit
Distribution of resting_blood_pressure
Train
Test
Random
Mean prediction for resting_blood_pressure
Mean Prediction
resting_blood_pressure
Mean prediction based on fixed values for resting_blood_pressure
resting_blood_pressure vs resting_blood_pressure +/- Standard Deviation
Class 0
Class 1
resting_blood_pressure + SD
resting_blood_pressure - SD
Distribution of cholesterol
Mean prediction for cholesterol
Mean prediction based on fixed values for cholesterol
cholesterol
cholesterol vs cholesterol +/- Standard Deviation
cholesterol + SD
cholesterol - SD

Figure 33 Sensitivity analysis for the mean prediction based on fixed values for resting blood pressure and cholesterol

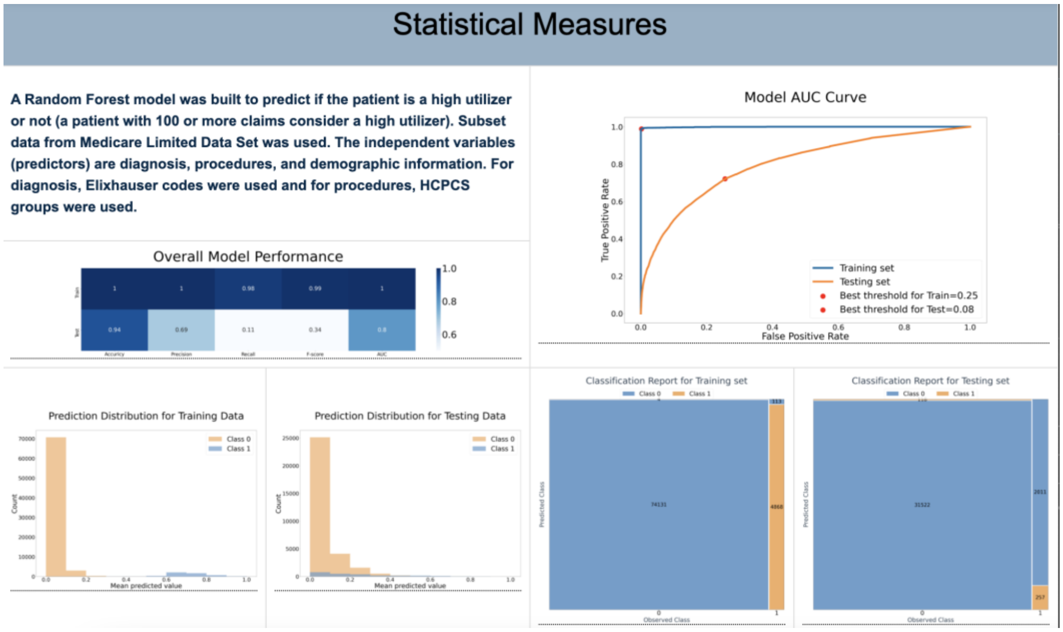
For all plots the predictions' means for testing data is higher than training data. From the classification report, the number of cases for patients with heart disease was higher in the testing data compared to training data that might explain this trend.

Claims Data

The second model used subset data from Medicare Limited Data Set (LDS). LDS contains beneficiary level health information. It is administered by the Centers for Medicare & Medicaid Services (CMS). They removed information that identifies the beneficiaries (Limited Data Set (LDS) Files | CMS, 2019). The model purpose is to predict if the patient will be high utilizer in the following year based on the patient's last year diagnosis and demographics information. A patient with 100 or more claims is considered a high utilizer. The input attributes are list of the Elixhauser comorbidity index for diagnosis, HCPCS groups for procedures, drugs, race, and age. The output attribute is 0 is the patient is not high utilizer and 1 if the patient is high utilizer.

Statistical Measures Section

Starting by the overall performance, the model has a high AUC 0.8. However, the recall is very low using the testing dataset (0.11). Clearly, the model overfit the training data since all values are close to 1. Moving to the ROC curve, the best threshold using testing data is 0.08. The classification plot and prediction distribution show that the data is highly imbalanced. Figure 34 shows the statistical measures section for high utilizers dataset.



Statistical Measures

A Random Forest model was built to predict if the patient is a high utilizer or not (a patient with 100 or more claims consider a high utilizer). Subset data from Medicare Limited Data Set was used. The independent variables (predictors) are diagnosis, procedures, and demographic information. For diagnosis, Elixhauser codes were used and for procedures, HCPCS groups were used.

Overall Model Performance

	Accuricy	Precision	Recall	F-score	AUC
Train	1	1	0.98	0.99	1
Test	0.94	0.69	0.11	0.34	0.8

Figure 34 The statistical measures section for high utilizers dataset

Features Importance Section

In the features' importance section, Elix14 (Deficiency Anemia), Elix6 (Renal Failure), and G-22 (Drugs administered other than oral method, chemotherapy drugs) were the highest three predictors correlated with the output. For the correlation between the predictors: Elix6 (Renal Failure) correlated with elix18 (Hypertension, Complicated) and Elix21(Diabetes, Complicated) correlated with elix9 (Diabetes, Uncomplicated). Even though age had the lowest correlation score with the output, age had the highest score using feature importance methods, followed by Elix14 (Deficiency Anemia) and Elix6 (Renal Failure). Using LASSO, Elix10 (Lymphoma) got a very high score; however, permutation and random forest scores are very low.

In the learning curve, the model AUC did not improve significantly by increasing the number of cases. For number of features, the AUC stopped improving after 50 features. Figure 35 shows the features' importance section for high utilizers dataset.

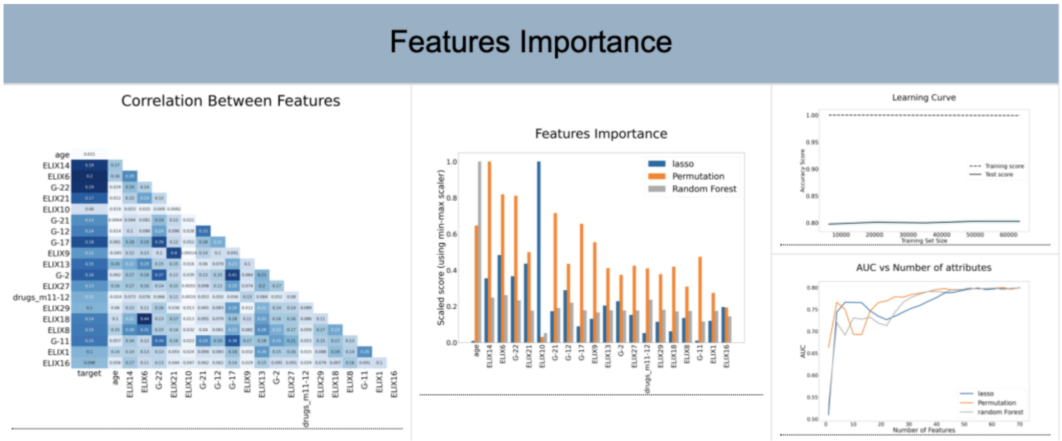
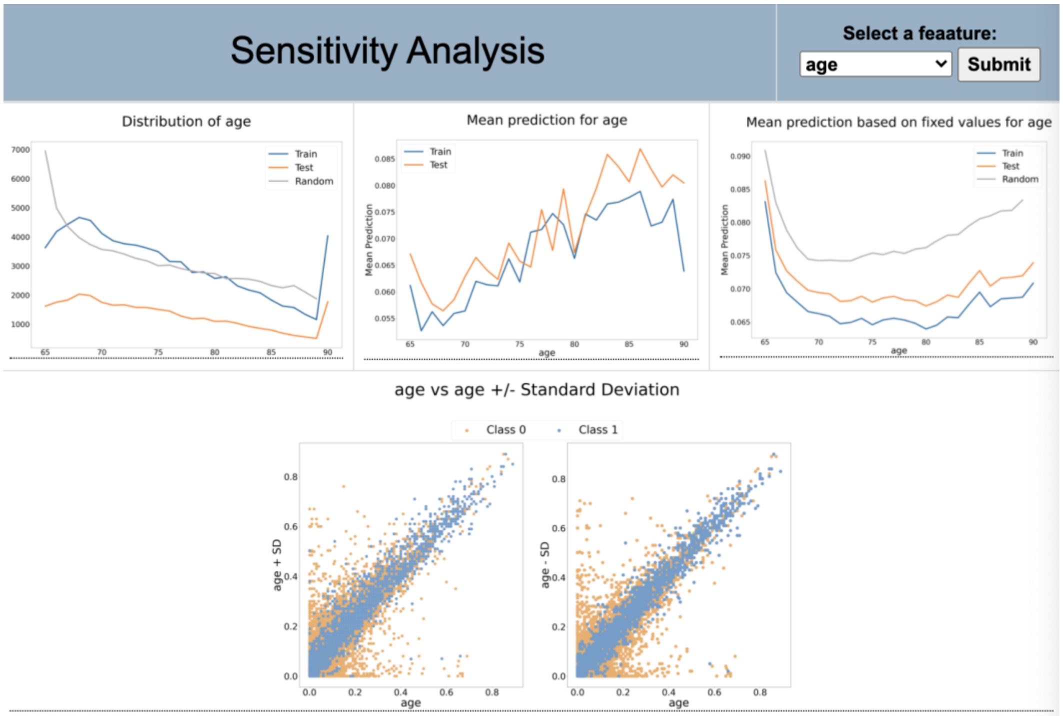


Figure 35 the features' importance section for high utilizers dataset

Sensitivity Analysis Section

Sensitivity analysis for Age. The data distribution shows that most cases are at age 65. The number of cases decreased when the age increased, except for age 90 the number of cases increased. The increase in the age of 90 is because in the dataset any patient aged greater than 90 was set to 90. When the patients’ age increased the mean prediction increased. However, using a fixed age shows totally different trend. Using the fixed values, the predictions’ means decreased till the age of 80 and then the predictions’ means start to increase. The model is slightly sensitive to change of age and is specifically not for high utilizer cases. Figure 36 shows the sensitivity analysis for age in the high utilizer dataset.



Sensitivity Analysis

Select a feaature: age

Figure 36 The sensitivity analysis for age in the high utilizer dataset

Figure 37 shows that the rest of the top 20 features are categorical attributes and have the same trends. All distribution is left skewed using training, testing, and random datasets.

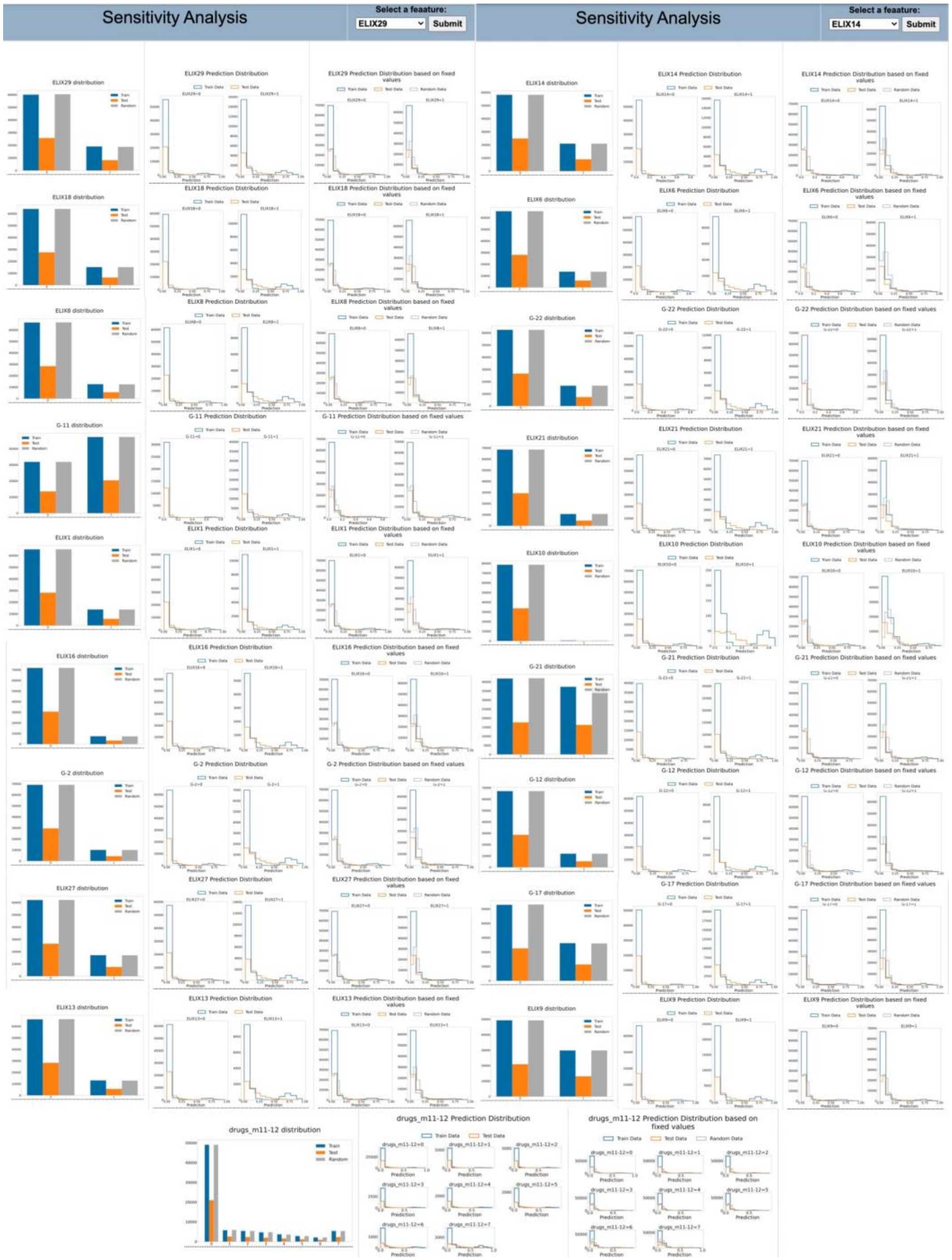
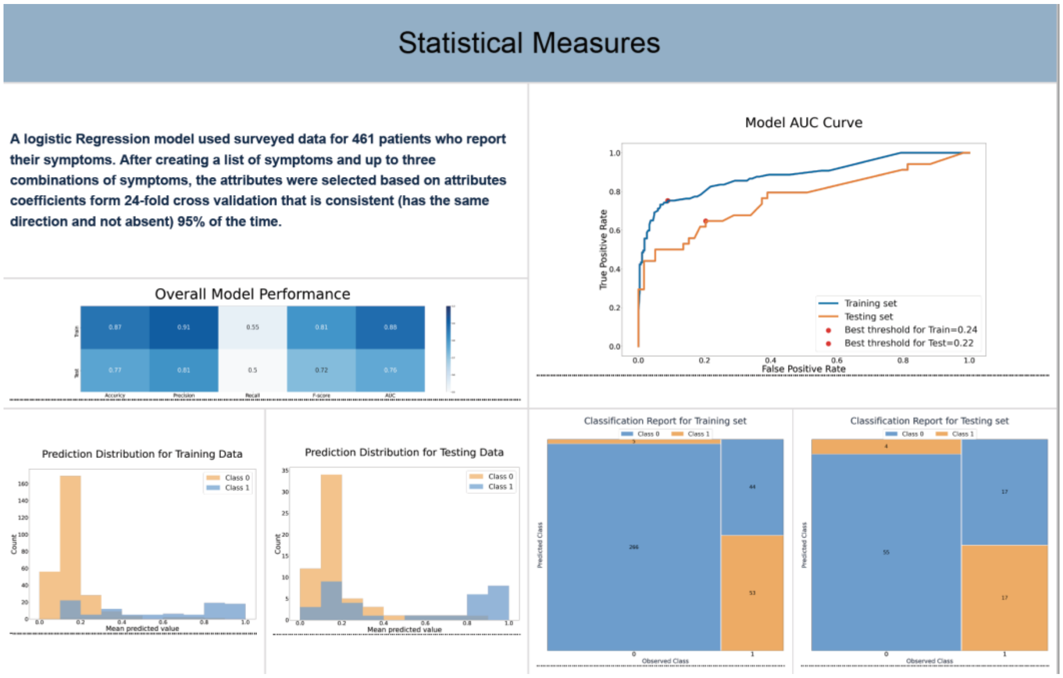

Sensitivity Analysis
Select a feaature:
ELIX29
Submit
Sensitivity Analysis
Select a feaature:
ELIX14
Submit
ELIX29 distribution
ELIX29 Prediction Distribution
ELIX29 Prediction Distribution based on fixed values
ELIX14 distribution
ELIX14 Prediction Distribution
ELIX14 Prediction Distribution based on fixed values
ELIX18 distribution
ELIX18 Prediction Distribution
ELIX18 Prediction Distribution based on fixed values
ELIX6 distribution
ELIX6 Prediction Distribution
ELIX6 Prediction Distribution based on fixed values
ELIX8 distribution
ELIX8 Prediction Distribution
ELIX8 Prediction Distribution based on fixed values
G-22 distribution
G-22 Prediction Distribution
G-22 Prediction Distribution based on fixed values
G-11 distribution
G-11 Prediction Distribution
G-11 Prediction Distribution based on fixed values
ELIX21 distribution
ELIX21 Prediction Distribution
ELIX21 Prediction Distribution based on fixed values
ELIX1 distribution
ELIX1 Prediction Distribution
ELIX1 Prediction Distribution based on fixed values
ELIX10 distribution
ELIX10 Prediction Distribution
ELIX10 Prediction Distribution based on fixed values
ELIX16 distribution
ELIX16 Prediction Distribution
ELIX16 Prediction Distribution based on fixed values
G-21 distribution
G-21 Prediction Distribution
G-21 Prediction Distribution based on fixed values
G-2 distribution
G-2 Prediction Distribution
G-2 Prediction Distribution based on fixed values
G-12 distribution
G-12 Prediction Distribution
G-12 Prediction Distribution based on fixed values
ELIX27 distribution
ELIX27 Prediction Distribution
ELIX27 Prediction Distribution based on fixed values
G-17 distribution
G-17 Prediction Distribution
G-17 Prediction Distribution based on fixed values
ELIX13 distribution
ELIX13 Prediction Distribution
ELIX13 Prediction Distribution based on fixed values
ELIX9 distribution
ELIX9 Prediction Distribution
ELIX9 Prediction Distribution based on fixed values
drugs_m11-12 distribution
drugs_m11-12 Prediction Distribution
drugs_m11-12 Prediction Distribution based on fixed values
Train
Test
Random
Train Data
Test Data
Random Data
Prediction

Figure 37 The sensitivity analysis for 19 attributes in the high utilizer dataset

COVID 19 Data

A logistic regression model used surveyed data for 461 patients who reported their symptoms. After creating a list of symptoms and up to three combinations of symptoms, the attributes were selected based on attributes coefficients from 24-fold cross validation that is consistent (has the same direction and not absent) 95% of the time (Tibshirani et al., 2012). The selected attributes are: Headaches and loss of taste; Chest pain, chills, and fatigue; Headaches and pinkeye; Chest pain and chills; Cough and runny nose; Abdominal pain and Muscle aches; Fatigue, headaches, and muscle aches; Cough and loss of appetite; Chills, fatigue, and wheezing; Fever and headaches; Excess sweat, fever, and loss of smell; Cough, excess sweat, and loss of smell; Chest pain, chills, and muscle aches; Headaches; Diarrhea, muscle aches, and runny nose; Runny nose; Diarrhea; Cough, loss of taste, and runny nose; History of respiratory symptoms.

Statistical Measures. Figure 38 reports that the model has 0.76 AUC and 0.5 recall using 0.5 default threshold. Based on the ROC Curve, the suggested threshold is 0.22. The number of positive cases represent 37% of the data. There are two peaks of the distribution of the positive cases, one in the 100% and the other in the 2%. In comparison, the negative cases have one peak of the distribution in the 2%.



Statistical Measures

A logistic Regression model used surveyed data for 461 patients who report their symptoms. After creating a list of symptoms and up to three combinations of symptoms, the attributes were selected based on attributes coefficients form 24-fold cross validation that is consistent (has the same direction and not absent) 95% of the time.

Overall Model Performance

	Accuricy	Precision	Recall	F-score	AUC
Train	0.87	0.91	0.55	0.81	0.88
Test	0.77	0.81	0.5	0.72	0.76

Figure 38 Statistical measure section for COVID 19 data

Features Importance. Figure 39 shows that chest pain and chills combinations are the top symptoms that correlated with positive test results. The top three attributes are chest pain and chills; chest pain, chills, and muscle aches; and chest pain, chills, and fatigue. However, using features importance methods, two of the chest pain and chills has the lowest score. The top three predictors are: cough & loss of taste & runny nose, excess sweat & fever & loss of smell, and headaches & pinkeye. In the learning curve, the training line did not stop increasing which suggest that increasing the number of cases may increase the model accuracy.

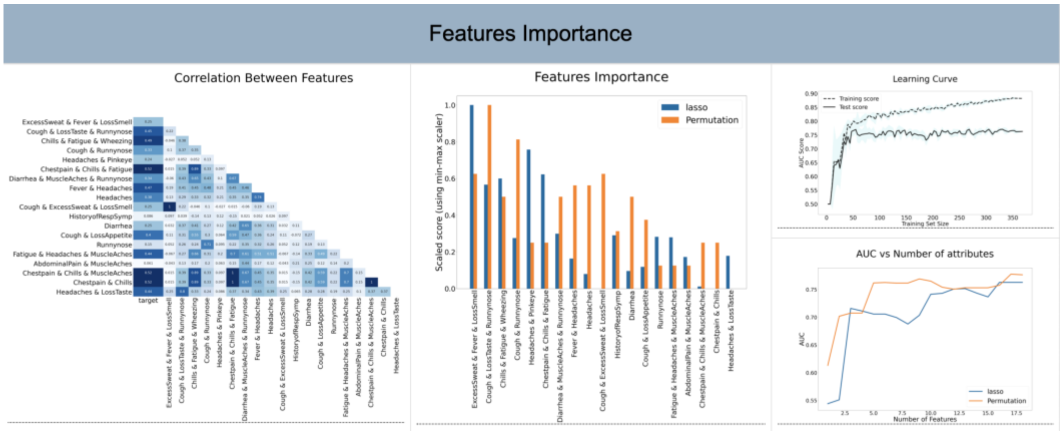

Features Importance
Correlation Between Features
ExcessSweat & Fever & LossSmell
Cough & LossTaste & Runnynose
Chills & Fatigue & Wheezing
Cough & Runnynose
Headaches & Pinkeye
Chestpain & Chills & Fatigue
Diarrhea & MuscleAches & Runnynose
Fever & Headaches
Headaches
Cough & ExcessSweat & LossSmell
HistoryofRespSymp
Diarrhea
Cough & LossAppetite
Runnynose
Fatigue & Headaches & MuscleAches
AbdominalPain & MuscleAches
Chestpain & Chills & MuscleAches
Chestpain & Chills
Headaches & LossTaste
target
Features Importance
Scaled score (using min-max scaler)
lasso
Permutation
Learning Curve
Training score
Test score
AUC Score
Training Set Size
AUC vs Number of attributes
AUC
lasso
Permutation
Number of Features

Figure 39 Feature importance section for COVID 19 data

Sensitivity Analysis. For all attributes except runny nose, the number of cases when the attribute is absent is higher than the number of cases when the attribute is present.

The following attributes have positive correlation with prediction positive COVID-19 test using the prediction distribution as shown in figure 40:

- Cough & loss of taste & runny nose: the prediction peak for random data and value 1 for all cases is 4%.
- Headaches & pinkeye → the prediction peak for random data and value 1 for all cases is 4%. One important note is that the number of positive cases is very small.
- Chills & fatigue & wheezing → the prediction peak for random data and value 1 for all cases is 4%

- Chest pain & chills & fatigue → the prediction peak for random data and value 1 for all cases is 2%
- Fatigue & headaches & muscle ache → the prediction peak for random data and value 1 for all cases is 2%
- Headaches & loss of taste → the prediction peak for random data and value 1 for all cases is 2%.
- Cough & excess sweat & loss of smell → the prediction peak for random data and value 1 for all cases is 2%. One important note is that the number of positive cases is very small.
- Cough & loss of appetite → the prediction peak for random data and value 1 for all cases is 2%.
- Chest pain & chills & muscle aches → the prediction peak for random data and value 1 for all cases is 2%.

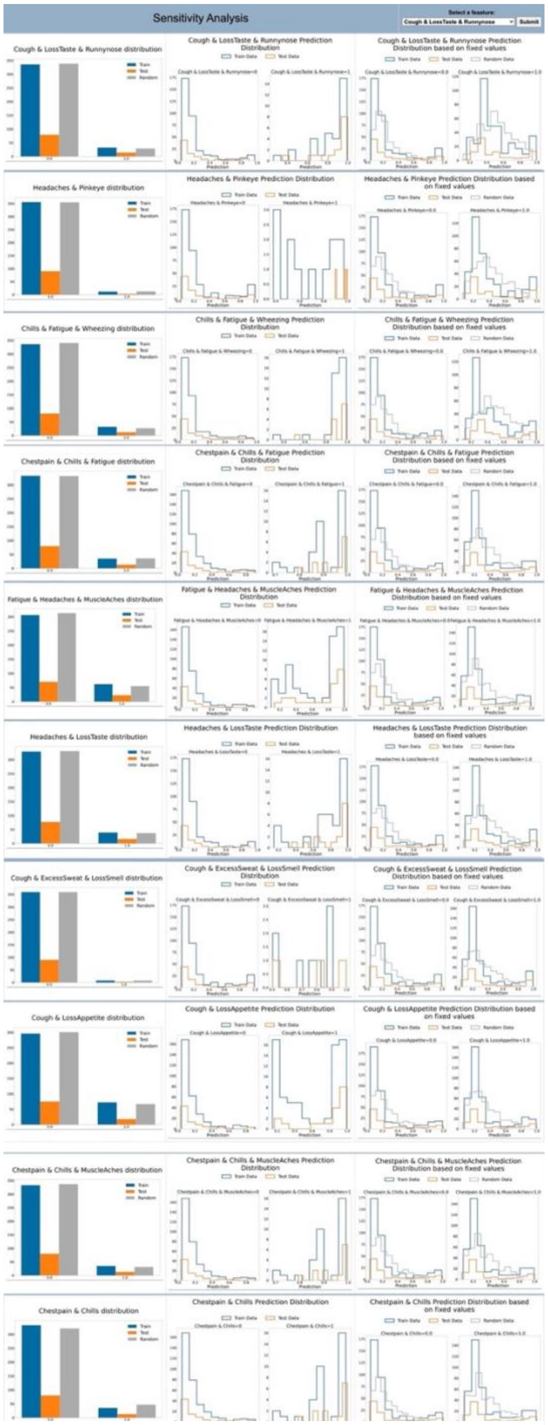
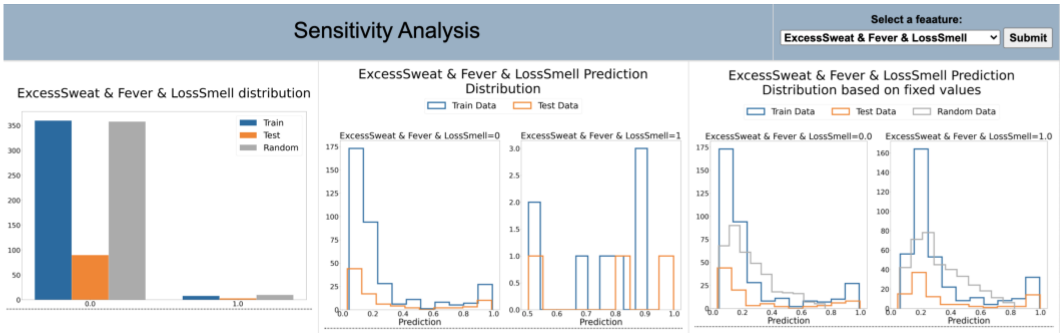

Sensitivity Analysis
Select a feature:
Cough & LossTaste & Runnynose
Submit
Cough & LossTaste & Runnynose distribution
Train
Test
Random
Cough & LossTaste & Runnynose Prediction Distribution
Train Data
Test Data
Cough & LossTaste & Runnynose=0
Cough & LossTaste & Runnynose=1
Prediction
Cough & LossTaste & Runnynose Prediction Distribution based on fixed values
Random Data
Cough & LossTaste & Runnynose=0.0
Cough & LossTaste & Runnynose=1.0
Headaches & Pinkeye distribution
Headaches & Pinkeye Prediction Distribution
Headaches & Pinkeye=0
Headaches & Pinkeye=1
Headaches & Pinkeye Prediction Distribution based on fixed values
Headaches & Pinkeye=0.0
Headaches & Pinkeye=1.0
Chills & Fatigue & Wheezing distribution
Chills & Fatigue & Wheezing Prediction Distribution
Chills & Fatigue & Wheezing=0
Chills & Fatigue & Wheezing=1
Chills & Fatigue & Wheezing Prediction Distribution based on fixed values
Chills & Fatigue & Wheezing=0.0
Chills & Fatigue & Wheezing=1.0
Chestpain & Chills & Fatigue distribution
Chestpain & Chills & Fatigue Prediction Distribution
Chestpain & Chills & Fatigue=0
Chestpain & Chills & Fatigue=1
Chestpain & Chills & Fatigue Prediction Distribution based on fixed values
Chestpain & Chills & Fatigue=0.0
Chestpain & Chills & Fatigue=1.0
Fatigue & Headaches & MuscleAches distribution
Fatigue & Headaches & MuscleAches Prediction Distribution
Fatigue & Headaches & MuscleAches Prediction Distribution based on fixed values
Headaches & LossTaste distribution
Headaches & LossTaste Prediction Distribution
Headaches & LossTaste Prediction Distribution based on fixed values
Cough & ExcessSweat & LossSmell distribution
Cough & ExcessSweat & LossSmell Prediction Distribution
Cough & ExcessSweat & LossSmell Prediction Distribution based on fixed values
Cough & LossAppetite distribution
Cough & LossAppetite Prediction Distribution
Cough & LossAppetite Prediction Distribution based on fixed values
Chestpain & Chills & MuscleAches distribution
Chestpain & Chills & MuscleAches Prediction Distribution
Chestpain & Chills & MuscleAches Prediction Distribution based on fixed values
Chestpain & Chills distribution
Chestpain & Chills Prediction Distribution
Chestpain & Chills Prediction Distribution based on fixed values

Figure 40 Sensitivity analysis for attributes with positive prediction correlation.

- Chest pain & chills → the prediction peak for random data and value 1 for all cases is between 2% and 3%.

Figure 41 shows that excess sweat, fever, and loss of smell has weaker positive correlation with predicting a positive COVID-19 test. The prediction peak for random data and value 1 for all cases is 2%. One important note is that the number of positive cases is very small.


Sensitivity Analysis
Submit
ExcessSweat & Fever & LossSmell distribution
Train
Test
Random
350
300
250
200
150
100
50
0
0.0
1.0
ExcessSweat & Fever & LossSmell Prediction Distribution
Train Data
Test Data
ExcessSweat & Fever & LossSmell=0
ExcessSweat & Fever & LossSmell=1
175
150
125
100
75
50
25
0
0.0
0.2
0.4
0.6
0.8
1.0
Prediction
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0.5
0.6
0.7
0.8
0.9
1.0
Prediction
ExcessSweat & Fever & LossSmell Prediction Distribution based on fixed values
Train Data
Test Data
Random Data
ExcessSweat & Fever & LossSmell=0.0
ExcessSweat & Fever & LossSmell=1.0
175
150
125
100
75
50
25
0
0.0
0.2
0.4
0.6
0.8
1.0
Prediction
160
140
120
100
80
60
40
20
0
0.0
0.2
0.4
0.6
0.8
1.0
Prediction

Select a feaature: ExcessSweat & Fever & LossSmell

Figure 41 Sensitivity analysis for Excess sweat & fever & loss

The following attributes have negative correlation with predicting a negative COVID-19 test result using the prediction distribution as shown in figure 42:

- History of respiratory symptoms: the prediction peak for random data and value 1 for all cases is 2%.
- Runny nose: the prediction peak for random data and value 1 for all cases is 1%
- Headaches: the prediction peak for random data and value 1 for all cases is 1%

- Abdominal pain & muscle ache: the prediction peak for random data and value 1 for all cases is 1%. One important note is that the number of positive cases is very small.

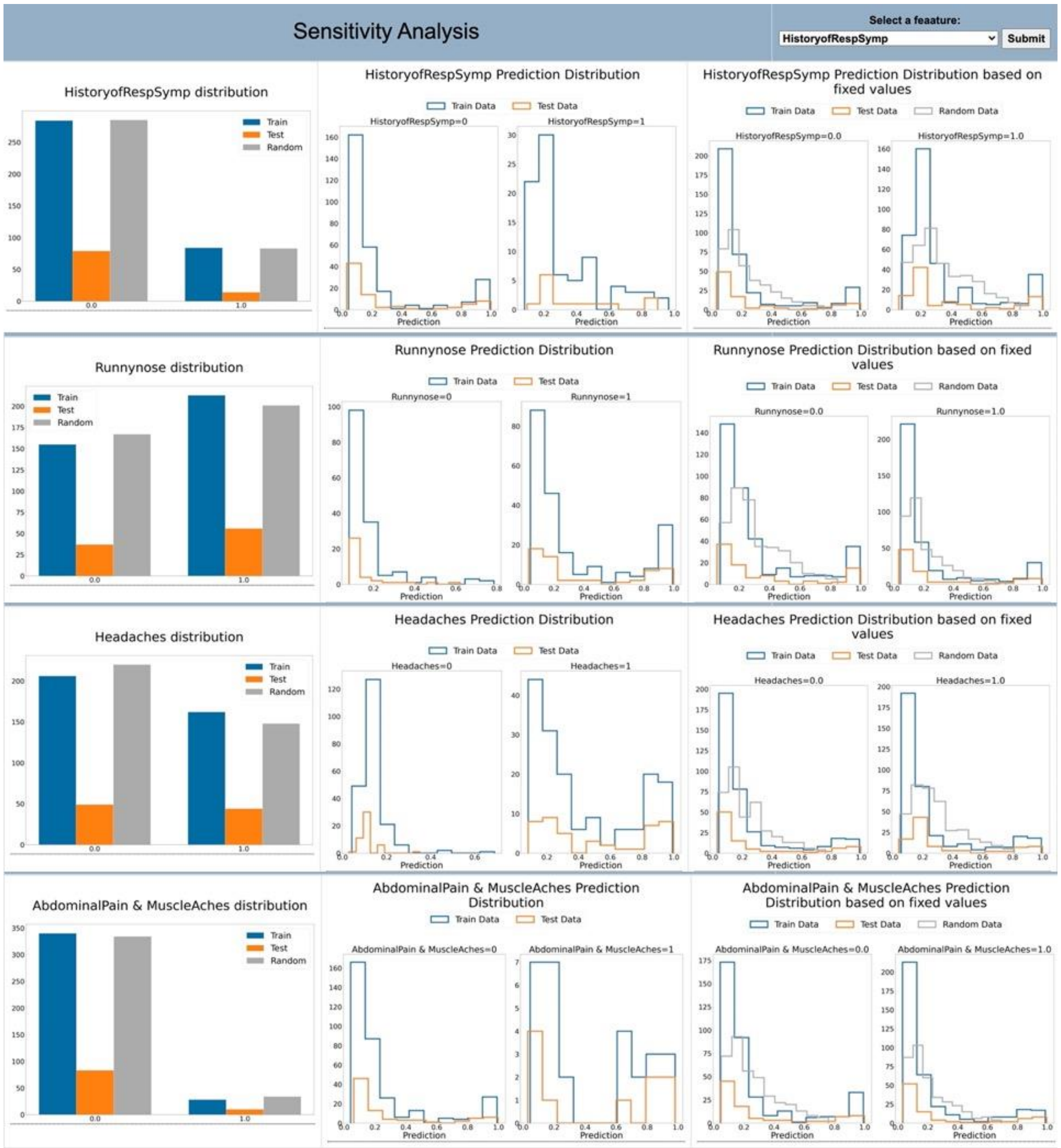


Figure 42 Sensitivity analysis for attributes with negative prediction correlation.

Figure 43 shows that cough & runny nose and diarrhea have a weaker negative correlation with prediction positive COVID-19 test. The prediction peak for random data and value 1 for all cases are 3% and 1%, respectively. One important note is that the number of positive cases is very small.

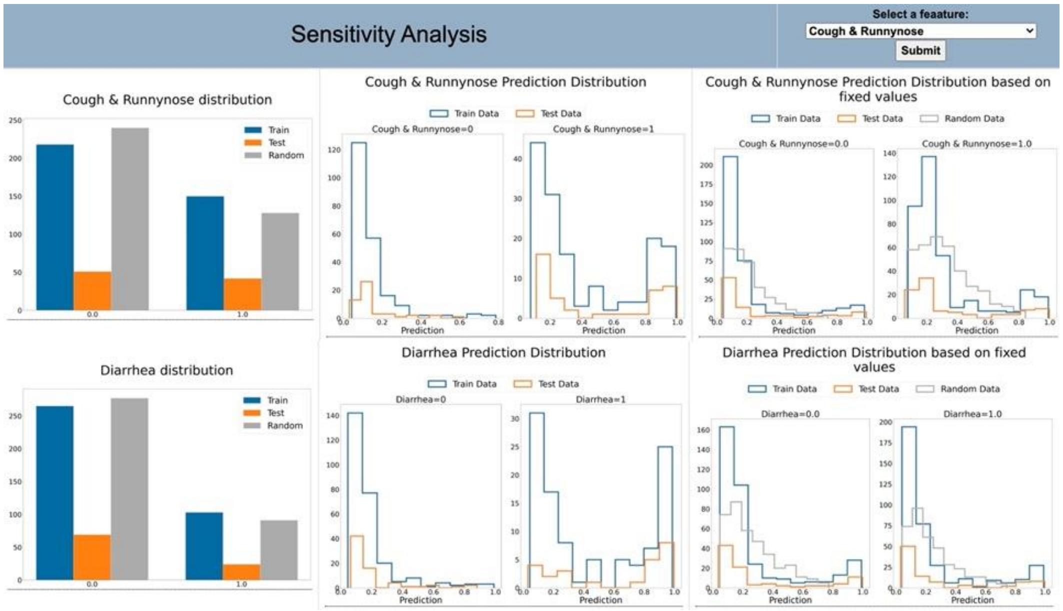
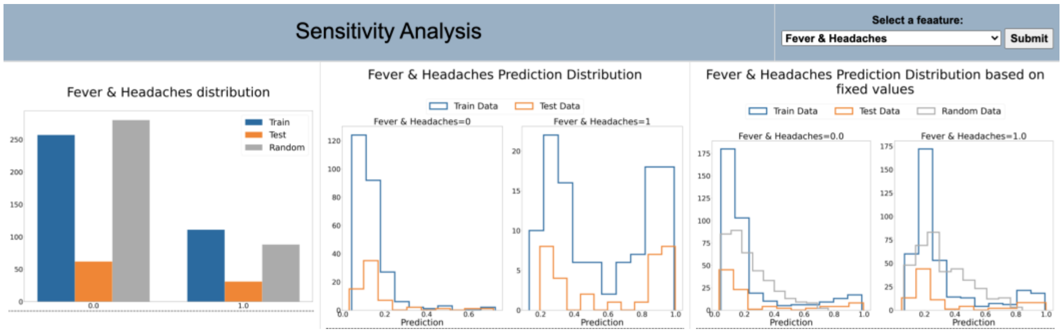


Figure 43 Sensitivity analysis for Cough & runny nose and diarrhea

Figure 44 shows that the prediction distribution of having fever & headaches has two peaks one near the 100% and the other near the 2%. Therefore, the relationship is not clear. Using the random data, the peak is around 2%.

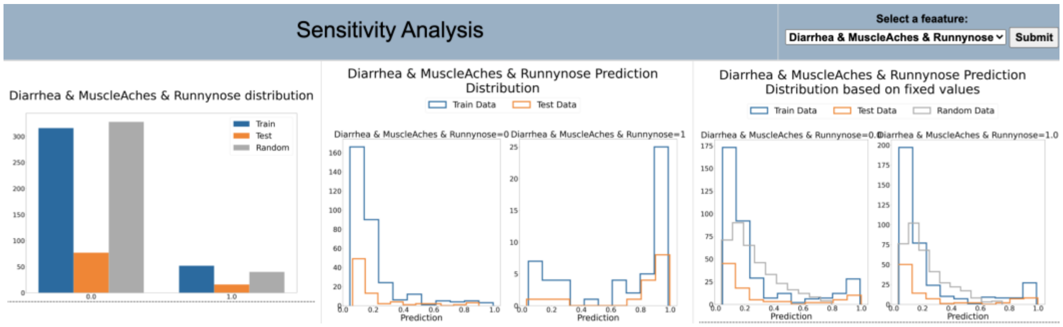


Sensitivity Analysis

Select a feaature: Fever & Headaches

Figure 44 Sensitivity analysis for fever & headaches

Figure 45 shows that the distribution of having diarrhea & muscle aches & runny nose is positive. Nonetheless, setting the value for diarrhea & muscle aches & runny nose to 1 for all data results one prediction peak at 0% and the distribution is similar for both values 0 and 1.



Sensitivity Analysis

Select a feaature: Diarrhea & MuscleAches & Runnynose

Submit
Diarrhea & MuscleAches & Runnynose distribution
Train
Test
Random
0
50
100
150
200
250
300
0.0
1.0
Diarrhea & MuscleAches & Runnynose Prediction Distribution
Train Data
Test Data
Diarrhea & MuscleAches & Runnynose=0
Diarrhea & MuscleAches & Runnynose=1
Prediction
Diarrhea & MuscleAches & Runnynose Prediction Distribution based on fixed values
Train Data
Test Data
Random Data
Diarrhea & MuscleAches & Runnynose=0.0
Diarrhea & MuscleAches & Runnynose=1.0
Prediction

Figure 45 Sensitivity analysis for diarrhea & muscle aches & runny nose

RESULTS

A small survey was distributed to gain feedback on and capture the usefulness of the dashboard. The survey provides the reviewers the three study cases described earlier and asked them to evaluate the dashboard's three sections. The number of responses was 15 faculty members or graduate students in data analytics, informatics, or health sciences. First, the survey asked about the position and area of work. Then the survey asked the user to evaluate the usefulness of the 3 sections of the dashboard. To protect reviewers' identity, the results split them into two categories: faculty members or students. The survey was approved by George Mason University IRB number 1766037-1.

Survey Results

15 people evaluated the dashboards: 5 faculty members and 10 students. 13 out of 15 people agreed that the dashboard visuals give a better understanding of the model behavior than other methods they have experienced. For evaluating each section's usefulness, Table 1 shows the summary of the responses scores. In general, most reviewers provided positive comments. They thought that the dashboard gives a comprehensive understanding of the dataset and the model, while some provided recommendation and other mentioned that some of the plots were not useful. Tables 2, 3, 4, and 5 show the comments divided into three sections: positive, neutral, and negative.

Statistical measures section. Most of the comments agreed that this section is important to understand how the model performs. This section was the most interesting section for one of the reviewers in terms of understanding. However, for "Prediction

Distribution and Classification Reports," one of the comments suggests that they are unnecessary.

Feature's importance section. Several reviews mentioned that this section is important to give an idea about the data. The correlation plot got the most attention; however, the size of the plots was too small to read.

Sensitivity analysis section. Most of the comments agreed that selecting a variable is very helpful to understand the performance. However, one of the comments found it hard to understand the categorical attributes plots.

Finally, most of the comments were positive. Comments related to the size of plots, typos, and rewording were reflected on the dashboard. The other suggestions would be considered as future work due to time limitations.

Table 1 Survey evaluation score per dashboard section

	statistical Measures	Features Importance	Sensitivity Analysis
Extremely useful	4	7	6
Very useful	7	5	4
Moderately useful	3	3	4
Slightly useful	0	0	1
Not at all useful	1	0	0
TOTAL	15	15	15

Table 2 Survey general comments

positive	neutral	negative
Model 3 was particularly clear and useful to visual dense data and results in an easy to understand format	I don't know much about ML, so any understanding of model behavior is better than what I have experienced (sorry that I	The display is too busy.

	am not very helpful in this way)	
The dashboard highlights the most important aspects about model evaluation.	If you have a little bit background in subject matter, it is much easier to quickly gain a lot of information.	Useful information is provided but I do not understand the purpose of creating it as there are many products that does the same. So, I'm not sure of the value being provided.
Dashboard 2 contains the most useful and understandable information.	I think there were some typos in the dashboard (did you mean "variable" where "vitiable" is written?). Also assumes a level of knowledge about these models that I, personally, don't have but perhaps the target audience does have.	I have seen these types of visuals before. These do not give me a better understanding than I have experienced before.
I really like the way the dashboard is created. It gives a comprehensive understanding of the dataset and the model. I felt using different colors for threshold values in train and test would help the viewer understand in an easier way.	Very similar dashboard for regression class ML models. Including dashboards to evaluate challenger vs. champion models performance will be great.	Suggest scaling back the visuals. Too much.
The first graph clearly layout most important metrics to evaluate classification models which is good.	I would have liked to know which dashboard I was viewing. The top of the page could have said Model Evaluation-Dashboard1. I got lost trying to compare the different models flipping back and forth.	The layout of dashboard is just clutter of different visualizations. It works for the people in field; however, for normal Joe, it would be more effective if these visualizations are accompanied by information about the purpose of visualization.

The visuals are clean and easy to understand. A dynamic dashboard that allows you to look at one feature at a time is always helpful for comparison.		
Yes, I learn and understand data/information better with a visual representation accompanying the textual information.		
It's easy to read		
Overall, the dashboard clearly represented model evaluation metrics that are required to quantify model performance. The choices of evaluation metrics on a given machine learning model give the most important information that needed. Coloring and dividing the dashboard into sections make it very organized. Last but not least, using hoover makes the graph clear and understandable.		
I thought each of the sections provided very useful information/data. I think perhaps there might be too many graphs in each section. I like the pull-down option enabling the user to select the graph/chart they would like to view. In addition, from my experience with software projects, often users do not like to scroll too much when looking at a dashboard - just feedback I have received over the years. I hope that was helpful - best of luck		

Table 3 Survey comments related to statistical measures section

positive	neutral	negative
presentation of results (conclusions) intuitively useful	It provided context to the project, providing a stage for the data.	Useful but a challenge to weight through the visual noise.
Representing AUC, Precision, and recall is clear beside with the best threshold of the model.	It provides quantified results.	From my perspective, the Prediction Distribution and Classification Reports don't contain very much useful information and do not seem necessary.
I thought this section was the most interesting in terms of understanding how well the model performed. I found dashboard 1 to be the easiest to follow. There were some results in dashboards 2 and 3 that I didn't completely follow.	Again, about the color coding for threshold in AUC can be thought of.	
Everything is perfect is what I feel.		
I am visual person. I do not like confusion in my graphic representations. I appreciate the term Confusion Matrix; however, this visual representation is easy to understand at quick look.		

Table 4 Survey comments related to features importance section

positive	neutral	negative
Dashboard2 was easier to view. Using ELIX as the identifiers kept the screen from being crowded.	Features Importance represent at least two FI depending on two different techniques. Also, having the learning curve and AUC vs. No.	Everything was very small, and I had a hard time reading it.

	of feature boost the usefulness.	
It allows us to see which features are most relevant; and then select analysis based on relevant features. Gets rid of clutter; also helps identify features that are not contributing to issue.	Renaming the x-axis of learning curve chart can be considered. It was a bit confusing. Also having Number of Features (x-axis of 4th chart) as Integers would make it look better instead of decimals.	I do believe the graphs/charts are interesting such as the "Features Importance". The AUC graph and learning graph were similar to the AUC chart in the statistical results section. I was not sure why I am seeing both - just a thought
I like to see the correlation between features and the feature importance.		
The correlation chart is interesting, but too small to read		
It gives me an understanding of the dataset. That is really great and important.		

Table 5 Survey comments related to sensitivity analysis section

Positive	neutral	negative
Being able to change the variables that may be more sensitive/related to a particular (i.e. untoward) outcome is VERY helpful! Also, very quick to compute for a variety of variables	Gives us quantified result.	I don't understand what I am looking at for the categorical data.

Selecting a specific feature makes the evaluation easier and it is clear when highlighted the distribution with fixed value.		
The sensitivity analysis line charts are especially clear ways to display the performance information.		
I really like the interactive feature. It helps us to get a better understanding of the variables.		
I liked I was able to select a feature.		
I found it useful, because I like the ability to explore the data quickly in a format that I was able to understand.		

CONCLUSION AND DISCUSSION

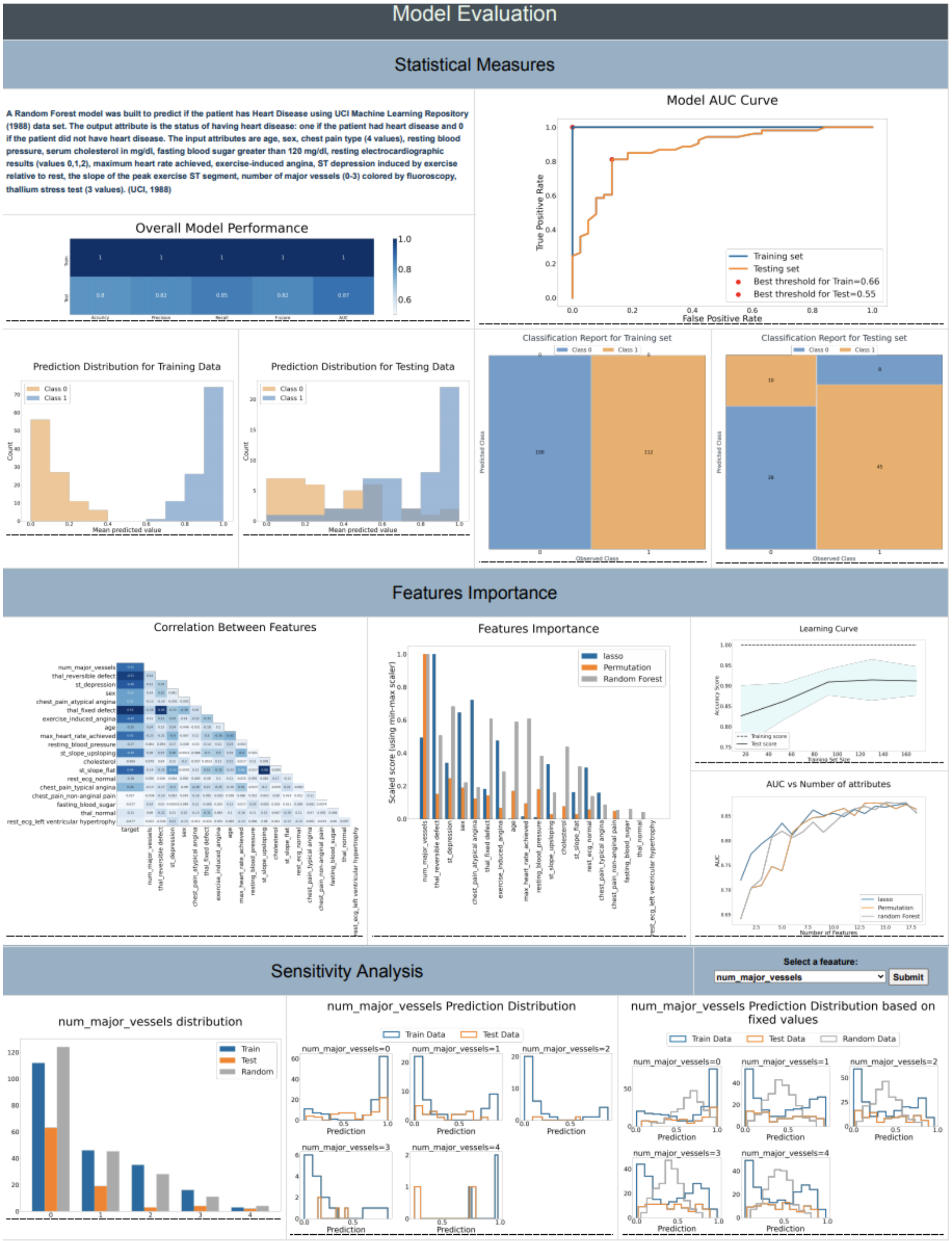
The present thesis was designed to demonstrate an approach to visualizing classification model performance in a dashboard with three sections: statistical measures which provide an overview of the model performance, feature importance which gives an overview of the data, and sensitivity analysis which identifies the relationship between the attribute and the prediction. The dashboard adds to a growing body of literature on understanding and evaluating classification learning. Most of the survey feedback found the dashboard useful and easy to understand.

Limitation and Future Work

The survey results cannot be generalized due to sample size limitation. However, the purpose of the survey was to understand how people interact with the dashboard, and the most interesting part was the reviewers' comments. Second, some design-related changes like the colors and sizes of the plots are recommended. For example, when the names of the columns are long, the size of the figures in the feature importance section becomes small, which required zooming in to read. Visualizing the regression model results and comparing models are the next part of this work.

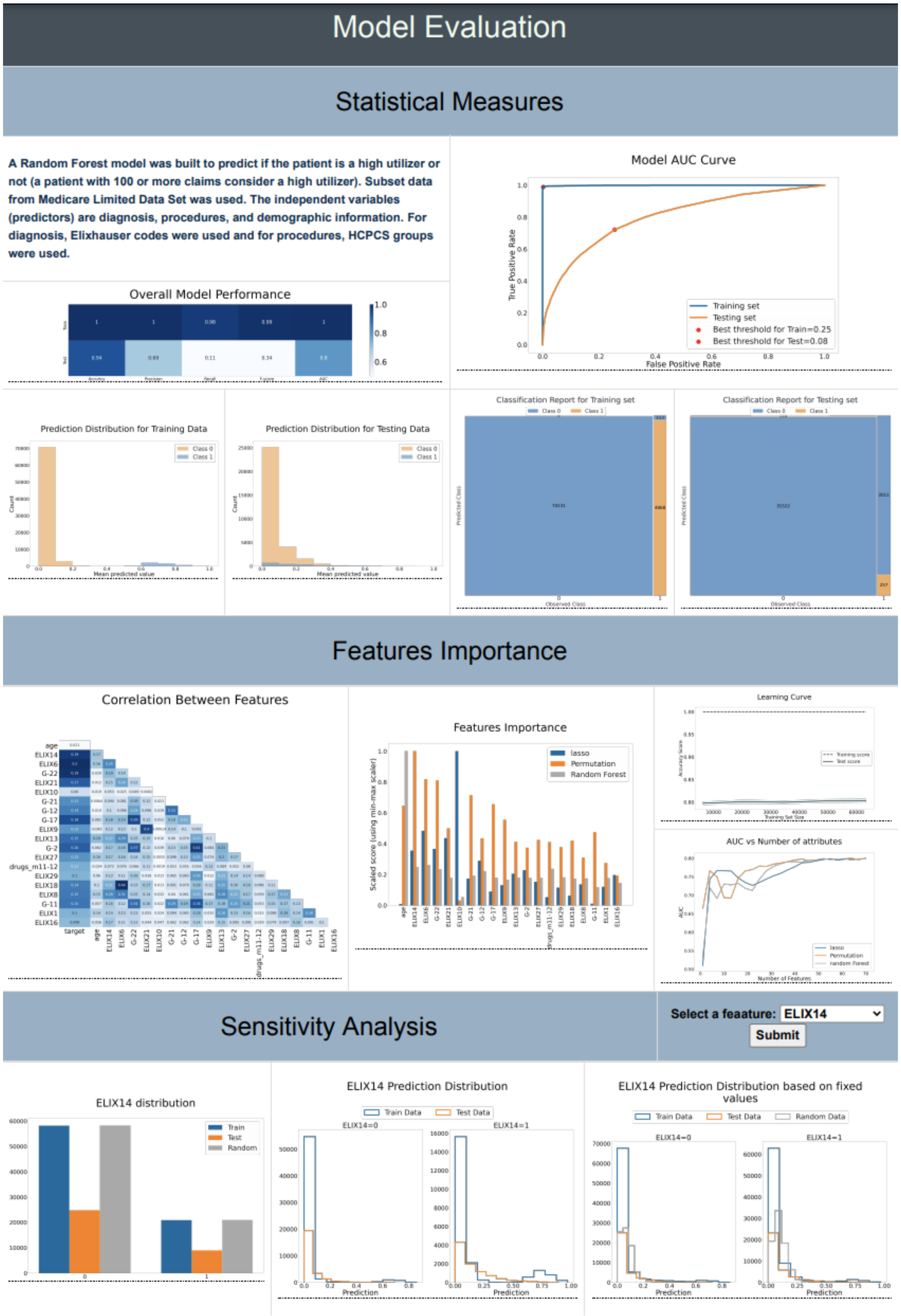
APPENDIX

Dashboard 1: heart disease dataset



A Random Forest model was built to predict if the patient has Heart Disease using UCI Machine Learning Repository (1988) data set. The output attribute is the status of having heart disease: one if the patient had heart disease and 0 if the patient did not have heart disease. The input attributes are age, sex, chest pain type (4 values), resting blood pressure, serum cholesterol in mg/dl, fasting blood sugar greater than 120 mg/dl, resting electrocardiographic results (values 0,1,2), maximum heart rate achieved, exercise-induced angina, ST depression induced by exercise relative to rest, the slope of the peak exercise ST segment, number of major vessels (0-3) colored by fluoroscopy, thallium stress test (3 values). (UCI, 1988)

Dashboard2: claims dataset



Model Evaluation

Statistical Measures

A Random Forest model was built to predict if the patient is a high utilizer or not (a patient with 100 or more claims consider a high utilizer). Subset data from Medicare Limited Data Set was used. The independent variables (predictors) are diagnosis, procedures, and demographic information. For diagnosis, Elixhauser codes were used and for procedures, HCPCS groups were used.

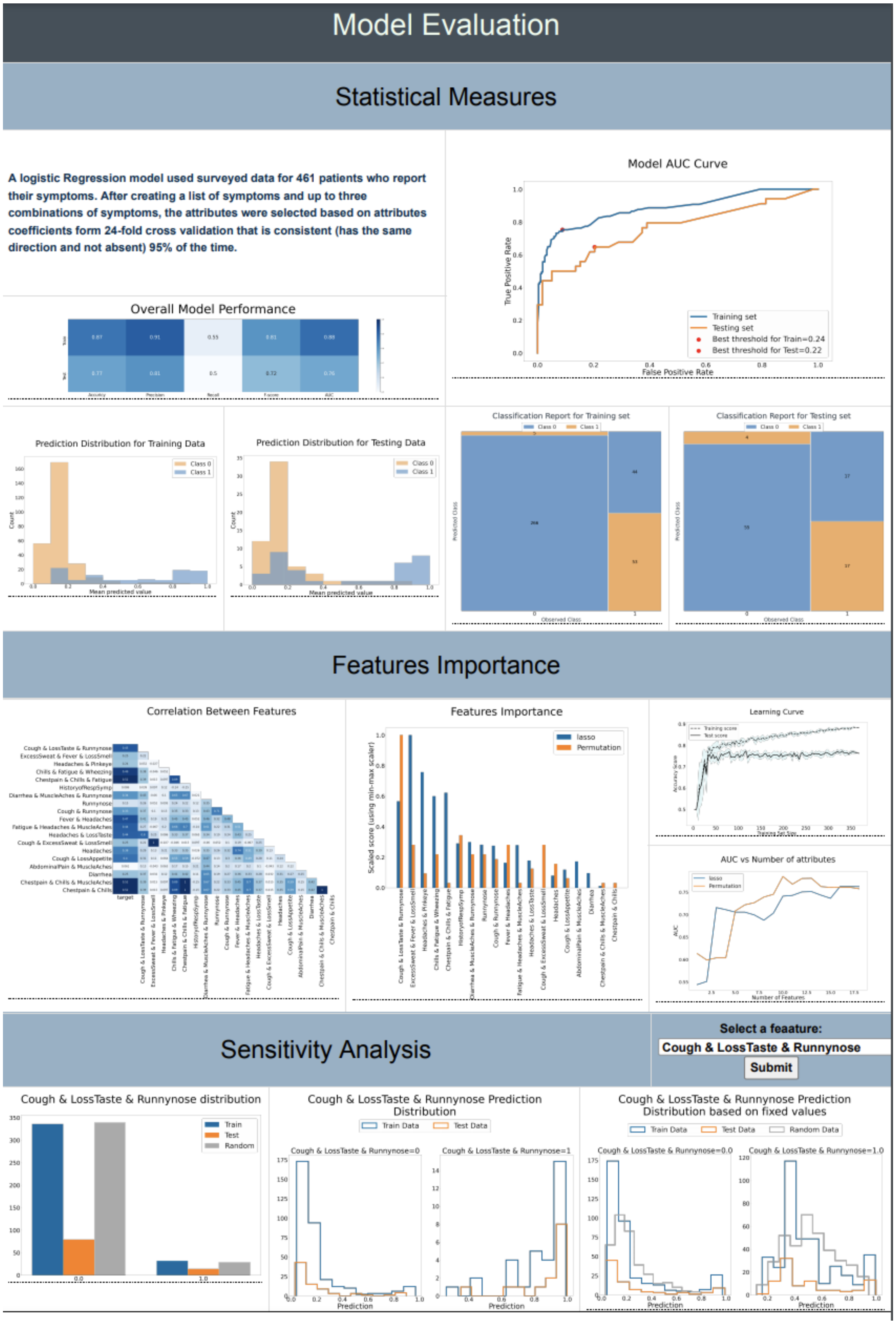
Features Importance

Sensitivity Analysis

Select a feaature: ELIX14

Submit

Dashboard3: covid-19 dataset



Model Evaluation

Statistical Measures

A logistic Regression model used surveyed data for 461 patients who report their symptoms. After creating a list of symptoms and up to three combinations of symptoms, the attributes were selected based on attributes coefficients form 24-fold cross validation that is consistent (has the same direction and not absent) 95% of the time.

Features Importance

Sensitivity Analysis

Select a feaature:

Cough & LossTaste & Runnynose

Submit

Overall Model Performance
Train
Test
0.87
0.91
0.55
0.81
0.88
0.77
0.81
0.5
0.72
0.76
Accuracy
Precision
Recall
F-score
AUC
Model AUC Curve
True Positive Rate
False Positive Rate
Training set
Testing set
Best threshold for Train=0.24
Best threshold for Test=0.22
Prediction Distribution for Training Data
Prediction Distribution for Testing Data
Class 0
Class 1
Count
Mean predicted value
Classification Report for Training set
Classification Report for Testing set
Predicted Class
Observed Class
5
44
266
53
4
37
55
37
Correlation Between Features
target
Features Importance
Scaled score (using min-max scaler)
lasso
Permutation
Cough & LossTaste & Runnynose
ExcessSweat & Fever & LossSmell
Headaches & Pinkeye
Chills & Fatigue & Wheezing
Chestpain & Chills & Fatigue
HistoryofRespSymp
Diarrhea & MuscleAches & Runnynose
Runnynose
Cough & Runnynose
Fever & Headaches
Fatigue & Headaches & MuscleAches
Headaches & LossTaste
Cough & ExcessSweat & LossSmell
Headaches
Cough & LossAppetite
AbdominalPain & MuscleAches
Diarrhea
Chestpain & Chills & MuscleAches
Chestpain & Chills
Learning Curve
Training score
Test score
Accuracy Score
Training Set Size
AUC vs Number of attributes
AUC
Number of Features
Cough & LossTaste & Runnynose distribution
Train
Test
Random
Cough & LossTaste & Runnynose Prediction Distribution
Train Data
Test Data
Cough & LossTaste & Runnynose=0
Cough & LossTaste & Runnynose=1
Prediction
Cough & LossTaste & Runnynose Prediction Distribution based on fixed values
Random Data
Cough & LossTaste & Runnynose=0.0
Cough & LossTaste & Runnynose=1.0

The Survey

Dashboard for Machine Learning Models in Health Care

Start of Block: Default Question Block

D1 **INFORMED CONSENT FORM**
RESEARCH PROCEDURES The purpose of the project is to collect one-time responses to survey questions related to the evaluation of the dashboard will be administered. The survey will ask about the participants' emails, positions, majors, evaluations, and comments regarding the dashboard. The participants are asked to:

- Grant the research team access to the survey responses
- Complete the survey

The dashboard summarizes the most important factors for evaluating any classification supervised ML model in one place using a dashboard. The dashboard is split into three main sections: statistical measures, features importance, and sensitivity analysis. The survey results will be used to explain the participant's opinions and evaluations of the dashboards.

D2 RISKS
There are no foreseeable risks.

D3 BENEFITS
There are no direct benefits to you for participating. The study aims to improve the communication and evaluation of the machine learning model results. The result of the study will be sent to the participants via email.

D4 CONFIDENTIALITY
No actual sensitive data will be included. The data will be stored in Qualtrics and then in the George Mason University DSHI secure server. After removing the identification information (Emails), the de-identified data will be in the researchers' personal computers and George Mason University DSHI secure server. The email is required to share the study results with the participants. Participants will receive by email a copy of the manuscript summarizing the research. If any of the identifiable information is provided in the responses, it will be removed. The de-identified data could be used for future research without additional consent from participants. While it is understood that no computer

transmission can be perfectly secure, reasonable efforts will be made to protect the confidentiality of your transmission.
The Institutional Review Board (IRB) committee that monitors research on human subjects may inspect study records during internal auditing procedures and are required to keep all information confidential.

D5 PARTICIPATION
Your participation is voluntary, and you may withdraw from the study at any time and for any reason. The survey will take about 15 minutes to complete. If you decide not to participate or if you withdraw from the study, there is no penalty or loss of benefits to which you are otherwise entitled. There are no costs to you or any other party.
To participate in the project, you need to be at least 18 years old.

D6 CONTACT
This research is being conducted as a thesis project by Wejdan Bagais, a master's student at the College of Health and Human Services at George Mason University who can be contacted by email at wbagais@gmu.edu for questions or to report a research-related problem and Supervised by Dr. Janusz Wojtusiak who can be contacted by email at jwojtusi@gmu.edu. You may contact the George Mason University Institutional Review Board (IRB) Office at 703-993-4121 or IRB@gmu.edu if you have questions or comments regarding your rights as a participant in the research. This research IRBNet number is 1766037-1.
This research has been reviewed according to George Mason University procedures governing your participation in this research.

D7 CONSENT
I have read this form and agree to participate in this study.

✱

Q1 Participant Email Address

Q2 Participant Signature

✱

Q3 Today's Date

End of Block: Default Question Block

Start of Block: Block 1

D8 **Introductions:** The purpose of the dashboard is to evaluate a machine learning model. The dashboard is split into three main sections: Statistical measures:

representation of the model accuracy measures Features' importance: representation of the most important features for the model Sensitivity analysis: representation of the effect of the input changes at the output results. The plot in this section is represented per feature.
You will review three examples of the dashboard, which are accessible by bellow links. When you review the dashboard, pay attention to each section's usefulness. The main purpose of the evaluation is to see if the dashboard gives important information that helps the reviewer better understand a machine learning model performance. Once you finish reviewing the dashboard, go to the survey and evaluate the usefulness of the plot's information.

D9 **Instructions You should review all three dashboards before answering any questions.** Open the dashboard looks at the overall performance of the model in the statistical measures section. Then check the top 20 features correlation and ranking in the feature's importance section. Finally, select a feature in the sensitivity analysis section to see how the model prediction behaves based on the selected feature.

D10 **Dashboards:**
Click here to access dashboard 1
Click here to access dashboard 2
Click here to access dashboard 3

End of Block: Block 1

Start of Block: Block 2

Q31
You should review all three dashboards (on the previous page) before answering any questions.

D11 Administrative questions

Q4 1. What is your position?

Q5 2. What is your major? (Area of work for facility, and study major for students)

D12 **Dashboard evaluation**

Q6 3. Do the dashboard visuals give you a better understanding of the model behavior than other methods you have experienced before? please explain your answer.

○ Yes (1)

○ No (2)

Q6.1 Explain

D13 **The next few questions ask about the specific sections in the dashboard (questions 4 to 6):**

Q7 4. Rate the **statistical results section** based on usefulness? please explain your answer.

○ Extremely useful (1)

○ Very useful (2)

○ Moderately useful (3)

○ Slightly useful (4)

○ Not at all useful (5)

Q7.1 Explain

Q8 5. Rate the **Feature importance section** based on usefulness? please explain your answer.

○ Extremely useful (1)

○ Very useful (2)

○ Moderately useful (3)

○ Slightly useful (4)

○ Not at all useful (5)

Q8.1 Explain

Q9 6. Rate the **sensitivity analysis section** based on usefulness? please explain your answer.

○ Extremely useful (1)

○ Very useful (2)

○ Moderately useful (3)

○ Slightly useful (4)

○ Not at all useful (5)

Q9.1 Explain

Q10 7. Is there anything else you would like to say about the dashboard (Ex. what is missing, what too much …)?

End of Block: Block 2

REFERENCES

Adams, N. M., & Hand, D. J. (1999). Comparing classifiers when the misallocation costs are uncertain. *Pattern Recognition*, *32*(7), 1139–1147. https://doi.org/10.1016/S0031-3203(98)00154-X

Alsallakh, B., Hanbury, A., Hauser, H., Miksch, S., & Rauber, A. (2014). Visual Methods for Analyzing Probabilistic Classification Data. *IEEE Transactions on Visualization and Computer Graphics*, *20*(12), 1703–1712. https://doi.org/10.1109/TVCG.2014.2346660

Amershi, S., Cakmak, M., Knox, W. B., & Kulesza, T. (2014). Power to the People: The Role of Humans in Interactive Machine Learning. *AI Magazine*, *35*(4), 105–120. https://doi.org/10.1609/aimag.v35i4.2513

Apley, D. W., & Zhu, J. (2020). Visualizing the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *82*(4), 1059–1086. https://doi.org/10.1111/rssb.12377

Brooks, M., Amershi, S., Lee, B., Drucker, S. M., Kapoor, A., & Simard, P. (2015). FeatureInsight: Visual support for error-driven feature ideation in text classification. *2015 IEEE Conference on Visual Analytics Science and Technology (VAST)*, 105–112. https://doi.org/10.1109/VAST.2015.7347637

Healy, K. J. (2019). *Data visualization: A practical introduction*. Princeton University Press.

Jasmina Dj. Novaković, Alempije Veljović, Siniša S. Ilić, Željko Papić, & Tomović Milica. (2017). Evaluation of Classification Models in Machine Learning. *Theory and Applications of Mathematics & Computer Science*, *7*(1). https://uav.ro/applications/se/journal/index.php/TAMCS/article/view/158

Jerome H. Friedman. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, *29*(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

Jiang, T., Gradus, J. L., & Rosellini, A. J. (2020). Supervised Machine Learning: A Brief Primer. *Behavior Therapy*, *51*(5), 675–687. https://doi.org/10.1016/j.beth.2020.05.002

Krause, J., Perer, A., & Ng, K. (2016). Interacting with Predictions: Visual Inspection of Black-Box Machine Learning Models. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (pp. 5686–5697). Association for Computing Machinery. https://doi.org/10.1145/2858036.2858529

Limited Data Set (LDS) Files | CMS. (2019, December 4). https://www.cms.gov/Research-Statistics-Data-and-Systems/Files-for-Order/LimitedDataSets/

Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 30* (pp. 4765–4774). Curran Associates, Inc.

http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf

Nussbaumer Knaflic, C. (2015). *Storytelling with Data: A Data Visualization Guide for Business Professionals* (1st ed.). John Wiley & Sons, Incorporated, Wiley, Wiley-Blackwell.

Raymaekers, J., Rousseeuw, P. J., & Hubert, M. (2020). Visualizing classification results. *ArXiv:2007.14495 [Cs, Stat]*. http://arxiv.org/abs/2007.14495

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). “Why Should I Trust You?”: Explaining the Predictions of Any Classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. https://doi.org/10.1145/2939672.2939778

Russell, S., & Norvig, P. (2021). *Artificial Intelligence: A Modern Approach* (4th Edition) [Kindle book]. Pearson Education. https://www.pearson.com/content/one-dot-com/one-dot-com/us/en/higher-education/program.html

Saltelli, A., Tarantola, S., Campolongo, F., & Ratto, M. (2004). *Sensitivity analysis in practice: A guide to assessing scientific models* (Vol. 1). Wiley Online Library.

Samek, W., Binder, A., Montavon, G., Lapuschkin, S., & Müller, K.-R. (2017). Evaluating the Visualization of What a Deep Neural Network Has Learned. *IEEE Transactions on Neural Networks and Learning Systems*, *28*(11), 2660–2673. https://doi.org/10.1109/TNNLS.2016.2599820

Swets, J. (1988). Measuring the accuracy of diagnostic systems. *Science*, *240*(4857), 1285. https://doi.org/10.1126/science.3287615

Tang, J., Alelyani, S., & Liu, H. (2014). Feature selection for classification: A review. *Data Classification: Algorithms and Applications*, 37.

Tibshirani, R., Bien, J., Friedman, J., Hastie, T., Simon, N., Taylor, J., & Tibshirani, R. J. (2012). Strong rules for discarding predictors in lasso-type problems. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, *74*(2), 245–266. PubMed. https://doi.org/10.1111/j.1467-9868.2011.01004.x

Tonekaboni, S., Joshi, S., McCradden, M. D., & Goldenberg, A. (2019). What Clinicians Want: Contextualizing Explainable Machine Learning for Clinical End Use. *ArXiv:1905.05134 [Cs, Stat]*. http://arxiv.org/abs/1905.05134

Tufte, E. R. (2001). *The visual display of quantitative information* (2nd ed.). Graphics Press.

UCI. (1988, July 1). *UCI Machine Learning Repository: Heart Disease Data Set* [Education]. UCI MAchine Learning Repository. https://archive.ics.uci.edu/ml/datasets/Heart+Disease

Wexler, J., Pushkarna, M., Bolukbasi, T., Wattenberg, M., Viégas, F., & Wilson, J. (2020). The What-If Tool: Interactive Probing of Machine Learning Models. *IEEE Transactions on Visualization and Computer Graphics*, *26*(1), 56–65. https://doi.org/10.1109/TVCG.2019.2934619

Wojtusiak, J. (2021). Reproducibility, Transparency and Evaluation of Machine Learning in Health Applications. *HEALTHINF*, 685–692.

Wojtusiak, J., Elashkar, E., & Mogharab Nia, R. (2018). C-LACE2: Computational risk assessment tool for 30-day post hospital discharge mortality. *Health and Technology*, *8*(5), 341–351. https://doi.org/10.1007/s12553-018-0263-1

Zhang, J., Wang, Y., Molino, P., Li, L., & Ebert, D. S. (2019). Manifold: A Model-Agnostic Framework for Interpretation and Diagnosis of Machine Learning Models. *IEEE Transactions on Visualization and Computer Graphics*, *25*(1), 364–373. https://doi.org/10.1109/TVCG.2018.2864499

BIOGRAPHY

Wejdan H Bagais graduated with a bachelor's degree in Information Technology. She received her Bachelor of Science from King Abdulaziz University, Jeddah, Saudi Arabia, in 2015. She was employed as a sales analyst in Saudi Arabia for two years, where she worked on implementing a new method for analyzing and predicting the sales trend performance, which improved the prediction accuracy. She is currently volunteering as a web developer at MAP clinics, where she developed a new feature for their OpenEMR website using SQL, Linux server, PHP, HTML, and CSS. She is responsible for creating patients' panels to follow patients with specific health care needs. Additionally, she works with the George Mason MLi Lab team on a project related to COVID-19 symptoms.