Heart Disease Prediction: Classification Models and Risk Factor Analysis

Group 19:Baihan Liu, Wencong Liang, Yiwei Mai, Yuwen He

04/28/2023

1 Introduction

Heart disease remains a leading cause of death in the United States, accounting for an estimated 655,000 fatalities each year (CDC, 2021). In light of its high prevalence, identifying the risk factors associated with heart disease and devising effective prevention strategies are of utmost importance. The present study aims to develop classification models that predict the likelihood of heart disease based on an individual's Body Mass Index (BMI), lifestyle, mental health, and other relevant factors. Additionally, the study will estimate the impact of prominent risk factors, such as smoking and alcohol consumption, on heart disease risk.

The dataset used in this study was sourced from Kaggle and originally collected by the Centers for Disease Control and Prevention (CDC) as part of the Behavioral Risk Factor Surveillance System (BRFSS). The BRFSS conducts annual telephone surveys to gather data on the health status of U.S. residents. The dataset's most recent version (as of February 15, 2022) features data from 2020 and comprises 401,958 rows and 279 columns. For this analysis, the most pertinent 18 variables were selected. The "HeartDisease" variable serves as a binary response, with "Yes" denoting that the respondent has heart disease and "No" indicating its absence.

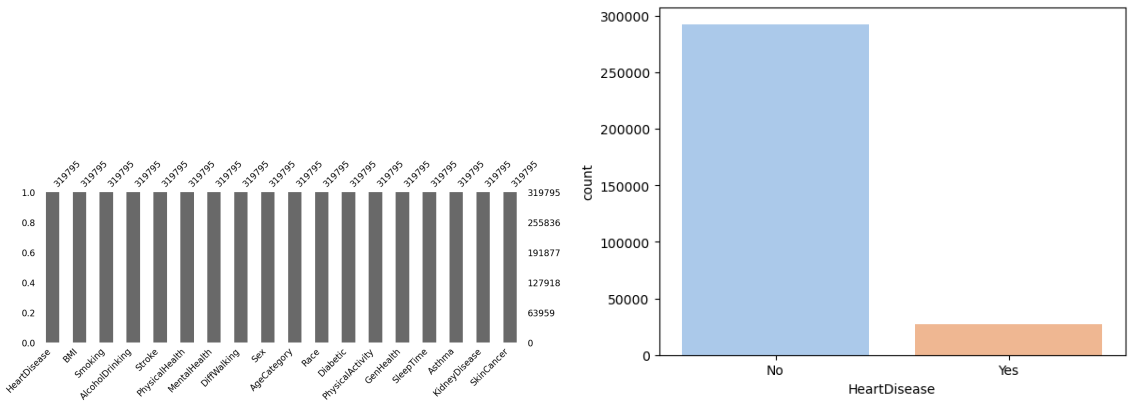
The primary objectives of this study are to develop classification models for heart disease prediction capable of accurately determining an individual's risk of developing heart disease based on their BMI, lifestyle, mental health, and other relevant factors, and to evaluate the effects of prominent risk factors on heart disease risk, such as smoking and alcohol consumption. By accomplishing these objectives, this study aims to provide valuable insights into heart disease risk factors and contribute to the development of effective prevention strategies.

2 Exploratory Data Analysis (EDA)

Before diving into the main analysis, it is crucial to explore the dataset to gain a better understanding of the data and identify any potential issues that may affect the subsequent steps. The EDA process involves the following steps:

2.1 Missing Values and Imbalanced Data

During the EDA, we checked for missing values and found that all variables had complete data. However, we noticed an imbalance in the number of records for HeartDisease, with an uneven data distribution between those with and without heart disease. To address this class mbalance, we applied Synthetic Minority Oversampling Technique (SMOTE) and considered using metrics such as Area Under the Curve (AUC) instead of classification accuracy.



(a) age (b) stroke

Figure 1: missing data and imbalenced labels

2.2 Numerical Variables

The distribution analysis of numerical variables reveals that BMI and SleepTime are asymptotically normally distributed, with their values more clustered in the center. On the other hand, PhysicalHealth and MentalHealth exhibit bimodal distributions, with values clustered at both extremes.Besides,judged by interquantile range,there are 10396 outliers in variable BMI, 47146 in PhysicalHealth,51576 in MentalHealth and 4543 in SleepTime.

Based on these observations, we will treat PhysicalHealth and MentalHealth as categorical variables for our subsequent analysis.

2.3 Categorical Variables

Among the categorical variables, we draw we bar plots on the variables and identified Stroke, DiffWalking, AgeCategory, Race, and several other variables as potentially important indicators for detecting heart disease. By focusing on these variables, we can potentially improve the accuracy of our predictive models and better understand the key risk factors associated with heart disease.

2.4 Correlation analysis

Lastly, we examined the correlation matrix to identify any potential multicollinearity issues. We found low to moderate correlations between variables, with no strong correlation above 0.5. This indicates that our dataset does not suffer from multicollinearity and can be used effectively in predictive modeling.

We also use hierarchical clustering and complete linkage method, as shown in Figure 5, to explore the correlations among features using dissimilar Matrix. One of the advantages of Complete linkage is that it suffers little from the noise of our data compared to single linkage. We can cluster the features into four clusters by the red cutting line. The features in a cluster have relatively high correlation. For example, we find that features BMI, Stroke, PhysicalHealth, MentalHealth, DiffWalking, Diabetic, Asthma, KidneyDisease have relatively high correlation with our predicted variable HeartDisease.

2.5 Principal Components Analysis

From the heatmap analysis above, we conclude that dimension reduction methods are not necessary for the data set. However, we can perform principal component analysis to have 2D and 3D visualizations, as

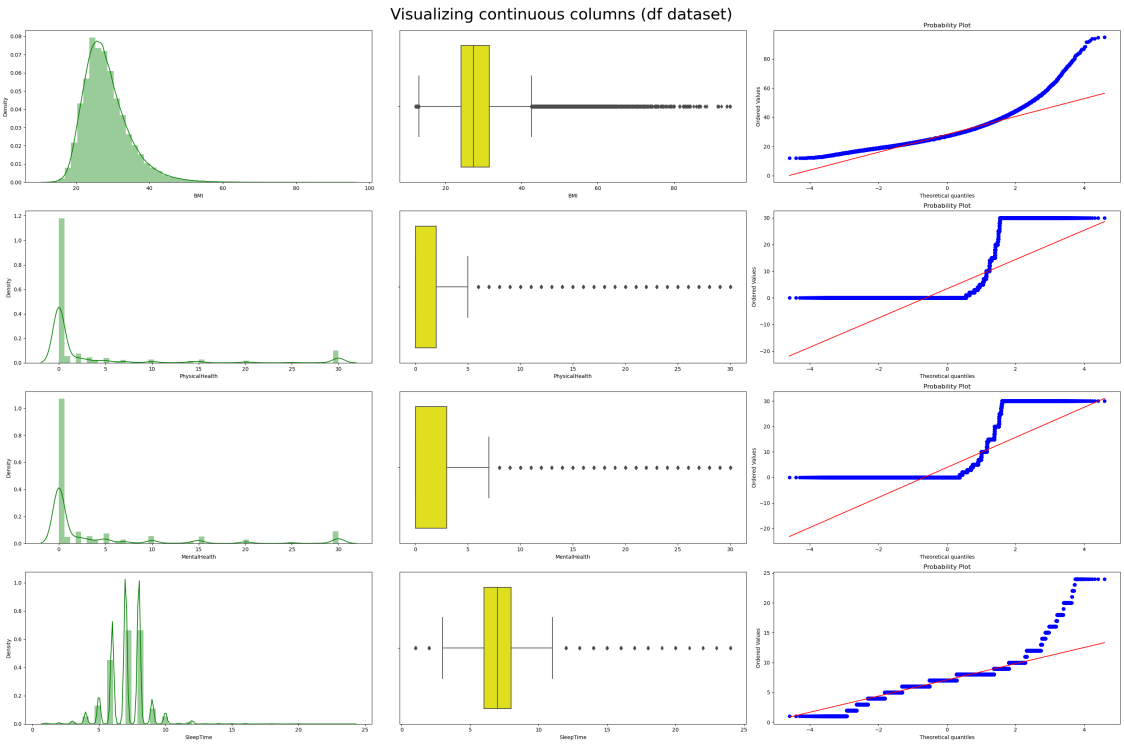


Figure 2: numerical variables

shown in Figure 6 of our data using four numerical variables. We also plot the scree plot which shows the variances explained by each principal component.

We cannot see a clear pattern in the two dimensions and three dimensions plot. This confirms our analysis that there is no principal component that dominates the variance. Thus, principal component analysis is not a good choice for numerical data in our data set.

3 Data Preprocessing

- Data Transformation

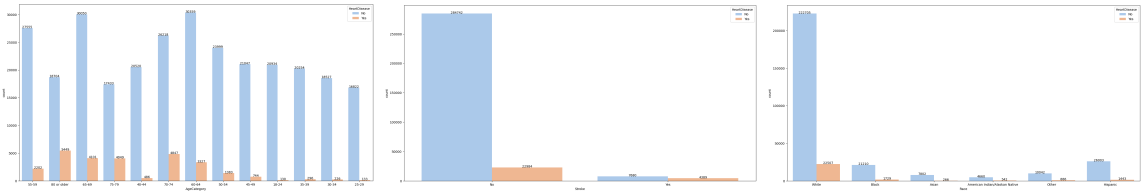
  For numerical variables with bimodal distributions, namely, PhysicalHealth and MentalHealth, we transform them into categorical variables by setting cutting values.

- Data Encoding

  We try different methods to encode the categorical variables in our dataset. The first one is ordinal encoding. Ordinal encoding converts each label into integer values and the encoded data represents the sequence of labels. The second one is dummy variable encoding. The Dummy coding scheme is similar to one-hot encoding. This categorical data encoding method transforms the categorical variable into a set of binary variables (also known as dummy variables). In the case of one-hot encoding, for N categories in a variable, it uses N binary variables. The dummy encoding is a small improvement over one-hot-encoding. Dummy encoding uses N-1 features to represent N labels/categories.

- Data Normalization

  For some methods that are distance-based, we need to standardize our data. We use StandardScaler to normalize our data, which standardizes a feature by subtracting the mean and then scaling to unit



(a) age (b) stroke (c) race

Figure 3: bar plot of three categorical variables

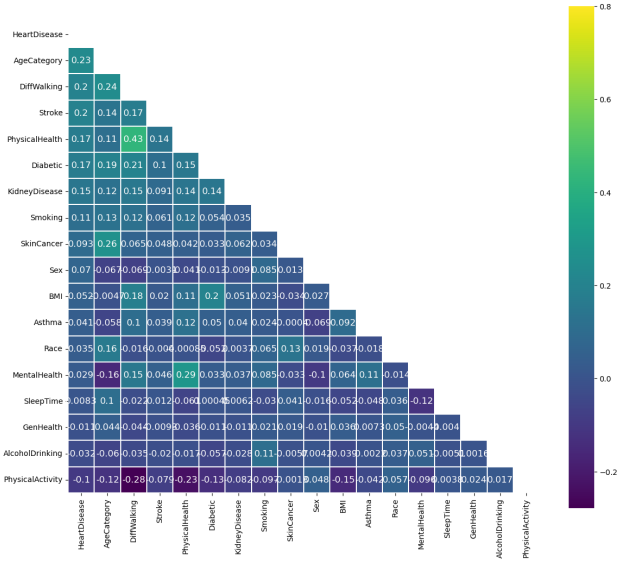


Figure 4: Heatmap matrix

variance.

- Data Augmentation

  We use Synthetic Minority Oversampling Technique (SMOTE) to handle the imbalanced data. It duplicates examples in the minority class.

- Data Splitting

  We randomly split our data into training set(80%) and test set(20%). Training set will be further divided into training and validation sets in the Cross-Validation approach in the our classification methods. Training set will be used to train our models, while validation set will be used to find hyperparameters of our models. Testing set will be used to evaluate the performance of our models.

4 Statistical Methods

To find the best way to predict heart disease with given and chosen indicators, we will try different classifier methods including: XGBoost, Support Vector Machine, K Neighborhood Nearest and Logistic Regression. Besides that our group will use confusion matrix, AUC score, BER score, precision, recall as well as F1 score to evaluate the effectiveness of the machine learning models

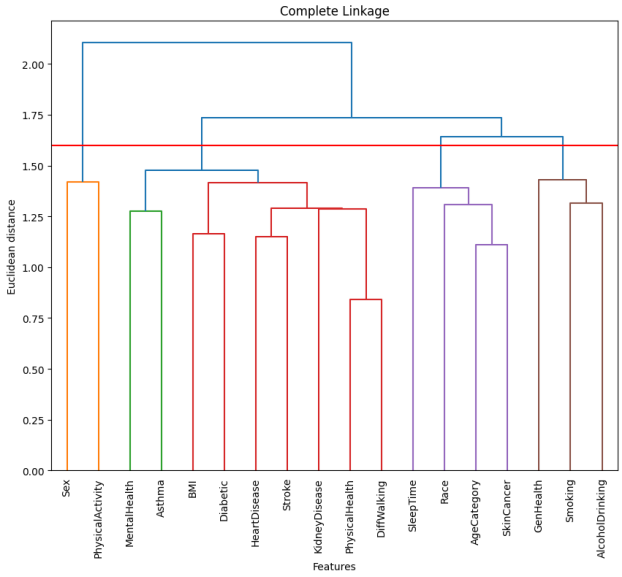
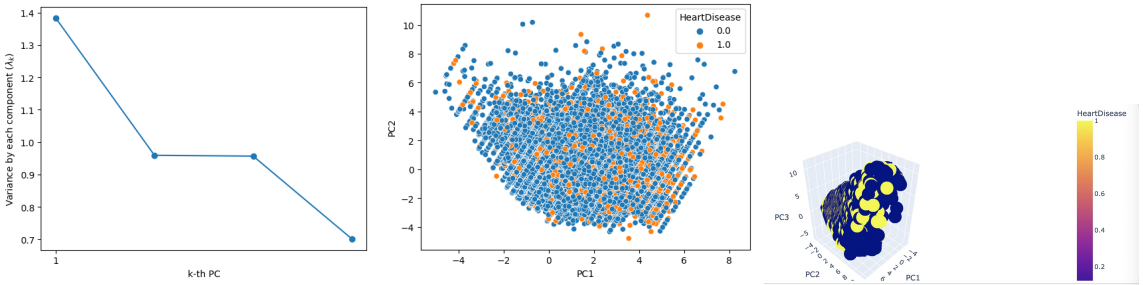


Figure 5: hierarchical clustering



(a) Scree plot (b) 2D visualization (c) 3D visualization

Figure 6: PCA analysis

We use GridSearch 5-fold Cross-validation to select the best hyperparameters based on the AUC score for the following three methods.

4.1 Model Construction

For K nearest neighbors, the hyperparameter is k. We choose a list of k that is around the square root of the number of data points. To reduce the computational cost, we use 10000 data points for parameters' searching and fit our model using all the data points.

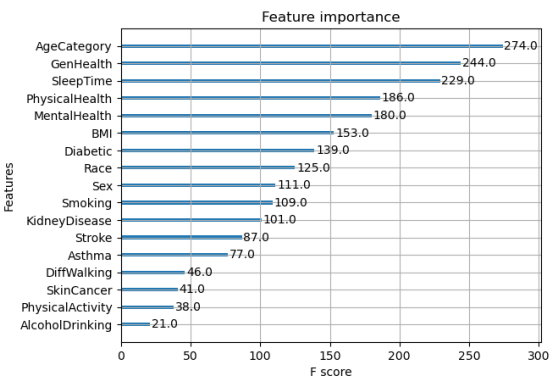
For logistic regression, we first perform the hypothesis testing and find the features' significance. We found that all variables are significant since their p-values are less than 0.05. The regularization hyperparameter C controls the size of the penalty term. For penalty, we have two options: l1-norm and l2-norm.

For support vector machine, the hyperparameters are regularization parameter C and gamma values for different types of kernel functions, specifically, linear kernel and rbf kernel.

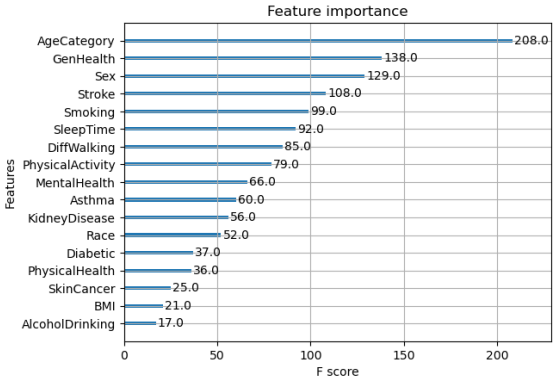
For XGBoost,the hyperparameters that we used to construct XGBoost are as followed: 'objective'='binary:logistic', $'max\_depth'$=8, $'eval\_metric'$='auc', and the rest are as default.

4.2 Model Interpretation

4.2.1 XGboost



(a) non smote



(b) smote

In this section, we will interpret the results of our XGBoost model by examining its feature importance plot. Since we have built models on three different encoded data, we got three feature importance plots. Since the plot for data with dummy variables is not explicit enough, we focus on the other two.

The left plot is for the data applying SMOTE but no dummy variables. From this plot, we can see that the most important feature in this model is 'AgeCategory', followed by 'GenHealth', which means how healthy a person feels for himself in general, and 'Sex'. These features have the highest weights and contribute the most to the prediction of heart disease. The next most important features are 'Stroke' and 'Smoking' (Meaning: Have you smoked at least 100 cigarettes in your entire life?), which also have a significant impact on the prediction of heart disease.

The right plot is for the data not applying SMOTE and no dummy variables. From this plot, we can see that the most important feature in this model is 'AgeCategory', followed by 'GenHealth' and 'SleepTime' (The average hours a person sleeps in a 24 hour period). These features have the highest weights and contribute the most to the prediction of heart disease. The next most important features are 'PhysicalHealth' and 'MentalHealth', which also have a significant impact on the prediction of heart disease.

4.2.2 Logistic Regression

$$\log \frac{P(\text{HeartDisease} = 1|X)}{P(\text{HeartDisease} = 0|X)} = \beta_0 + X^T\beta$$

The p-values show that all variables are statistically significant. If "AgeCategory" increased by 1 unit, log odds of the heart risk would decrease by 0.1745. If "SleepTime" increased by 1 unit, log odds of the heart risk would decrease by 0.3658.

Therefore,based on the logistic regression summary,the increase in Smoking,Stroke higher the risk of heart disease while the increase in PhysicalActivity and MentalHealth can prevent people from heart disease.

Also,it shows that men are at greater risk of heart disease.Diabetic and people with SkinCancer are more likely to suffering from heart disease.These indications correspond to our EDA analysis.

4.3 Performance Comparison

Table 2 and Figure 7 shows that XGBoost outperforms the other models in terms of most of the evaluation metrics. It has the lowest BER score, highest precision score, and highest recall score. It also has a relatively low log loss score.On the other hand, KNN has the highest BER score and lowest F1 score, indicating that

Table 1: logistic regression

	coef	std err	z	P>—z—	[0.025	0.975]
Smoking	0.2711	0.013	20.117	0.000	0.245	0.298
AlcoholDrinking	-0.4207	0.033	-12.717	0.000	-0.486	-0.356
Stroke	1.2281	0.023	54.023	0.000	1.184	1.273
PhysicalHealth	0.0148	0.001	19.750	0.000	0.013	0.016
MentalHealth	-0.0115	0.001	-13.413	0.000	-0.013	-0.010
DiffWalking	0.2904	0.018	16.594	0.000	0.256	0.325
Sex	0.4386	0.013	32.741	0.000	0.412	0.465
AgeCategory	0.1745	0.002	72.921	0.000	0.170	0.179
Race	-0.2453	0.005	-53.567	0.000	-0.254	-0.236
Diabetic	0.2437	0.008	31.574	0.000	0.229	0.259
PhysicalActivity	-0.6013	0.014	-43.108	0.000	-0.629	-0.574
GenHealth	-0.1447	0.005	-30.695	0.000	-0.154	-0.135
SleepTime	-0.3658	0.003	-107.547	0.000	-0.372	-0.359
Asthma	0.0843	0.019	4.504	0.000	0.048	0.121
KidneyDisease	0.7502	0.024	30.746	0.000	0.702	0.798
SkinCancer	0.4532	0.019	23.669	0.000	0.416	0.491

it may not be the most effective model for predicting heart disease. Logistic Regression and SVM perform similarly to each other, but not as well as XGBoost.

Overall, in our project,XGBoost appears to be the best model for predicting heart disease based on these evaluation metrics.

Table 3 compares the performance of three XGBoost models with different data encoding methods. The XGBoost model with SMOTE and dummy encoding outperforms the other models with the highest AUC Score, precision score, recall score, and the lowest BER score and log loss score. On the other hand, the XGBoost model with SMOTE encoding but without dummy encoding has the lowest precision and recall scores, as well as the highest BER score and log loss score. The XGBoost model without SMOTE and dummy encoding has the lowest f1 score and the second-highest BER score.

Table 2: performances of models on different metrics

Methods	AUC Score:	BER Score	f1 Score	Precision Score	Recall Score	Log Loss Score
KNN	0.8	0.28	0.31	0.2	0.71	9.1
Logistic Regression	0.77	0.31	0.29	0.19	0.66	9.5
Svm	0.77	0.3	0.29	0.19	0.67	9.74
XGBoost	**0.876**	0.196	0.802	0.846	0.762	0.459

Table 3: performances of 3 data encoding XGBoost models

Methods	AUC Score:	BER Score	f1 Score	Precision Score	Recall Score	Log Loss Score
SMOTE,no Dummy	0.814	0.323	0.345	0.474	0.272	0.331
no SMOTE,no Dummy	0.841	468	0.123	0.069	0.559	0.232
SMOTE,Dummy	0.876	0.196	0.802	0.846	0.762	0.459

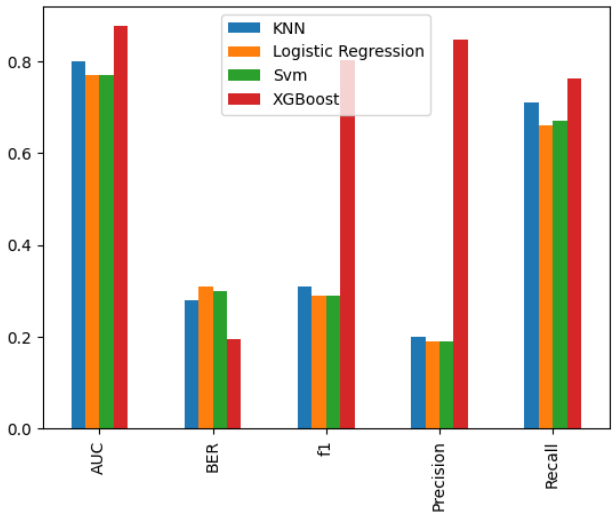


Figure 7: Experiment Results

5 Conclusion

The analysis revealed that age category and general health were consistently among the most important features for predicting heart disease in both scenarios. However, the importance of other features varied depending on the inclusion of SMOTE and dummy variables. In addition to age category and general health, features such as sleep time, sex, stroke history, mental health, BMI, and others also demonstrated significant importance in predicting heart disease. These results suggest that a combination of demographic, health-related, and behavioral factors are important for predicting heart disease. Overall, the findings highlight the importance of considering a broad range of factors in predicting heart disease and developing targeted interventions to reduce its risk.

In conclusion, preventing cardiovascular disease requires individuals to adopt several essential measures. These include maintaining a healthy diet, controlling weight and abdominal circumference, engaging in appropriate physical activity, quitting smoking, alcohol, and drug addiction, managing stress, monitoring blood pressure, cholesterol, and sugar levels, and scheduling preventive examinations and regular visits to a cardiologist. By following these guidelines, individuals can significantly reduce their risk of developing cardiovascular diseases and maintain good heart health. It is essential to prioritize heart health through lifestyle changes and medical interventions to prevent cardiovascular disease and ensure a healthy life.

6 Future work

In future work, we will remove the outliers of our data in the preprocessing step. We will also consider the evaluation metric MCC(Matthews correlation coefficient) to evaluate our model performance. The MCC is a measure of the correlation between predicted and actual samples, with a range of values from -1 to 1. It is more robust to imbalanced categories compared to accuracy-based metrics.

7 References

- Deep, Mala. 2022. “Easy Way of Finding and Visualizing Missing Data in Python.”Medium, January 6, 2022.
- Nkitgupta. 2022. “Advance Data Preprocessing.” Kaggle, April.

- Georgyzubkov. 2022. “Heart Disease. Exploratory Data Analysis.” Kaggle, August.
- Berge, Eirik, PhD. 2022. “Visualizing Missing Values in Python Is Shockingly Easy.” Medium, January 4, 2022.
- “Personal Key Indicators of Heart Disease.” 2022. Kaggle. February 16, 2022.
- The 5 Classification Evaluation metrics every Data Scientist must know. September 17, 2019.