Feature Selection:

* Given that the dataset solely contains numerical values, a VIF test was conducted to identify the variables that are multicollinear so that they can be omitted before creating the models.
* After exploring various combination, the final models used the following features, which had a VIF score below 10.

|  |  |  |
| --- | --- | --- |
|  | Features | VIF score |
| 1. | HighBP | 2.376639 |
| 2. | HighChol | 2.075802 |
| 3. | Smoker | 1.983754 |
| 4. | Stroke | 1.108968 |
| 5. | Diabetes | 1.419802 |
| 6. | PhysActivity | 4.144036 |
| 7. | Fruits | 2.828395 |
| 8. | Veggies | 5.287618 |
| 9. | HvyAlcoholConsump | 1.092803 |
| 10. | NoDocbcCost | 1.219131 |
| 11. | MentHlth | 1.469514 |
| 12. | PhysHlth | 2.009715 |
| 13. | DiffWalk | 1.847508 |
| 14. | Sex | 1.888215 |

Modeling on the imbalanced dataset:

* The goal was to determine what features caused a person to have a heart disease or attack.
* As the dataset is heavily imbalanced, the models were performed well when it came to predicting no heart disease or attack.
* The following table shows how the models performed on the imbalanced test dataset.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Name of the model | Recall | | Accuracy | AUC |
| **0 – no heart disease or attack** | **1 – yes heart disease or attack** |
| Naïve Bayes | 0.85 | 0.55 | 0.81 | 0.69 |
| Logistic Regression | 0.99 | 0.11 | 0.89 | 0.55 |
| Decision Tree Classifier | 0.99 | 0.08 | 0.89 | 0.53 |
| Random Forest Classifier | 1.00 | 0.39 | 0.90 | 0.53 |
| XGBoost Classifier | 0.99 | 0.18 | 0.89 | 0.55 |

* 1. Looking at the table, it is clear that Random Forest Classifier performs the best when compared with other models with respect to recall. But there is a high possibility that the model is being over-fit. The reason for this is due to the heavy imbalance within the dataset.
  2. Focusing on recall is important since the model should be able to accurately predict true positives while minimizing false negatives.

Balancing Technique:

* SMOTE or Synthetic Minority Oversampling Technique is an improved alternative for oversampling.
* By generating artificial data points based on the real data points, the SMOTE algorithm conducts data augmentation.
* The advantage of SMOTE is that, it is not generating duplicates, but rather creating synthetic data points that are slightly different from the original data points.

The SMOTE algorithm works as follows:

1. Draw a random sample from the minority class.
2. For the observations in the sample, identify the k-nearest neighbors.
3. Consider one of those neighbors and identify the vector beterrn the current data point and the selected neighbor.
4. Multiply the vector by a random number between 0 and 1.
5. To obtain the synthetic data point, add the multiplied vector to the current data point.

* SMOTE attempts to balance the dataset into a 50:50 ratio because the existing dataset is unbalanced, with yes heart disease or attack having 23717 datapoints and no heart disease or attack having 206064 datapoints.
* After balancing there 108186 data points present in both yes heart disease or no heart disease or attack.

Modeling on balanced dataset:

* The objective was to identify the features that contributed to heart disease or attack in an individual.
* The following are the results of all the models on the balanced dataset.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Name of the model | Recall | | Accuracy | AUC |
| **0 – no heart disease or attack** | **1 – yes heart disease or attack** |
| Naïve Bayes | 0.78 | 0.61 | 0.76 | 0.69 |
| Logistic Regression | 0.73 | 0.71 | 0.72 | 0.71 |
| Decision Tree Classifier | 0.77 | 0.56 | 0.74 | 0.66 |
| Random Forest Classifier | 0.70 | 0.72 | 0.70 | 0.71 |
| XGBoost Classifier | 0.85 | 0.43 | 0.80 | 0.63 |

1. From the above table, it can be concluded that Random Forest Classifier performs better for both yes heart disease or attack and no heart disease or attack.