**About the Dataset:**

Diabetes is one of the most prevalent diseases in India, where I come from and in the US where I live. It impacts hundreds of millions of people around the world. It is a chronic disease that affects how your body produces and uses insulin. Insulin is a hormone that helps your body turn glucose (sugar) from food into energy.

There are two main types of diabetes:

**Type 1 diabetes:** This type is caused by the body's immune system attacking and destroying the insulin-producing cells in the pancreas. People with type 1 diabetes must take insulin injections or use a pump to manage their blood sugar levels.

**Type 2 diabetes:** This type is caused by the body becoming resistant to insulin or not producing enough insulin. People with type 2 diabetes may need to take medication, follow a healthy diet, and exercise regularly to manage their blood sugar levels.

The dataset I am using in this project is **Diabetes Health Indicators Dataset** from Kaggle. This dataset has 21 independent variables and a class variable (Diabetes\_012). It contains 253,680 rows of data. The file is in csv format.

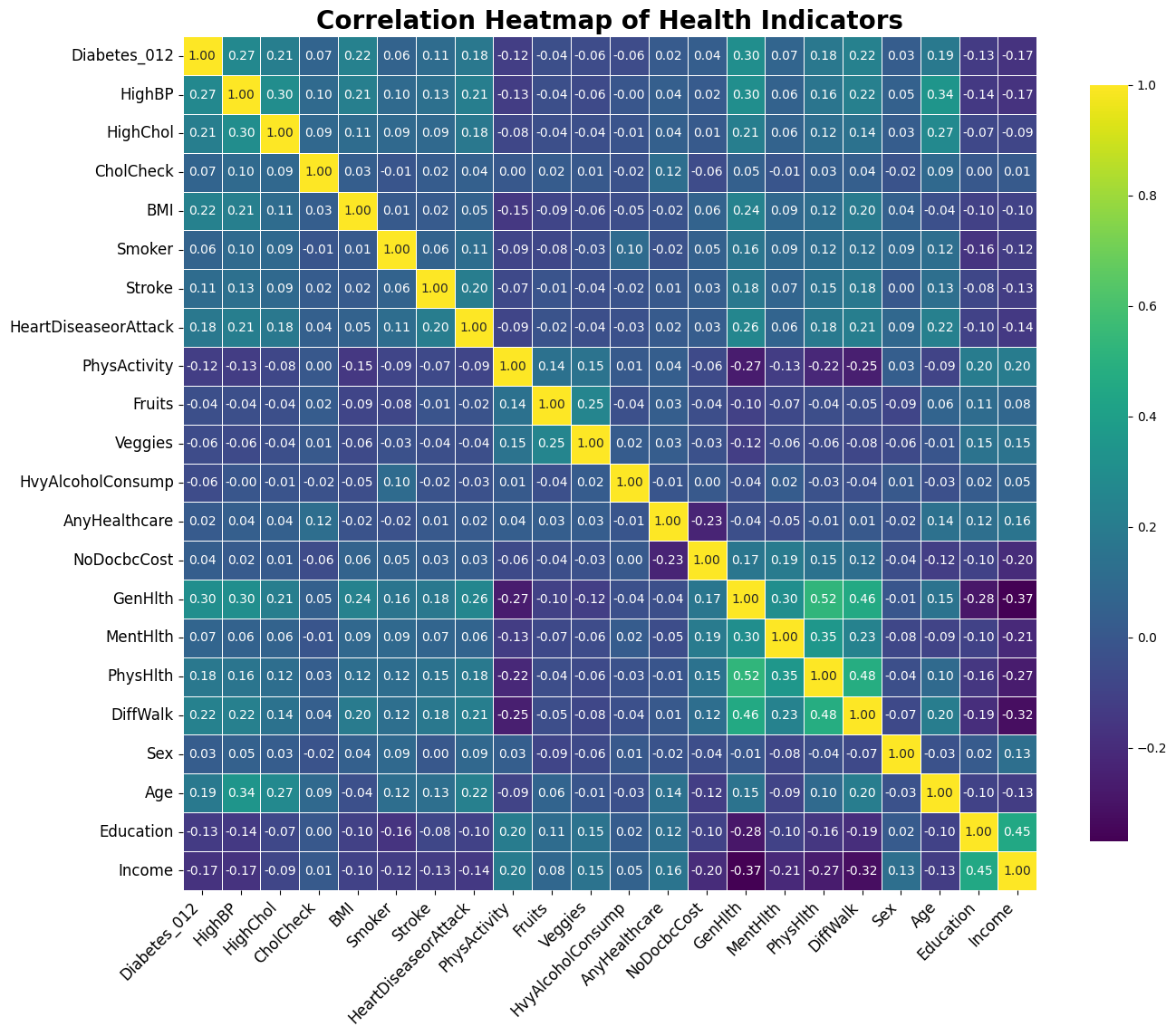
|  |  |
| --- | --- |
| **Feature** | **Description** |
| Diabetes\_012 | 0 = no diabetes 1 = prediabetes 2 = diabetes. |
| High blood pressure | 0 = no high BP 1 = high BP. |
| high cholesterol | 0 = no high cholesterol 1 = high cholesterol. |
| Cholesterol Check | 0 = no cholesterol check in 5 years 1 = yes cholesterol check in 5 years. |
| BMI | Body Mass Index. |
| Smoker | Have you smoked at least 100 cigarettes in your entire life? 0 = no 1 = yes. |
| Stroke | you had a stroke. 0 = no 1 = yes |
| HeartDiseaseorAttack | coronary heart disease (CHD) or myocardial infarction (MI) 0 = no 1 = yes |
| PhysActivity | physical activity in past 30 days - not including job 0 = no 1 = yes |
| Fruits | Consume Fruit 1 or more times per day 0 = no 1 = yes |
| Veggies | Consume Vegetables 1 or more times per day 0 = no 1 = yes. |
| HvyAlcoholConsump | Heavy drinkers (adult men having more than 14 drinks per week and adult women having more than 7 drinks per week) 0 = no 1 = yes. |
| AnyHealthcare | Have any kind of health care coverage, including health insurance, prepaid plans such as HMO, etc. 0 = no 1 = yes. |
| NoDocbcCost | Was there a time in the past 12 months when you needed to see a doctor but could not because of cost? 0 = no 1 = yes. |
| GenHlth | Would you say that in general your health is: scale 1-5 1 = excellent 2 = very good 3 = good 4 = fair 5 = poor. |
| MentHlth | mental health scale, scale 1-30 days |
| PhysHlth | Now thinking about your physical health, for how many days during the past 30 days was your physical health not good? scale 1-30 days |
| DiffWalk | Do you have serious difficulty walking or climbing stairs? 0 = no 1 = yes. |
| Sex | 0 = female 1 = male |
| Age | 13-level age category  1 = 18-24  2 = 25-29  9 = 60-64  13 = 80 or older |
| Sex | 0 = female 1 = male |
| Education | Education level (EDUCA see codebook) scale 1-6 1 = Never attended school or only kindergarten, 2 = Grades 1 through 8 (Elementary) ,3 = Grades 9 through 11 (Some high school), 4 = Grade 12 or GED (High school graduate), 5 = College 1 year to 3 years (Some college or technical school) , 6 = College 4 years or more (College graduate). |
| Income | Income scale, scale 1-8 1 = less than $10,000 5 = less than $35,000 8 = $75,000 or more. |

**Distribution of Data:**

Most of the data is of class ‘No Diabetes’. Only 1.8% of the data is of class ‘Pre-Diabetes’. So, I ran machine learning models without and with handling the imbalance of the data. I found the results to be surprising, which I will discuss later in this report.

A close-up of a graph

Description automatically generated

**Correlation of the features**:

The above heatmap shows the correlation of the target variable **Diabetes\_012** with other health indicators. The features with relatively higher correlation (positive or negative) are:

1. **HighBP (0.27)**: **Diabetes\_012** shows a moderate positive correlation with high blood pressure (HighBP), suggesting that individuals with diabetes are more likely to have elevated blood pressure. This is consistent with clinical evidence showing that hypertension is a common comorbidity in diabetes.
2. **HighChol (0.21)**: There is a weak positive correlation between diabetes and high cholesterol (HighChol). While the relationship is not as strong, it still indicates that people with diabetes may have higher cholesterol levels, which can lead to further health complications.
3. **BMI (0.22)**: The positive correlation between BMI and diabetes indicates that individuals with higher BMI tend to have diabetes. This aligns with well-known research that links obesity to a higher risk of developing type 2 diabetes.
4. **GenHlth (-0.30)**: There is a moderate negative correlation between diabetes and general health, indicating that individuals with diabetes often report poorer general health, which is expected given the impact of the disease on overall well-being.
5. **DiffWalk (0.22)**: The moderate positive correlation between diabetes and difficulty walking highlights the mobility challenges that individuals with diabetes might face, possibly due to complications like neuropathy.

**Machine Learning Models:**

The goal is to accurately predict **No Diabetes**, **Pre-Diabetes**, and **Diabetes**. So, this a multinomial target variable. I ran various machine learning models with different Train/Test splits to get the best performance. Initially I did not balance the data or do feature reduction. On the Linear regression I ran different iterations, regularization parameters and landed on 100 iterations with a mix of L1 and L2 regularization, mostly L1 with some L2. I ran different iterations depths for other classifiers as well and landed on the ones that gave best results. The results are below:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Train-Test Split** | **Model** | **Accuracy** | **F1-Score** | **Precision** | **Recall** |
| 50% train, 50% test | Logistic Regression | 0.846 | 0.7984 | 0.7966 | 0.846 |
|  | Decision Tree | 0.8481 | 0.8027 | 0.8014 | 0.8481 |
|  | Random Forest | 0.8488 | 0.7991 | 0.8052 | 0.8488 |
|  | Naive Bayes | 0.7571 | 0.7642 | 0.7731 | 0.7571 |
|  | SVM | 0.8423 | 0.7702 | 0.7094 | 0.8423 |
|  | XGBClassifier | 0.842 | 0.8104 | 0.7976 | 0.842 |
| 60% train, 40% test | Logistic Regression | 0.8463 | 0.7984 | 0.7968 | 0.8463 |
|  | Decision Tree | 0.8488 | 0.8033 | 0.8029 | 0.8488 |
|  | Random Forest | 0.8488 | 0.7985 | 0.8052 | 0.8488 |
|  | Naive Bayes | 0.7565 | 0.7639 | 0.773 | 0.7565 |
|  | SVM | 0.8425 | 0.7705 | 0.7098 | 0.8425 |
|  | XGBClassifier | 0.8432 | 0.8106 | 0.7979 | 0.8432 |
| 70% train, 30% test | Logistic Regression | 0.8479 | 0.8004 | 0.7996 | 0.8479 |
|  | Decision Tree | 0.8498 | 0.8047 | 0.804 | 0.8498 |
|  | Random Forest | 0.8503 | 0.8008 | 0.8073 | 0.8503 |
|  | Naive Bayes | 0.7577 | 0.7656 | 0.7749 | 0.7577 |
|  | SVM | 0.8435 | 0.7719 | 0.7115 | 0.8435 |
|  | XGBClassifier | 0.8458 | 0.8136 | 0.8017 | 0.8458 |
| 80% train, 20% test | Logistic Regression | 0.8484 | 0.8011 | 0.7993 | 0.8484 |
|  | Decision Tree | 0.8505 | 0.8056 | 0.804 | 0.8505 |
|  | Random Forest | 0.8509 | 0.8028 | 0.8064 | 0.8509 |
|  | Naive Bayes | 0.7601 | 0.7679 | 0.7773 | 0.7601 |
|  | SVM | 0.8446 | 0.7734 | 0.7133 | 0.8446 |
|  | XGBClassifier | 0.8475 | 0.8151 | 0.8034 | 0.8475 |

**Overall performance:**

* Most models perform relatively well, with accuracies generally above 0.80 across different train-test splits.
* XGBClassifier and Random Forest tend to be among the top performers consistently.

**Impact of train-test split:**

* As the training set size increases from 50% to 80%, there's a general trend of very slight improvement in model performance across most metrics. It didn’t make much difference overall.

**Model comparisons:**

* XGBClassifier often achieves the highest F1-Score and Precision across different splits.
* Random Forest and Decision Tree perform consistently well, often close to XGBClassifier.
* Logistic Regression shows stable performance across different splits.
* Naive Bayes consistently underperforms compared to other models, with the lowest accuracy and F1-Score.
* SVM shows moderate performance but doesn't stand out in any metric.

Precision is generally lower than Recall for most models, indicating a tendency towards false positives rather than false negatives. I think since we are dealing in medical context, Precision is more important than recall.

Now let’s look at confusion matrix of worst model and with model at 80-20 split:

A blue and white box with numbers

Description automatically generated A blue and white graph

Description automatically generated

This tells a different story. After all Naïve Bayes isn’t that bad at all. It did better than XGBoost at classifying Pre-Diabetes and Diabetes. It did worse with ‘No Diabetes’, but overall, it looked worse. If we care about Pre-Diabetes and Diabetes, we can say that Naïve Bayes did so much better than XGBoost.

Now, let’s fix the class imbalance problem and introduce feature reduction to improve Precision.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Train-Test Split** | **Model** | **Accuracy** | **F1-Score** | **Precision** | **Recall** |
| 50% train, 50% test | Logistic Regression | 0.6462 | 0.7117 | 0.8329 | 0.6462 |
|  | Decision Tree | 0.6052 | 0.6802 | 0.8345 | 0.6052 |
|  | Random Forest | 0.6137 | 0.6908 | 0.8356 | 0.6137 |
|  | Naive Bayes | 0.6309 | 0.706 | 0.8207 | 0.6309 |
|  | SVM | 0.7778 | 0.7939 | 0.8244 | 0.7778 |
|  | XGBClassifier | 0.7565 | 0.7788 | 0.8129 | 0.7565 |
| 60% train, 40% test | Logistic Regression | 0.6428 | 0.711 | 0.833 | 0.6428 |
|  | Decision Tree | 0.6068 | 0.6801 | 0.8338 | 0.6068 |
|  | Random Forest | 0.6124 | 0.692 | 0.8368 | 0.6124 |
|  | Naive Bayes | 0.627 | 0.7048 | 0.8218 | 0.627 |
|  | SVM | 0.775 | 0.7922 | 0.8255 | 0.775 |
|  | XGBClassifier | 0.7492 | 0.7754 | 0.8157 | 0.7492 |
| 70% train, 30% test | Logistic Regression | 0.6478 | 0.713 | 0.8331 | 0.6478 |
|  | Decision Tree | 0.5876 | 0.6715 | 0.8395 | 0.5876 |
|  | Random Forest | 0.6119 | 0.6925 | 0.8388 | 0.6119 |
|  | Naive Bayes | 0.6304 | 0.7062 | 0.8228 | 0.6304 |
|  | SVM | 0.7725 | 0.7909 | 0.8274 | 0.7725 |
|  | XGBClassifier | 0.744 | 0.773 | 0.8187 | 0.744 |
| 80% train, 20% test | Logistic Regression | 0.6465 | 0.7136 | 0.8344 | 0.6465 |
|  | Decision Tree | 0.6306 | 0.7031 | 0.8342 | 0.6306 |
|  | Random Forest | 0.6119 | 0.6925 | 0.8388 | 0.6119 |
|  | Naive Bayes | 0.6304 | 0.7062 | 0.8228 | 0.6304 |
|  | SVM | 0.7725 | 0.7909 | 0.8274 | 0.7725 |
|  | XGBClassifier | 0.744 | 0.773 | 0.8187 | 0.744 |

**Overall performance changes:**

* The precision scores have improved across all models, which is particularly important for medical applications.
* But the accuracy, f1, and recall have decreased for most models compared to the previous results.

**Model comparisons:**

* XGBClassifier and SVM are the best model in terms of accuracy, F1-Score, and recall across all splits.
* Logistic Regression, Decision Tree, Random Forest, and Naive Bayes now show lower performance compared to SVM and XGBClassifier.

**Precision focus:**

* Given the importance of precision in medical contexts, it's noteworthy that all models now show high precision scores (above 0.81).
* Random Forest consistently achieves the highest precision across all splits, followed closely by Decision Tree.

Now let’s look at confusion matrix of worst model and with model at 80-20 split, with handling class imbalance and feature selection:

A blue and white chart with numbers

Description automatically generatedA blue squares with black numbers

Description automatically generated

Prediction of Pre-Diabetes and Diabetes improved a lot in both the models. Naïve Bayes still does a better job with predicting these two, even though overall accuracy is lowest.

These results demonstrate the impact of addressing class imbalance and feature reduction. While they've improved precision, which is crucial in medical applications, they've also led to some decrease in recall. It also demonstrated that different models perform better with different aspects of the data. While initially we ruled out Naïve Bayes as the worst performer, considering the goal of predicting the people who are already with Diabetes and about to get Diabetes, it did best.