LifeFactors: Unveiling Disease Probability with Machine Learning

Senior Project I

Co-Authored By

Anurag Karki
Bachelor of Information Technology, Assumption University, 6411318

Sanjana Subrahmanya
Bachelor of Information Technology, Assumption University, 6411242

Manal Mahmood
Bachelor of Computer Science, Assumption University, 6340044

Under the Advising of

Dr. Anilkumar Kothalil Gopalakrishnan
Lecturer, Department of Computer Science, Assumption University

Acknowledgement

The participants of this project would like to thank Assumption University of Thailand for the completion of this project, for this learning opportunity, and valuable life lessons including soft skills as well as technical skills.

Secondly, we would like to convey our deepest and most sincere gratitude towards Dr. Anilkumar Kothalil Gopalakrishnan, for providing unwavering guidance, support and advice as well as suggestions on our project. It has been an honor and a privilege to work as well as study under his guidance.

Table of Content

Abstract

Health and IT have been in collaboration with each other since the very start of the Computer Science field. One of the highest uses of IT has been seen in the use of Machine Learning. This paper emphasizes the use of IT, especially machine learning and the algorithms in implementing a health focused web application. The use of machine learning here to predict a highly prevalent disease, diabetes, and a highly specific disease, heart disease (myocardial infarction or coronary heart disease). This paper delves into the code used in the machine learning application and shows each of their effectiveness and performance in predicting the respective diseases. The paper aims to aid in medical research.

I. Introduction

Diabetes is one of the most prevalent diseases in the continent of North America, and impacts millions. It is also very prevalent in other continents, especially where the knowledge and implementation of a good diet or a good lifestyle is not enforced or taken care of. The dataset used for Diabetes comes from the BRFSS (Behavioral Risk Factor Surveillance System), which is a telephone survey collected annually by the CDC, Centre for Disease Control in the USA (CDC,2015). On the other hand myocardial infarction has been a leading cause of death in the world due to its nature of causing inflammation in the vascular walls. This, unlike diabetes, may be an acute disease or a lifelong disease and it is the first symptom of coronary heart diseases. Myocardial infarctions are hard to detect unless very specific tests of ST changes of the heart along with ECG of resting and post stress test of the heart is done. Hence in the case of this project, there are two ways to detect heart diseases, one is through a dataset that has highly specific markers needed to detect the disease while on the other hand, the dataset from the CDC of 2022, is broken down to have general lifestyle and background markers to detect heart diseases. This is a project and it should not be used to prescribe or diagnose any user with any disease. It is simply a prediction system that aims to aid in the field of Health and IT.

II. Datasets

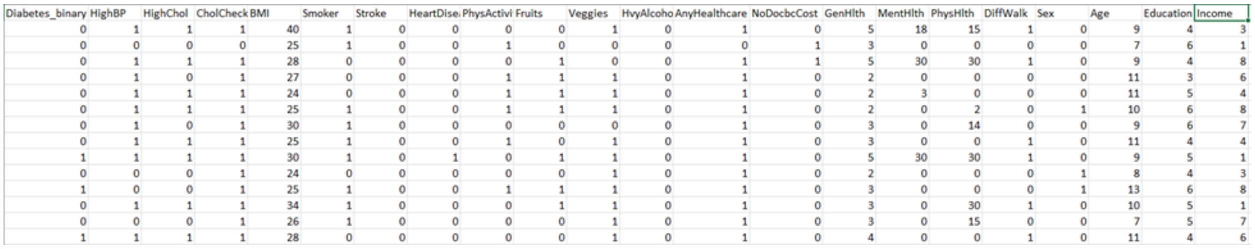
The survey collects more than 400,000 replies, and this survey is from 2015. There are 2 datasets used for diabetes. First is the dataset used for testing and training which includes 254,680 survey responses, with the target variable having 2 classes; 1 for Prediabetic or Diabetic and 0 for no Diabetes. The dataset contains 21 features. The second dataset used is unseen data from the total 400,000 responses, where it contains 70,692 responses and has an equal number of diabetic and nondiabetic responses.

The columns include:

1. General health (A rating of 1 to 5, 5 being the highest on the responders general health)

2. High Blood Pressure (Yes/No for having High BP)

3. BMI (Continuous float value)
4. Age (Continuous integer value, grouped into 13 bins)
5. High Cholesterol (Yes/No for having high Cholesterol)
6. Difficulty Walking (Yes/No for having Difficulty Walking)
7. Income (Integer grouped into 8 bins)
8. Heart Disease or Attack (Yes/No for experiencing Coronary Heart Disease or Myocardial Infarction)
9. Physical Health (Personal rating of Physical health)
10. Physical Activity (Yes/No for Physical Activity done in the past 30 days)
11. Education (6 bins for the education level one has received)
12. Stroke (Yes/No for ever having a stroke)
13. Cholesterol Check within last 5 years (Yes/No)
14. Heavy Alcohol Consumption (Yes/No, Heavy Alcohol Consumption is more than 14 Standard drinks a week for male and more than 7 Standard drinks for female)
15. Mental Health (Integer value for the number of days where responder has had poor mental health, within the past 30 days)
16. Fruit Consumption (Yes/No, consumption of fruits at least once a day)
17. Vegetable Consumption (Yes/No, consumption of vegetables at least once a day)
18. Smoker (Yes/No, Consuming of more than 100 cigarettes in lifetime; 5 packets of standard cigarettes)
19. Sex (Gender of the responder)
20. Any Health Care (Yes/No having a health care plan)

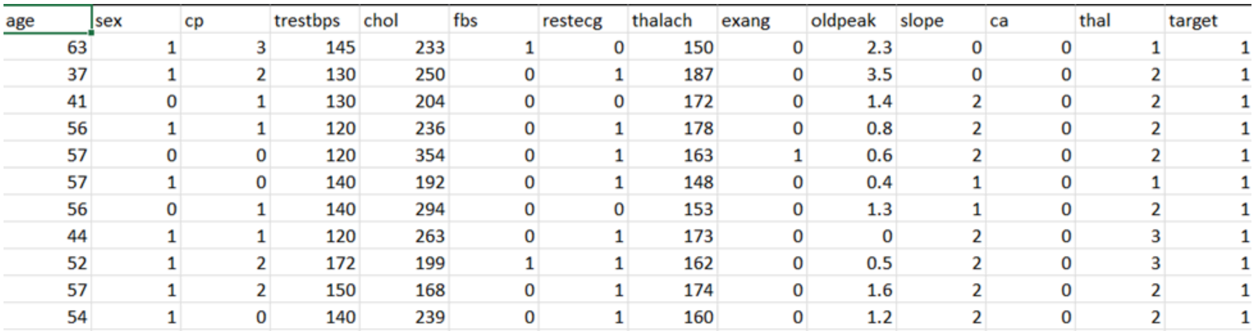


Diabetes_binary	HighBP	HighChol	CholCheck	BMI	Smoker	Stroke	HeartDise	PhysActivi	Fruits	Veggies	HvyAlcoho	AnyHealthcare	NoDocbcCost	GenHlth	MentHlth	PhysHlth	DiffWalk	Sex	Age	Education	Income
0	1	1	1	40	1	0	0	0	0	1	0	1	0	5	18	15	1	0	9	4	3
0	0	0	0	25	1	0	0	1	0	0	0	0	1	3	0	0	0	0	7	6	1
0	1	1	1	28	0	0	0	0	1	0	0	1	1	5	30	30	1	0	9	4	8
0	1	0	1	27	0	0	0	1	1	1	0	1	0	2	0	0	0	0	11	3	6
0	1	1	1	24	0	0	0	1	1	1	0	1	0	2	3	0	0	0	11	5	4
0	1	1	1	25	1	0	0	1	1	1	0	1	0	2	0	2	0	1	10	6	8
0	1	0	1	30	1	0	0	0	0	0	0	1	0	3	0	14	0	0	9	6	7
0	1	1	1	25	1	0	0	1	0	1	0	1	0	3	0	0	1	0	11	4	4
1	1	1	1	30	1	0	1	0	1	1	0	1	0	5	30	30	1	0	9	5	1
0	0	0	1	24	0	0	0	0	0	1	0	1	0	2	0	0	0	1	8	4	3
1	0	0	1	25	1	0	0	1	1	1	0	1	0	3	0	0	0	1	13	6	8
0	1	1	1	34	1	0	0	0	1	1	0	1	0	3	0	30	1	0	10	5	1
0	0	0	1	26	1	0	0	0	0	1	0	1	0	3	0	15	0	0	7	5	7
1	1	1	1	28	0	0	0	0	0	1	0	1	0	4	0	0	1	0	11	4	6

Figure 1: BRFSS dataset, CDC, 2015

Unlike diabetes, heart disease can be caused by underlying factors that may or may not always be caused by lifestyle factors. While there was data available for heart disease based on one's lifestyle factors, these datasets were not reliable and robust, hence the dataset chosen here is highly complex and one would need to get a bloodwork done with the doctor to input their statistics. The heart dataset contains 15 columns which include:

1. Age (age of the responder)
2. Sex (Gender)
3. Chest Pain Type (0: Typical angina | 1: Atypical angina | 2: Non-anginal pain | 3: Asymptomatic)
4. Resting Blood Pressure in mm/Hg
5. Serum Cholesterol (mg/dl)
6. Fasting Blood Sugar level (1 for higher than 120 mg/dl, else 0)
7. Resting Electrocardiographic Results (0: Normal | 1: Having ST-T wave abnormality |
8. 2: Showing probable or definite left ventricular hypertrophy
9. Maximum Heart Rate achieved during a stress test
10. Exercise-induced Angina (Yes/No)
11. ST depression induced by exercise relative to rest
12. Slope of peak exercise ST segment (0: Upsloping | 1: Flat | 2:Downsloping)
13. Major Vessels (0-4) colored by fluoroscopy
14. Thallium Stress Test Result (0: Normal | 1: Fixed Defect | 2: Reversible Defect | 3: Not described)
15. Heart disease Status (0 = no disease | 1 = presence of disease)



age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
57	0	0	120	354	0	1	163	1	0.6	2	0	2	1
57	1	0	140	192	0	1	148	0	0.4	1	0	1	1
56	0	1	140	294	0	0	153	0	1.3	1	0	2	1
44	1	1	120	263	0	1	173	0	0	2	0	3	1
52	1	2	172	199	1	1	162	0	0.5	2	0	3	1
57	1	2	150	168	0	1	174	0	1.6	2	0	2	1
54	1	0	140	239	0	1	160	0	1.2	2	0	2	1

Figure 2: Heart Blood work specific input dataset, 2018

The heart disease prediction is done again for myocardial infarction and coronary heart disease, but the dataset chosen in this case does not delve into the heart-specific blood work but rather behavioral risk factors as defined by the CDC in their 2020 Behavioral Risk Factor Surveillance survey. The dataset has around 350,000 rows, from which a random sample of 20,000 rows are chosen with those having no heart disease, and 20,000 rows for those responses having heart disease. These are concatenated to make up the training and testing dataset. A total of 40,000 rows are chosen because having the algorithms train on a dataset of a high magnitude is computationally expensive and is not compatible with the machine and equipment used for this research. Out of the 40,000 balanced rows. The dataset in this case consists of 16 independent variables and 1 dependent variable. The rows include:

1. BMI (float value)
2. Sex
3. Diabetic (Whether the person was diabetic or prediabetic: 1, or non-diabetic: 0
4. KidneyDisease (Whether the responder ever has had kidney disease 0 for No, 1 for yes)
5. SkinCancer (Whether the responder ever has had Skin Cancer 0 for No, 1 for yes)
6. KidneyDisease (Whether the responder ever has had Asthma 0 for No, 1 for yes)
7. Smoking (Whether the responder has smoked more than 100 singular cigarettes in their lifetime, 0 for No, 1 for Yes)
8. Stroke (Yes/No for ever having a stroke)
9. Heavy Alcohol Consumption (Yes/No, Heavy Alcohol Consumption is more than 14 Standard drinks a week for male and more than 7 Standard drinks for female)
10. Mental Health (Integer value for the number of days where responder has had poor mental health, within the past 30 days)
11. Sleep Time (How many hours in a day the responder sleeps | Integer)
12. GenHealth (A rating of 1 to 5, 5 being the highest on the responders own general health)
13. Race (The responders race background, from a choice of

    'White'

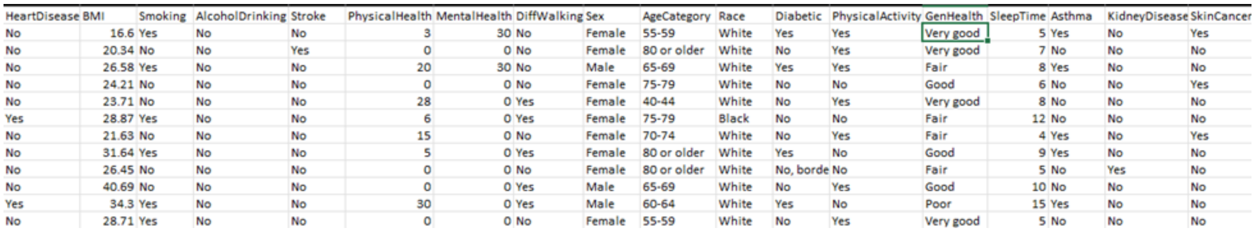
    'Black'

    'Hispanic,

    'Asian'

    'American Indian/Alaskan Native', 'Other')

14. DiffWalk (Whether responder has serious difficulty walking, 0 for NO, 1 for YES)
15. Age Category (The responders age, grouped into 13 bins, 5 years apart)
16. PhysicalHealth (The number of days the responder has fallen sick in the last 30 days)
17. HeartDisease (Whether the responder has a heart disease or not, 0 for NO, 1 for YES)



HeartDisease	BMI	Smoking	AlcoholDrinking	Stroke	PhysicalHealth	MentalHealth	DiffWalking	Sex	AgeCategory	Race	Diabetic	PhysicalActivity	GenHealth	SleepTime	Asthma	KidneyDisease	SkinCancer
No	16.6	Yes	No	No	3	30	No	Female	55-59	White	Yes	Yes	Very good	5	Yes	No	Yes
No	20.34	No	No	Yes	0	0	No	Female	80 or older	White	No	Yes	Very good	7	No	No	No
No	26.58	Yes	No	No	20	30	No	Male	65-69	White	Yes	Yes	Fair	8	Yes	No	No
No	24.21	No	No	No	0	0	No	Female	75-79	White	No	No	Good	6	No	No	Yes
No	23.71	No	No	No	28	0	Yes	Female	40-44	White	No	Yes	Very good	8	No	No	No
Yes	28.87	Yes	No	No	6	0	Yes	Female	75-79	Black	No	No	Fair	12	No	No	No
No	21.63	No	No	No	15	0	No	Female	70-74	White	No	Yes	Fair	4	Yes	No	Yes
No	31.64	Yes	No	No	5	0	Yes	Female	80 or older	White	Yes	No	Good	9	Yes	No	No
No	26.45	No	No	No	0	0	No	Female	80 or older	White	No, borde	No	Fair	5	No	Yes	No
No	40.69	No	No	No	0	0	Yes	Male	65-69	White	No	Yes	Good	10	No	No	No
Yes	34.3	Yes	No	No	30	0	Yes	Male	60-64	White	Yes	No	Poor	15	Yes	No	No
No	28.71	Yes	No	No	0	0	No	Female	55-59	White	No	Yes	Very good	5	No	No	No

Figure 3: (BRFSS Dataset, CDC, 2020)

III. Machine Learning algorithms

Machine Learning or ML as used in this documentation is an approach to "teach' machines and computers to handle data in an efficient way to achieve an output, in a more proper and efficient manner than a human could do on their own. The main purpose of ML algorithms was initially to extract information from data that cannot be easily extracted by humans. There are two types of ML algorithms:

> Supervised Learning, which has in its grasp a set of inputs and a label or labels that are outputs and maps these inputs to the outputs based on example pairs
>
> And Unsupervised ML algorithms, which do not have any target outputs, and instead are geared towards finding trends in the data or patterns that may exist.
>
> In this documentation and this project, most of the algorithms used and those that are most efficient are supervised learning algorithms, and these are the ones to be focused on.

Naïve Bayes

This supervised form of machine learning is based on the Bayes Probability theorem, which is based on one primary assumption: that the presence of every feature is independent of those of all the other features. Using this following formula, the algorithm calculates the mean and the standard deviation of each of the predictor variables in each class, then calculates the probability of F(i) using the gauss Density equation :

Where, in training dataset T,

Naive Bayes Formula:

$$P(c|x) = \frac{P(x\,|c)\,P(c)}{P(x)}$$

$P(c|X) = P(x_1|c) \times P(x_2|c) \times ... \times P(x_n|c) \times P(c)$

- $P(c|x)$ → Posterior Probability
- $P(x|c)$ → Likelihood
- $P(c)$ → Class Prior Probability
- $P(x)$ → Predictor Prior Probability

$F = f_1, f_2, f_3, f_4 \dots f_n$, where f_n refers to the value of predictor variables in the testing dataset

The general probability density function:

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

Support Vector Machine

SVM has also been used in this project, and is widely used in the ML field as well. The algorithm works well with the datasets in the given project, as in this case, the classification is linear as most of the columns that are used in the algorithm show a linear co-relationship with the target, whether positive in the case of BMI, smoker, Education, or negative in case of Veggies, PhysHlth (physical health rating), MentHlth (Mental Health rating), GenHlth (General Health rating), and research indicates that SVMs work efficiently on both linear and non-linear relationships (Mahesh, 2019).

Gradient Boosting

The Gradient Boosting Machine or the Gradient Boosting Algorithm has been known as one of the most effective machine learning algorithms as the principle revolves around minimizing the bias error of a model. Unlike other techniques used in this project the GBM uses a machine learning ensemble technique which is the use of multiple learning algorithms (usually decision trees) in a sequential manner. The GBM then improves the performance of the model by continually updating the weights based on previous errors, similar to the Back-propagation method (Mahesh, 2019).

K Nearest Neighbors

The algorithm in the case for K nEarest Neighbors is different from the other algorithms used in this project as it uses the data directly for classification and does not build any model initially. The K in the name is the number of parameters to be calculated for the algorithm. Because there is no need for any model building the algorithm is pretty fast, and it will group the data based on the number of N_neighbors chosen. In the case of this project, since the dataset is binary classification, the number of neighbors chosen is 3, as opposed to two to avoid the voting tie. The major drawback for KNN is the fact that it uses distance metric from the neighbor mean as the main criteria and this makes the relative importance of data to be lost (Dreiseitl & Ohno-Machado, 2002).

Logistic Regression

The logistic regression model is a highly useful method as not only can it classify linear data, but also does well with the issue of linear separability if the dataset does not tend to be a linear dataset when compared with the target variable. The $P(y|x)$ or the prediction given the variable, is provided as a function $f(x, \alpha)$ and the alpha is determined based on the dataset, using the maximum likelihood estimation. The basic premise of the algorithm is that it predicts the binary outcome probability such as if an email is spam or not spam, or a yes.

The logistic regression formula is :

$$y = \frac{e^{b_0 + b_1 \cdot x}}{1 + e^{b_0 + b_1 \cdot x}}$$

Where:

x is the input value

y is the predicted output

b_0 is the bias or intercept term

b_1 is the coefficient for the single input value (x)

(Dreiseitl & Ohno-Machado, 2002)

Random Forest Classifier and Decision trees

Decision Trees are the foundations of Random Forest Classifier. The Decision tree is a supervised machine learning model that is made up of root nodes, internal nodes and leaf nodes, which is where the decision tree ends, where each node represents a certain feature and the branches represent the values. The root node and the split of the nodes are determined by the feature that best divides the data according to their purity score. The process is applied recursively to each feature, and the probability of each node is calculated till it reaches the leaf nodes. The root node represents the entire dataset, with the successive nodes, being known as the decision nodes due to the most pure split, will branch out and divide the data into smaller subsets, creating further branches and each branch is a possible route through the tree. This process is repeated till it reaches the leaf nodes. Decision Trees consist of a few assumptions such as the binary split which suggests that the node divides the data into two subsets, based on a single condition, making it seem as if each decision is a binary choice.

The random Forest has the advantage over decision trees of overfitting and are less sensitive to outlier data and variable or feature importance is generated automatically. It consists of selecting random data from the entire dataset and creating decision trees. In a Random Forest, the features are randomly selected in each decision split. Instead of selecting the most important feature while splitting a node, it searches for the best feature among a random subset of features, allowing for better results.

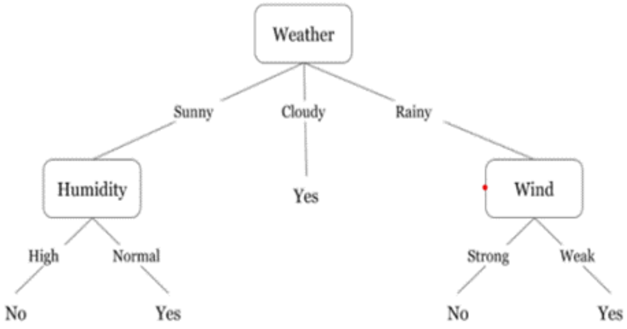


Figure 4: Decision Tree in a Random Forest (KUMAR, 2020)

Random forest algorithm uses a bagging technique which means that it creates a training subset from random data, and the final result depends upon the majority voting, for example if two models or decision trees give out the output as “1” and 5 give output as “0” or “No”, the majority voting will make it so that the final result is also “No”. The hyperparameters used includes n_estimators, which is the number of trees the algorithm builds, max features which is the number of features that the model considers on splitting and criterion which is the method in which the node is split. The criteria used is Entropy, which means the one with the lowest entropy is chosen (Ali et al., 2012).

IV. Code Breakdown:

Both of the datasets being used are binary classifications. First the data is explored to check the frequency distribution of the features against the target variable in both cases.

Diabetes Dataset:

Since the diabetes dataset has 22 columns in total with more than 200,000 rows, this would be computationally expensive in the case of this project and hence a sample of 50,000 is taken from the seen and unseen datasets. The sample is randomly taken from the dataset with the pd.sample(50,000) function. Furthermore, a block of code is run to know the weight of evidence of each of the features and its information value. In this block of code, known as the function iv_woe, it takes in the data, the target, bins and show_woe (Boolean = False) as parameters. The target is the y column, the bins is the number of bins for the continuous variables and the show_woe is a flag used to determine whether to display the Weight of Evidence table.

A new dataset newDF and woeDF is initialized and will store the information value and WOE values. The code then loops through the independent variable to check if they are of numeric type and if the unique values of these numeric types are more than 10. If both conditions are fulfilled, it will bin the value using quantiles (pd.qcut). If one of the conditions or both are not met, it will use the raw values themselves.

After converting the x column into dataframe d0 of string type, it groups the dataframe d0 by counting the number of occurrences in the y column or target variable. The dataframe d will include Events, Non-events and Cutoff. Events are where diabetes is present, Nonevents are where it is not present or ‘0’. Cutoff refers to the edges of the bins when binning continuous variables like age, BMI, etc. Then the % of events and nonevents is calculated along with their proportion. Finally the Weight of Evidence and the Information value is calculated for each group using the respective formulae.

Weight of evidence has the formula:

This is a measure between categorical independent variable and the target variable and shows how well a particular group differentiates between the event and the nonevent. Positive WoE shows that the group is more likely to have nonevents and a negative WoE shows higher likelihood of the event. This is an inverse relationship

Weight of Evidence Formula

$$WoE = ln(\frac{\% Non-Events}{\% Events})$$

- $\% Non - Events$ → The percentage of non-events in a group
- $\% Events$ → The percentage of events in a group

And Information value has the formula:

$$IV = \sum_{i=1}^{n} (\%NonEvents_i - \%Events_i) \times WoE_i$$

- n: The number of categories or bins in the variable
- $\% Non - Events$: The percentage of non-events in the i-th category
- $\% Events$: The percentage of events in the i-th category
- WoE_i: The WoE value for the i-th category

The Information value is what is needed to find the predictive power of the categorical variable by transforming the variable into a linear form and showing the influence it has on the target variable. The higher the Information value, the more predictive power it has on the prediction.

The information Value is what is needed to figure out the predictive power of the variables, and hence, it is shown here:

Columns in the dataset:

```
collist = ['HighBP', 'HighChol', 'CholCheck',
'BMI', 'Smoker', 'Stroke',
'HeartDiseaseorAttack', 'PhysActivity', 'Veggies',
'HvyAlcohol Consump', 'GenHlth',
'MentHlth','PhysHlth', 'Diffwalk',
'Age','Education', 'Income']
```

Information Value of columns:

```
              Variable        IV
               GenHlth  0.711537
                HighBP  0.561243
                   BMI  0.426943
                   Age  0.387188
              HighChol  0.338062
              Diffwalk  0.291940
                Income  0.187527
              PhysHlth  0.183964
  HeartDiseaseorAttack  0.180077
             Education  0.089460
          PhysActivity  0.085989
             CholCheck  0.069697
                Stroke  0.056556
     HvyAlcoholConsump  0.043313
              MentHlth  0.026132
                Smoker  0.020964
               Veggies  0.018431
                   Sex  0.011707
                Fruits  0.006902
           NoDocbcCost  0.005414
         AnyHealthcare  0.004927
```

The Columns then chosen are from the whole column list are up to 'Veggies' with an Information Value of 1.8%.

Now that the columns and the features have been chosen, the preprocessing requires certain columns to be standardized. The need for scaling in preprocessing has been evident in research. Firstly, when it comes to Machine Learning, having data in features that have values close together such as between -1 and 1 or 0 and 1 allows for better training, and on the other hand, data that are further apart takes longer to understand for the ML program (Sharma, 2022).

The Scaling used in the project is the Standard Scaler Library, that uses the z-score normalization where the data point x is converted into z, a standard value, using the formula given below.

$$z = \frac{x-\mu}{\sigma}$$

Where:

- z is the standardization value
- x is the original value
- μ is the mean of the feature
- σ is the standard deviation of the feature

This formula converts the data point into a value in such a way that the mean of the variable becomes 0 and the standard deviation is 1. The need for scaling is especially important for large datasets such as the one used in this project, especially in the diabetes dataset, with more than 200,000 rows and values ranging from 0 up to more than 200.

The use of scaling allows for faster convergence and a better performance of the ML algorithm which makes better predictions.

And finally, after preprocessing the machine learning algorithms are applied. The algorithms are already explained in the previous section. This section will delve into the hyperparameters and the parameters used in the algorithms.

For the diabetes dataset, the random forest classifier, which is prone to overfitting, uses a parameter search using the GridSearchCV. The max features refers to the number of features to consider for splitting a node during construction of the tree, and it will consider the values 2, 7, and 12 (2, 2+5, 2+5+5).

Secondly, the n_estimatiors is the number of trees in the forest which is 500,1000,1500, and the max depth is the maximum depth of the trees in the forest. The gridSearchCV function is used for cross validation and it finds the best combination of hyperparameters to find the best set based on the cross validated performance.

In the case of Logistic regression, a solver is used, called saga, or Stochastic Average Gradient Descent. This is used for large datasets such as the one in diabetes. Max_iter is set to 1000, to reduce the chances of overfitting. And finally, a random state with a random number, but consistent to all algorithms is used to ensure reproducibility of the results.

In the case of the Support vector Machine, a regularization parameter is used called C. This parameter is used to achieve a tradeoff between low testing error and low training error. Smaller values will lead to a large margin but allow training points to be misclassified, while larger values will allow for narrower margins but will penalize misclassifications heavily. The Kernel function is there to capture the relationship between data, where rbf is for nonlinear relationships and poly for polynomial relationships. The GridSearch as before will find the best combination of hyperparameters for the Support Vector machine.

Finally after training and testing the data, the accuracy score of each of the algorithm in training and testing, their precision score for ‘Yes’ or having diabetes and precision score for ‘No’, not having diabetes is recorded in an array with the names of each of the algorithm, and their incorrect and correct classification percentages are calculated form the confusion matrix as well. This is to find the best performing algorithm on the dataset, which later will predict on the unseen data and then on user input.

Heart Disease Prediction Code:

The heart prediction dataset is a relatively small dataset, hence each of the algorithms has parameter tuning with grid search, and this is not as computationally expensive as in the case, if used in the diabetes dataset.

In the case of each algorithm, a parameter grid is created and a grid search searches for the best combination from the given parameters in the parameter grid to find the best performing set of parameter, the hyperparameters is described:

Decision tree:

The criterion is the criteria used for splitting the nodes of the features. The entropy criteria is used which will measure the randomness or the surprise or the impurity of the feature and split according to the hierarchy of the most pure feature. The max depth is the maximum depth of the tree, and in this case a higher value would lead to better performance but is prone to overfitting. Minimum sample split is the number of samples at minimum required to split the internal node and it sets the threshold at which the node will not be split further. The minimum sample leaf parameter sets a threshold for the minimum number of samples required to create a leaf node, while small values can create granular leaves, it is prone to overfitting.

Gradient Boosting:

A parameter learning rate is used which determines the contribution of each tree to the overall model and a small learning rate makes the model robust but would need a higher n_estimator parameter to achieve higher accuracy. The n estimator parameter is the boosting stages to be fitted. While a high value is more robust it is more prone to overfitting. Max depth is the maximum depth of each tree in the algorithm; it is kept shallow at 2, to avoid the issue of overfitting.

Support vector Machine: Similar to the algorithm used in the diabetes dataset, this algorithm adds svm__gamma which controls the shape of the decision boundary. A decision boundary in SVM is determined by support vectors. In this algorithm there is a hyperplane that best separates the classes while minimizing the distance to the nearest data point of each class. This is crucial for making predictions. When scale is used, the decision boundary is less complex, and it is suitable for simple datasets, but here auto is chosen by the grid search as the data set is complex, and needs a flexible and adaptive decision boundary.

Random Forest Classifier and Logistic Regression use the same hyperparameters as used in the case of the diabetes dataset.Finally, the models are fitted and trained then tested. The model with the highest accuracy score is loaded into a file with the joblib library and is then used to predict from unseen data, which in this case is user input data.

Accuracy Tables

Accuracy table for Diabetes dataset:

Algorithm	Precision of Yes	Precision of No	Correctly Classified %	Incorrectly Classified %	Accuracy
Random Forest Classifier	0.7360	0.7654	74.48	25.52	74.48
K-Nearest Neighbors	0.7281	0.6503	69.88	30.12	69.88
Gaussian Naïve Bayes	0.7748	0.6459	71.69	28.32	71.69
Logistic Regression	0.7613	0.7473	75.64	24.36	75.64
Gradient Boosting Classifier	0.7498	0.7460	74.85	25.14	74.85
Support Vector Machine	75.04%	0.7594	75.33	24.67	75.33

Accuracy Table for Heart Dataset predicted from heart-specific bloodwork

Algorithm	Precision of Yes	Precision of No	Correctly Classified %	Incorrectly Classified %	Accuracy
Random Forest Classifier	0.8529	0.8519	85.24	14.75	85.24%

K Nearest Neighbor	0.7272	0.6786	70.49	29.51	70.49%
Logistic Regression	0.7941	0.7778	78.69	21.31	78.69%
Gradient Boosting Classifier	0.8378	0.9167	86.89	13.11	86.89%
Gaussian Naïve Bayes	0.8235	0.8148	81.97	18.03	81.97%
Support Vector Machine	0.8611	0.9200	88.52	11.48	88.52%

Accuracy Table for Heart dataset from BRFSS 2020

Algorithm	Precision of Yes	Precision of No	Correctly Classified %	Incorrectly Classified %	Accuracy
Random Forest Classifier	0.7897	0.7759	78.55	21.45	78.55%
K-Nearest Neighbors	0.7230	0.6553	69.58	30.42	69.58%
Gaussian Naïve Bayes	0.7734	0.6622	72.20	30.55	72.20%
Logistic Regression	0.8073	0.7471	78.64	21.35	78.64%
Gradient Boosting Classifier	0.7921	0.7606	78.20	21.80	78.20%

Support Vector Machine	0.7905	0.7696	78.40	21.60	78.40%

V. Discussion

In both the cases for BRFSS datasets and the case for the bloodwork based dataset, the Random Forest Classifier, Logistic Regression, Gradient Boosting Classifier and Support Vector Machine and Gaussian Naïve Bayes perform very well. Although this cannot be a definitive conclusion due to the full dataset not being used as the computational budget was limited, it was found that the accuracy of the algorithms stays in a range (-3,+3) from the values given above, when the random state is changed. This can be attributed to the algorithms characteristic of being able to handle complex relationships rather than just linear ones. Here the algorithms that are “winners" are those that do not have the issue of linear separability. It cannot be definitively concluded that one algorithm would be the best suited for machine learning application in the health sector, as this depends upon the data set as well. These results are consistent with evidence in machine learning research papers as well, where datasets were fed into supervised machine learning algorithms and the results shown were better performance from the mentioned machine learning algorithms (Battineni et al., 2020). Research also suggests that KNN suffers from high dimensionality curse, where, as the number of dimensions or the number of features increases, the performance of KNN suffers due to the increase in sparsity of the data as input spaces increase with respect to the number of features (Bellman, 1957)

VI. Literature Review

Research literature is consistent with the project scope as research suggests that prediction models are best utilized and show the most promise during the early intervention stages of a disease rather than in the latter stages when the disease has progressed towards a more chronic stage (Wiens & Shenoy, 2017).

Similar projects of machine learning applications on healthcare, specifically diabetes and heart condition (myocardial infarction), show that supervised learning models such as the Random Forest model, Logistic regression and Support Vector Machine model are the most effective in terms of their accuracy. However, this is the case, given that the data is reliable. Reliability of data in healthcare in such projects refers to the population size, the way the data is balanced (the proportion of the population that have the target variable positive and the proportion of those with negative target variable).

Peer reviewed journal on the prediction of diabetes and cardiovascular diseases followed a standard for the preprocessing portion which is also implemented in the LifeFactors project. This consists of encoding the categorical variables, encoding binary classification to the target variable (where 0 denotes non cases and 1 denotes cases), followed by normalization of the dataset. Models used in such peer reviewed journals were also consistent consisting of mostly supervised learning models and one or at most two boosting or ensemble methods (Dinh et al., 2019) Diabetes dataset prediction models such as Logistic regression, random forest, linear Support Vector Machine model, Rbf Support Vector Machine Model and Naïve Bayes model performing better than other models (Mohamed et al., 2024). When the Behavioral Risk Factor Surveillance Survey dataset (2014) was used, it was seen that with different parameters or features used and no feature extraction conducted, Random Forest model achieved a 86.60% accuracy. However, these peer reviewed studies used the entire dataset with 253,680 entries from the survey (Ullah et al., 2022).

Below is a summarized table of the literature review showing accuracy of the models chosen in this project, being used on the BRFSS datasets by peer reviewed journals in the past:

This table shows accuracy of the models on predicting diabetes from BRFSS datasets

Classifier	**(Mohamed et al., 2024)**	**(Dinh et al., 2019)**	**(Xie et al., 2019)**
Random Forest	79.27%	83.6%	79.27%
Logistic Regression	73%	82.7%	80.86%
Boosting Model	86%	83.8%	-
KNN	-	-	-
Naïve Bayes	-	-	77.56%
Support Vector Machine	80.82%	82.5%	81.78%

This table shows accuracy of models predicting heart complication, namely myocardial infarction using the models:

Classifier	**(Mohapatra et al., 2023)**	**(Boudali et al., 2024)**	**(Güler et al., 2023)**
Random Forest	77.9%	74%	99%
Logistic Regression	79.3%	-	73%
Boosting Model (XGBoost / GradientBoost)	77.5%	75%	91%
KNN	78.3%	-	85.37%
Naïve Bayes	68%	-	64.13%
Support Vector Machine	73.4%	-	-

In both of these tables, the studies that are quoted are using BRFSS dataset, however, unlike the LifeFactor Project, these studies use the entire dataset with all records, ranging from 200,000 rows to 253,000 depending on the study (after preprocessing; removing null values, normalizing and feature selection).

The current project of LifeFactors uses 42000 rows from the BRFSS dataset, by using a random sample method using python code.

VII. Industry Opinions

We consulted a few doctors throughout the process of creating this project. Doctors we believed would give us constructive comments that could help offer insights into our project's development; aspects of the project that only experts could tell were in need of adjustment.These collaborative efforts have influenced our approach, and these are two very concise emails of feedback we received right after our Senior Project I exhibition. Firstly Dr. Saju Pradhan, who has been Medical Director of Nepal Orthopedic Hospital since 2008.

The main takeaway from his notes was that our project would highly benefit from working alongside blood tests. This is something that we also previously discussed as our project is only here to aid medical testing not work as a substitute.

Dr. Saju Pradhan

He states:
"HBA1C a blood test is in fact a not so expensive test to predict prediabetes. A value of between 5.7-6.4% is a strong predictor of prediabetes in itself. Likewise a triglyceride/LDL ratio greater than 3 is a strong predictor of coronary heart disease and stroke.

Also you could add that patients be advised to take a test for coronary artery calcium (CAC). This is a CT scan of the heart that tries to compute the amount of calcium deposited in the vessels. This is a highly sensitive test to predict a likelihood of having coronary artery disease or myocardial infarction. A calcium score of 0 means no evidence of heart disease, mildly increased risk 1-99, moderately increased risk 100-299 and above that a high risk. You can look up literature to verify.

Overall I think your machine learning predictor is easy to fill and it could be valuable in predicting impending diseases.
I think you should also interact with other professionals for better feedback. All the best."

HBA1C testing was already something we believed would help significantly with our Heart-Specific Blood Work algorithm, so we were glad that this notion was backed by Dr. Pradhan.

Dr. Sanjaya Khanal MD, FACC

We were next able to get a hold of Dr. Sanjaya Khanal who works out of Lancaster, California. He broke down his advice into two main parts. Firstly:

"1. Please define the diseases (therefore the output) more precisely. Diabetes mellitus has a very precise definition- finding of Hemoglobin A1C of 6.5% or more in the blood or fasting/post prandial blood sugars of 126/200 mg/dl. Similarly myocardial infarction has a precise definition but coronary heart disease can range from minimal atherosclerosis to obstructive coronary disease. Also please elaborate about who the target users are for your model (clinicians, patients, payors, researches etc)"

This first part of his advice we had already had the answers to. Our target audience and diabetes-prediabetes query is something we've also clarified, unfortunately we weren't able to provide a full report at the time, just portions of it.
Note: Being more specific about who our target users are is something we will cover shortly in this report (Part VIII)

“2. Please clarify if you are trying to predict whether a person has the disease currently with the set of input parameters or trying to predict whether they will develop the disease in the future (Classifier vs. Predictive tool). The first is not very helpful for the clinician because there are precise tests to detect the diseases without the input parameters. The second might be helpful for clinicians, patients, and payors to assess overall risk of developing disease in a certain timeframe. If this is what you are planning to do, you need to define a certain timeframe (e.g. 10 years) for the prediction. Also elaborate if there are other tools that are doing this. 3. One general advice in choosing an Al project would be to think of the output action pairing. Start with the output and assess if it is unique/accurate, if the data is reasonable, whether the output is actionable (practical, useful, reproducible, financially viable, and scalable). There are a lot of Al solutions that are waiting to be discovered in health and medicine.
Good luck!”

A really great point made had to do with setting a timeline.
“(Classifier vs. Predictive tool)”
Our project was mainly to assess overall risk of developing disease, but we hadn't set an exact timeline. More so we believed the risk of development was something that would change as a person's individual features changed such as weight or other lifestyle factors; so setting a timeframe for the prediction would definitely depend on person to person and the resulting probability they receive after using our algorithm.

The first response we ever got was a medical response from Dr Amod Rayamajhi, MBBS, Nepal Medical Association. This was during the preliminary days of the project creation when we more so just had the concept, goals and began incorporating the algorithms. His advice also mentioned *“Diabetes Mellitus, Angina Pectoris, Myocardial Infarctions”*
He stated that he believed a project like this could help the everyday citizen “mitigate their individual risk factors” which became the core foundation for this project

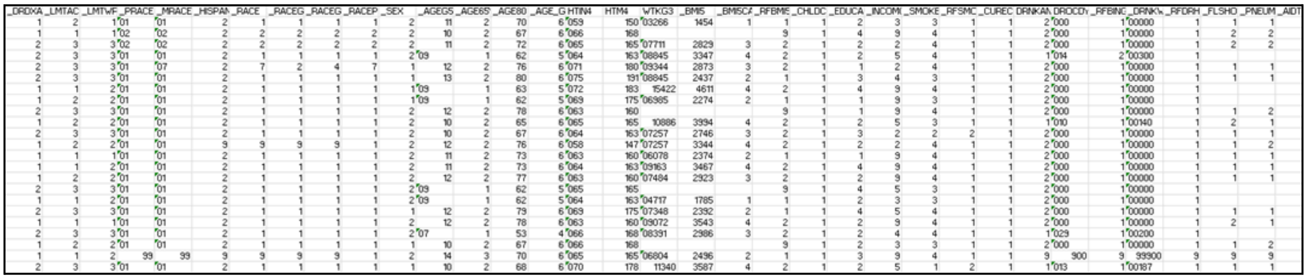
VIII. Pre-diabeties & More Recent Data Sets

With much consideration after the user exhibition and taking in the complaints from users in the project exhibition, it was clear that there were concerns on the medical data for diabetes being from 2015, and not recent years. While the treatments for diseases change over time, and new innovations allow the medical field to better treat patients, the underlying disease and its research are considered quite robust in the medical field, and seldom get outdated.

However, due to concerns from users, the project also incorporates datasets from 2021 and 2022, from the CDC, a highly reputed government agency, for a more recent data based diabetes prediction. This prediction also addresses the second concern of the people or users of the exhibition, which was the uncertainty of their bloodwork. Hence, this new prediction model takes in inputs that are instantly available to users as it is completely based on day-to-day routines. However, due to the very nature of medicine and biology, it is not possible to predict diabetes just from factors of day-to-day lives, hence, this model predicts whether the user may be pre-diabetic or borderline diabetic based on factors that are correlated with pre-diabetic patients.

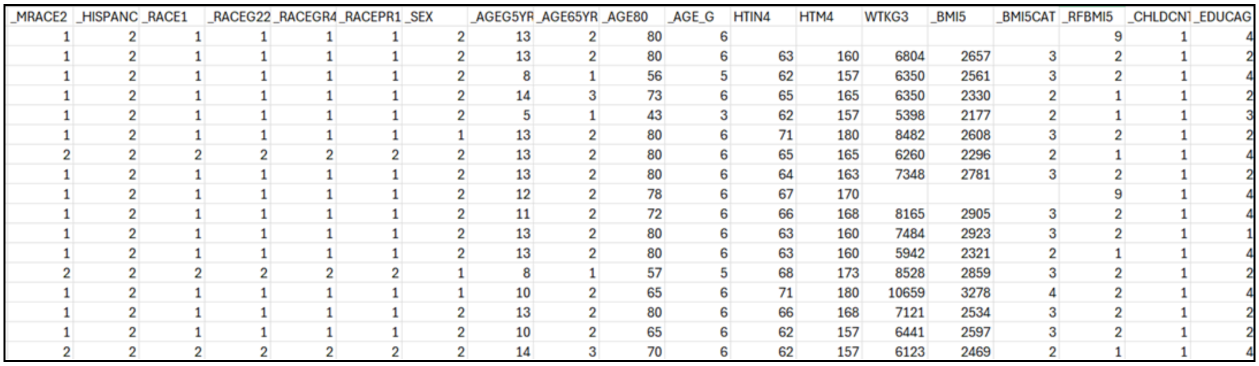
1. 'Stroke': Whether the patient has had a stroke;
2. 'DepressionEpisodes': Whether the patient used to or is currently experiencing depressive episodes,
3. 'DiffWalk': Whether the patient has difficulty walking,
4. 'Race': The patients racial background,
5. 'Sex': Gender of the patient at birth,
6. 'AgeCategory': The patient's age,
7. 'BMI': BMI of the patient,
8. 'SmokeStatus': The patients cigarette smoking habits,
9. 'VapeStatus': The patient's e-cigarette smoking habits,
10. 'BingeDrinker': The patient's binge drinking habit,
11. 'DrinkPerWeek': Number of alcoholic beverages the patient drinks per week,
12. 'HealthRating': The patient's rating for their own General health,
13. 'PhysicalHealth': Number of days the patient got sick in the last month,
14. 'MentalHealth': Number of days the patient got mental health problems in the last month,
15. 'PhysAct': Whether the patient exercised for more than majority of the month,
16. 'Pre_Diabetic': Whether the patient is considered prediabetic or borderline diabetic.

The datasets are raw datasets from the CDC according to their Behavioral Risk Factor Surveillance Survey of 2021 and 2022, respectively. The columns mentioned above each have a codeword column name and each of the values also have their corresponding value labels, which is extracted from the CDC codebook, and the two datasets are concatenated to make one dataset that includes both 2021 and 2022 datasets.



_DRDXA	_LMTAC	_LMTWF	_PRACE	_MRACE	_HISPAN	_RACE	_RACEG	_RACEG	_RACEP	_SEX	_AGEG5	_AGE65	_AGE80	_AGE_G	HTIN4	HTM4	WTKG3	_BMI5	_BMI5CA	_RFBME	_CHLDC	_EDUCA	_INCOM	_SMOKE	_RFSMC	_CUREC	DRNKAN	DROCDY	_RFBING	_DRNKW	_RFDRH	_FLSHO	_PNEUM	_AIDT
1	2	1	01	01	2	1	1	1	1	2	11	2	70	6	059	150	03266	1454	1	1	1	2	3	3	1	1	2	000	1	00000	1	1	1	
1	1	1	02	02	2	2	2	2	2	2	10	2	67	6	066	168				9	1	4	9	4	1	1	2	000	1	00000	1	2	2	
2	3	3	02	02	2	2	2	2	2	2	11	2	72	6	065	165	07711	2829	3	2	1	2	2	4	1	1	2	000	1	00000	1	2	2	
2	3	3	01	01	2	1	1	1	1	2	09	1	62	5	064	163	08845	3347	4	2	1	2	5	4	1	1	1	014	2	00300	1			
2	3	3	01	07	2	7	2	4	7	1	12	2	76	6	071	180	09344	2873	3	2	1	1	2	4	1	1	2	000	1	00000	1	1	1	
2	3	3	01	01	2	1	1	1	1	1	13	2	80	6	075	191	08845	2437	2	1	1	3	4	3	1	1	2	000	1	00000	1	1	1	
1	1	2	01	01	2	1	1	1	1	1	09	1	63	5	072	183	15422	4611	4	2	1	4	9	4	1	1	2	000	1	00000	1			
1	2	2	01	01	2	1	1	1	1	1	09	1	62	5	069	175	06985	2274	2	1	1	1	9	3	1	1	2	000	1	00000	1			
2	3	3	01	01	2	1	1	1	1	2	12	2	78	6	063	160				9	1	1	9	4	1	1	2	000	1	00000	1	1	2	
1	2	2	01	01	2	1	1	1	1	2	10	2	65	6	065	165	10886	3994	4	2	1	2	5	3	1	1	1	010	1	00140	1	2	1	
2	3	3	01	01	2	1	1	1	1	2	10	2	67	6	064	163	07257	2746	3	2	1	3	2	2	2	1	2	000	1	00000	1	1	1	
1	2	2	01	01	9	9	9	9	1	2	12	2	76	6	058	147	07257	3344	4	2	1	2	2	4	1	1	2	000	1	00000	1	1	2	
1	1	1	01	01	2	1	1	1	1	2	11	2	73	6	063	160	06078	2374	2	1	1	1	9	4	1	1	2	000	1	00000	1	1	1	
1	1	2	01	01	2	1	1	1	1	2	11	2	73	6	064	163	09163	3467	4	2	1	4	9	4	1	1	2	000	1	00000	1	1	1	
1	2	2	01	01	2	1	1	1	1	2	12	2	77	6	063	160	07484	2923	3	2	1	2	9	4	1	1	2	000	1	00000	1	1	1	
2	3	3	01	01	2	1	1	1	1	2	09	1	62	5	065	165				9	1	4	5	3	1	1	2	000	1	00000	1			
1	1	2	01	01	2	1	1	1	1	2	09	1	62	5	064	163	04717	1785	1	1	1	2	3	3	1	1	2	000	1	00000	1			
2	3	3	01	01	2	1	1	1	1	1	12	2	79	6	069	175	07348	2392	2	1	1	4	5	4	1	1	2	000	1	00000	1	1	1	
1	1	1	01	01	2	1	1	1	1	2	12	2	78	6	063	160	09072	3543	4	2	1	2	9	4	1	1	2	000	1	00000	1	2	1	
2	3	3	01	01	2	1	1	1	1	2	07	1	53	4	066	168	08391	2986	3	2	1	2	4	4	1	1	1	029	1	00200	1			
1	2	2	01	01	2	1	1	1	1	1	10	2	67	6	066	168				9	1	2	3	3	1	1	2	000	1	00000	1	1	2	
1	1	2	99	99	9	9	9	9	1	2	14	3	70	6	065	165	06804	2496	2	1	1	3	9	4	1	1	9	900	9	99900	9	9	9	
2	3	3	01	01	2	1	1	1	1	1	10	2	68	6	070	178	11340	3587	4	2	1	2	5	1	2	1	1	013	1	00187	1	1	1	

(BRFSS, CDC 2021)



_MRACE2	_HISPANC	_RACE1	_RACEG22	_RACEGR4	_RACEPR1	_SEX	_AGEG5YR	_AGE65YR	_AGE80	_AGE_G	HTIN4	HTM4	WTKG3	_BMI5	_BMI5CAT	_RFBMI5	_CHLDCNT	_EDUCAG
1	2	1	1	1	1	2	13	2	80	6						9	1	4
1	2	1	1	1	1	2	13	2	80	6	63	160	6804	2657	3	2	1	2
1	2	1	1	1	1	2	8	1	56	5	62	157	6350	2561	3	2	1	4
1	2	1	1	1	1	2	14	3	73	6	65	165	6350	2330	2	1	1	2
1	2	1	1	1	1	2	5	1	43	3	62	157	5398	2177	2	1	1	3
1	2	1	1	1	1	1	13	2	80	6	71	180	8482	2608	3	2	1	2
2	2	2	2	2	2	2	13	2	80	6	65	165	6260	2296	2	1	1	4
1	2	1	1	1	1	2	13	2	80	6	64	163	7348	2781	3	2	1	2
1	2	1	1	1	1	2	12	2	78	6	67	170				9	1	4
1	2	1	1	1	1	2	11	2	72	6	66	168	8165	2905	3	2	1	4
1	2	1	1	1	1	2	13	2	80	6	63	160	7484	2923	3	2	1	1
1	2	1	1	1	1	2	13	2	80	6	63	160	5942	2321	2	1	1	4
2	2	2	2	2	2	1	8	1	57	5	68	173	8528	2859	3	2	1	2
1	2	1	1	1	1	1	10	2	65	6	71	180	10659	3278	4	2	1	4
1	2	1	1	1	1	2	13	2	80	6	66	168	7121	2534	3	2	1	2
1	2	1	1	1	1	2	10	2	65	6	62	157	6441	2597	3	2	1	2
2	2	2	2	2	2	2	14	3	70	6	62	157	6123	2469	2	1	1	4

(BRFSS, CDC 2022)

Accuracy table based on lifestyle for Pre-diabetes or Borderline Diabetes

Algorithm	Precision of Yes	Precision of No	Correctly Classified %	Incorrectly Classified %	Accuracy
Random Forest Classifier	0.6708	0.6467	65.67	34.33	65.67
K-Nearest Neighbors	0.5990	0.6110	60.55	39.45	60.55
Gaussian Naïve Bayes	0.6260	0.5901	60.26	39.74	60.26
Logistic Regression	0.6732	0.6443	65.26	34.39	65.26
Gradient Boosting Classifier	0.6655	0.6659	66.57	33.43	66.57
Support Vector Machine	0.6647	0.6491	65.58	34.42	65.58

IX. Real-Life Significance

In the healthcare industry, the current traditional methods of diabetes and heart disease detection have relied on standardized tests and clinical assessments. These tests and assessments require so much time and hospital resources, and often include invasive medical procedures. Biopsies to assess tissue samples for signs of damage or disease; Angiographies where contrast dye is injected into blood vessels to visualize blockages or abnormalities in the arteries; Endoscopic procedures used to examine the digestive tract for signs of complications related to the disease; & Tissue Sampling For both heart disease and diabetes, for further analysis. These procedures for disease detection are the current norm when it comes to confirming heart disease or diabetes.

These procedures, while necessary, are more so done once there is a strong suspicion for disease. Healthcare providers after retrieving results from these procedures are able to intervene and come up with treatment plans; but we must consider how sometimes these tests come up as misdiagnosed or negative. How patients must fund the testing. The time and effort it takes for test results for these procedures to be complete. It not only inconveniences patients but also how it strains hospital resources and staff.

With the use of machine learning, if an individual is identified as high risk for heart disease or diabetes based on their personalized risk probability, they can begin to implement lifestyle changes or begin targeted interventions such as medication or dietary adjustments. This approach reduces the likelihood of disease progression, thereby decreasing the need for invasive diagnostic procedures like biopsies or angiographies. As a result, patients experience less physical discomfort, avoid potential complications associated with invasive tests, and contribute to reduced resource utilization within hospitals, alleviating burdens on healthcare systems.

The data needed to complete algorithm risk tests are typically gathered through routine blood tests, which are minimally invasive compared to procedures like biopsies or angiographies. These blood tests provide valuable insights into our set inputs that assess an individual's disease risk probability. We prioritize leveraging existing clinical data. To make healthcare more accessible to the average person, we want both patients and healthcare professionals to make informed decisions and tailor interventions to each patient's unique needs.

X. Conclusion

LifeFactors leverages machine learning algorithms trained on comprehensive datasets to predict the likelihood of developing diabetes and heart disease. By analyzing multiple health-related variables, including behavioral risk factors, our predictive models offer personalized risk assessments without the need for invasive tests. Our project aims to minimize patient discomfort, reduce healthcare costs, and streamline the diagnosis process.

Our project represents a significant advancement in data-driven healthcare, reshaping the detection and management of diabetes and heart disease. By combining traditional diagnostic methods with modern predictive analytics, we are moving towards a future where healthcare is more personalized, proactive, and accessible to all.

References

Ali, J., Maqsood , I., Khan, R., & Ahmad , N. (2012). Random forests and decision trees. International Journal of Computer Science Issues (IJCSI), 9(5), 272.

Battineni, G., Chintalapudi, N., & Amenta, F. (2020). Performance analysis of different machine learning algorithms in breast cancer predictions. EAI Endorsed Transactions on Pervasive Health and Technology, 6(23), 166010. https://doi.org/10.4108/eai.28-5-2020.166010

Bellman, R. 1957. Dynamic Programming. Princeton University Press

Center for Disease Control, CDC. (2015). Behavioral Risk Factor Surveillance Survey 2015

Center for Disease Control, CDC. (2020). Behavioral Risk Factor Surveillance Survey 2020

Dreiseitl, S., & Ohno-Machado, L. (2002). Logistic regression and Artificial Neural Network Classification models: A methodology review. Journal of Biomedical Informatics, 35(5–6), 352–359. https://doi.org/10.1016/s1532-0464(03)00034-0

KUMAR, D. M. D. (2020, August 3). *Learn decision tree from predicting whether a cricket match will be played or not and code by...* Medium. https://medium.com/@dilmihirdil123/learn-decision-tree-from-predicting-whether-a-cricket-match-will-be-played-or-not-and-code-by-bc603213d50d

Mahesh, B. (2019). Machine Learning Algorithms -A Review. International Journal of Science and Research (IJSR). https://doi.org/ 10.21275/ART20203995

Patel, A. (2021, June 19). *Naive Bayes algorithm*. Medium. https://imakash3011.medium.com/naive-bayes-algorithm-558d18d52c40

Sharma, V. (2022). A study on data scaling methods for machine learning. International Journal for Global Academic &amp; Scientific Research, 1(1). https://doi.org/10.55938/ijgasr.v1i1.4

Boudali, I., Chebaane, S., & Zitouni, Y. (2024). A predictive approach for myocardial infarction risk assessment using machine learning and Big Clinical Data. *Healthcare Analytics*, *5*, 100319. https://doi.org/10.1016/j.health.2024.100319

Dinh, A., Miertschin, S., Young, A., & Mohanty, S. (2019). A data-driven approach to predicting diabetes and cardiovascular disease with machine learning. *BMC Medical Informatics and Decision Making*, *19*(1). https://doi.org/10.1186/s12911-019-0918-5

Güler, H., Ulaş, M., & Santur, Y. (2023). Comparison of machine learning algorithms to predict cardiovascular heart disease risk level. *International Journal of Advanced Natural Sciences and Engineering Researches*. https://doi.org/10.59287/as-ijanser.7

Mohamed, M. H., Khafagy, M. H., Kamel, N. M. M., & Said, W. (2024). DIABETIC MELLITUS PREDICTION WITH BRFSS DATA SETS. *Journal of Theoretical and Applied Information Technology* , *102*(03), 883–897.

Mohapatra, S., Maneesha, S., Patra, P. K., & Mohanty, S. (2023). Heart diseases prediction based on stacking classifiers model. *Procedia Computer Science*, *218*, 1621–1630. https://doi.org/10.1016/j.procs.2023.01.140

Ullah, Z., Saleem, F., Jamjoom, M., Fakieh, B., Kateb, F., Ali, A. M., & Shah, B. (2022). Detecting high-risk factors and early diagnosis of diabetes using machine learning methods. *Computational Intelligence and Neuroscience*, *2022*, 1–10. https://doi.org/10.1155/2022/2557795

Wiens, J., & Shenoy, E. S. (2017). Machine Learning for Healthcare: On the verge of a major shift in healthcare epidemiology. *Clinical Infectious Diseases*, *66*(1), 149–153. https://doi.org/10.1093/cid/cix731

Xie, Z., Nikolayeva, O., Luo, J., & Li, D. (2019). Building risk prediction models for type 2 diabetes using Machine Learning Techniques. *Preventing Chronic Disease*, *16*. https://doi.org/10.5888/pcd16.190109

Appendix

https://github.com/AnuragKarki720/SeniorProjectAssumptionUniversity.git