**CHANGE LOG FOR diabetes \_ binary \_ 5050split \_ health \_ indicators \_ BRFSS2015. DATASET**

Making use of the **diabetes \_ binary \_ 5050split \_ health \_ indicators \_ BRFSS2015** dataset because of its class balance as against the other two datasets **diabetes \_ 012 \_ health \_ indicators \_ BRFSS2015** and **diabetes \_ binary \_ health \_ indicators \_ BRFSS2015** that accompanied it which had a class imbalance according to the collaborator Alex Teboul.

1. Renamed the headers columns of the dataset for readability

Below are the column headers and what their metrics

* **Diabetes\_binary** - 0 = no diabetes 1 = prediabetes 2 = diabetes
* **HighBP -** 0 = no high BP 1 = high BP
* **HighChol** - 0 = no high cholesterol 1 = high cholesterol
* **CholCheck** - 0 = no cholesterol checks in 5 years 1 = yes cholesterol check in 5 years
* **BM**I - Body Mass Index
* **Smoker** - Have you smoked at least 100 cigarettes in your entire life? [Note: 5 packs = 100 cigarettes] 0 = no 1 = yes
* **Stroke**  - (Ever told) 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** - (adult men >=14 drinks per week and adult women>=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** - days of poor mental health scale 1-30 days
* **PhysHlth** - physical illness or injury days in past 30 days scale 1-30
* **DiffWalk** - Do you have serious difficulty walking or climbing stairs? 0 = no 1 = yes
* **Sex** - 0 = female 1 = male
* **Age** - 13-level age category (\_AGEG5YR see codebook) 1 = 18-24 9 = 60-64 13 = 80 or older
* **Education** - Education level (EDUCA see codebook) scale 1-6 1 = Never attended school or only kindergarten 2 = elementary etc.
* **Income** - Income scale (INCOME2 see codebook) scale 1-8 1 = less than $10,000 5 = less than $35,000 8 = $75,000 or more

1. Substituted the age numbers representing the ages of patients for their actual age intervals.
2. Substituted in the income interval formerly represented by numbers
3. Substituted in the different diabetes category formerly represented by numbers
4. Substituted the different Bp categories formerly represented by numbers
5. Substituted the different Cholesterol categories formerly represented by numbers
6. Substituted in cholesterol check responses formerly represented by numbers
7. Substituted in the different categories of GenHlth formerly represented by numbers
8. Substituted in the sex formerly represented by numbers
9. Substituted in the education levels formerly represented by numbers

**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)

9 Refused

**Income: Income scale**

1 Less than $10,000

Notes: If "no," code 02

2 Less than $15,000 ($10,000 to less than $15,000)

Notes: If "no," code 03; if "yes," ask 01

3 Less than $20,000 ($15,000 to less than $20,000)

Notes: If "no," code 04; if "yes," ask 02

4 Less than $25,000 ($20,000 to less than $25,000)

Notes: If "no," ask 05; if "yes," ask 03

5 Less than $35,000 ($25,000 to less than $35,000)

Notes: If "no," ask 06

6 Less than $50,000 ($35,000 to less than $50,000)

Notes: If "no," ask 07

7 Less than $75,000 ($50,000 to less than $75,000)

Notes: If "no," code 08

8 $75,000 or more

**Age : 13-level age category**

1 Age 18 to 24

Notes: 18 <= AGE <= 24

2 Age 25 to 29

Notes: 25 <= AGE <= 29

3 Age 30 to 34

Notes: 30 <= AGE <= 34

4 Age 35 to 39

Notes: 35 <= AGE <= 39

5 Age 40 to 44

Notes: 40 <= AGE <= 44

6 Age 45 to 49

Notes: 45 <= AGE <= 49

7 Age 50 to 54

Notes: 50 <= AGE <= 54

8 Age 55 to 59

Notes: 55 <= AGE <= 59

9 Age 60 to 64

Notes: 60 <= AGE <= 64

10 Age 65 to 69

Notes: 65 <= AGE <= 69

11 Age 70 to 74

Notes: 70 <= AGE <= 74

12 Age 75 to 79

Notes: 75 <= AGE <= 79

13 Age 80 or older

Notes: 80 <= AGE <= 99

Model Building and Evaluation

1. Total of 21 features were used in predicting the target variable including the following:

* **HighBP -** 0 = no high BP 1 = high BP
* **HighChol** - 0 = no high cholesterol 1 = high cholesterol
* **CholCheck** - 0 = no cholesterol checks in 5 years 1 = yes cholesterol check in 5 years
* **BM**I - Body Mass Index
* **Smoker** - Have you smoked at least 100 cigarettes in your entire life? [Note: 5 packs = 100 cigarettes] 0 = no 1 = yes
* **Stroke**  - (Ever told) 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** - (adult men >=14 drinks per week and adult women>=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** - days of poor mental health scale 1-30 days
* **PhysHlth** - physical illness or injury days in past 30 days scale 1-30
* **DiffWalk** - Do you have serious difficulty walking or climbing stairs? 0 = no 1 = yes
* **Sex** - 0 = female 1 = male
* **Age** - 13-level age category (\_AGEG5YR see codebook) 1 = 18-24 9 = 60-64 13 = 80 or older
* **Education** - Education level (EDUCA see codebook) scale 1-6 1 = Never attended school or only kindergarten 2 = elementary etc.

1. No feature was removed
2. The data was splitted using the stratify and shuffle methods of the train\_test\_split function
3. The split was divided into train, validation and test data
4. Used MinMaxScaler() function to scale the data as the models overfitted with the original dataset

* Before the MinMaxScaling, Decision tree classifier had the following
* Training set accuracy score: 0.9952266023062484
* Validation set accuracy score: 0.6523487773487774

1. All models fell in the range of 0.7 – 0.8 accuracy scores and other metrics

* Hypertuning were carried out on some of the models

1. The best 2 models based on the hyperparameter tweaks are:

On the SVM

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | F1 | Accuracy | Precision | Recall |
| Train\_SVM | 0.77 | 0.74 | 0.72 | 0.83 |
| Validation\_SVM | 0.77 | 0.75 | 0.72 | 0.82 |
| Test\_SVM | 0.77 | 0.75 | 0.72 | 0.83 |
|  |  |  |  |  |

On the Decision Tree Classifier

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | F1 | Accuracy | Precision | Recall |
| Train\_SVM | 0.74 | 0.72 | 0.71 | 0.77 |
| Validation\_SVM | 0.75 | 0.73 | 0.72 | 0.77 |
| Test\_SVM | 0.74 | 0.72 | 0.70 | 0.77 |

1. Conclusively Support Vector Machine (SVM), attaining a remarkable recall score of 83%, an F1 score of 77%, and an accuracy score of 73%.