Got Health?

Machine Learning

Dataset

BRFSS survey

- .ASC file contains 441,456 records and 330 columns
- each record contains an individual's BRFSS survey responses
- each column contains a specific feature derived from the survey question

Heart Disease Health Indicators Dataset (Kaggle)

- subset of features from the *BRFSS survey 2015*
- selection based on important risk factors for heart disease and other chronic illnesses like diabetes
- Dataset has been largely modified and cleaned making it suitable for ML

Dataset

Dependent Variable

- *Heart disease*: Respondents that have ever reported having coronary heart disease or myocardial infarction - binary

Independent Variables

- *Ordinal (label encoding)*: age, BMI, general health, mental health, physical health, Household Income
- *Binary (one hot coding)*: blood pressure, smoking, cholesterol (high), cholesterol check, Sex, physical activity, fruits, vegetables, alcohol consumption, health care, stroke, health costs, walk difficulty – binary

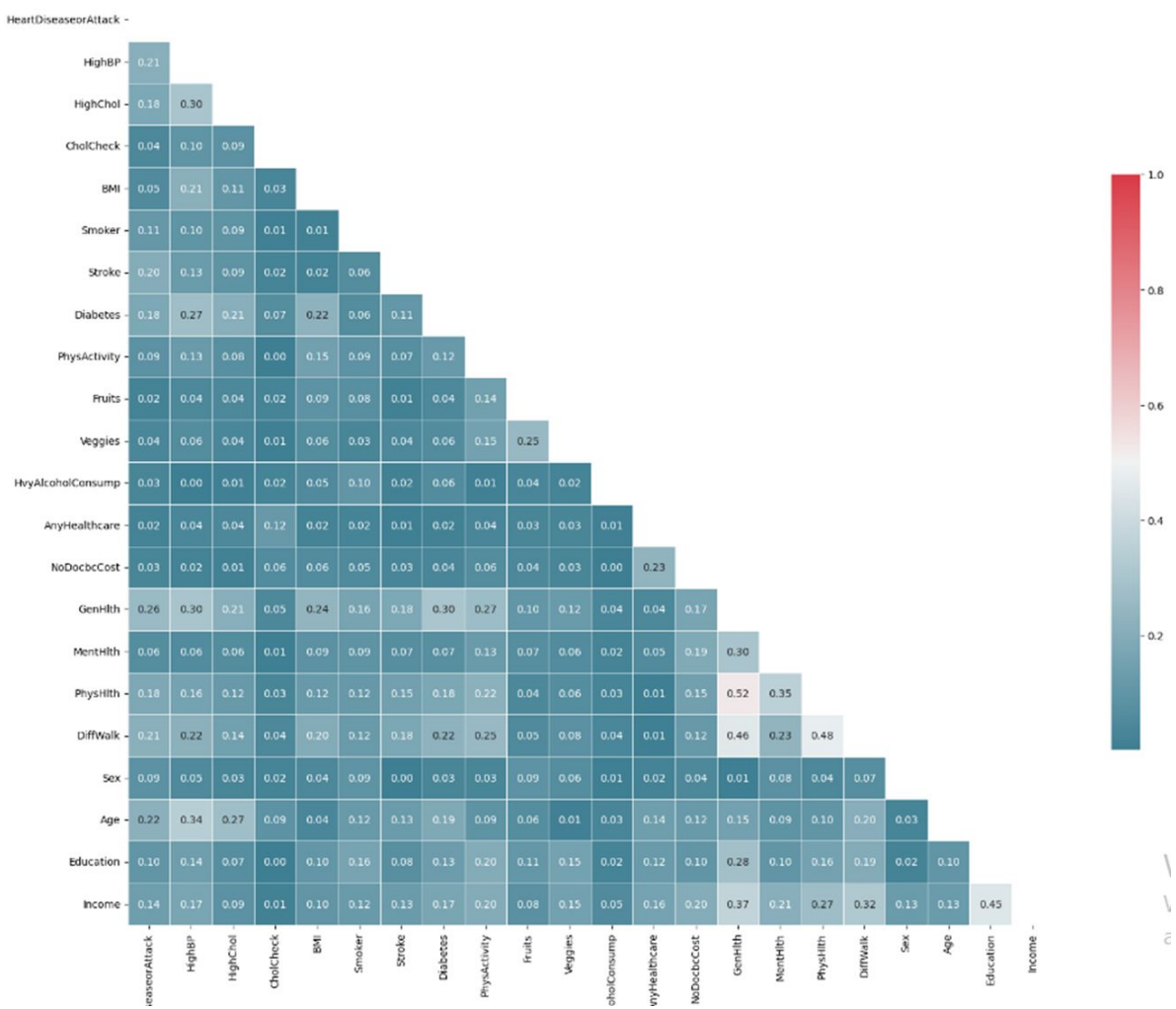
253,680 survey responses from cleaned BRFSS 2015

Dataset

253,680 Rows
21 Features

HighBP	HighChol	CholCheck	BMI	Smoker	Stroke	Diabetes	PhysActivity	Fruits	Veggies	...	AnyHealthcare	NoDocbcCost	GenHlth	MentHlth	PhysHlth
0	0	1	21	0	0	0	0	1	1	...	1	0	3	3	7
1	1	1	28	0	0	0	1	1	1	...	1	0	3	0	0
0	0	1	24	0	0	0	1	1	1	...	1	0	1	0	0
0	0	1	27	1	0	0	1	0	1	...	1	0	2	3	0
0	1	1	31	1	0	0	0	1	1	...	1	1	4	27	27
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1	0	1	29	1	0	2	1	1	1	...	1	0	3	0	10
0	0	1	25	0	0	0	1	1	1	...	1	0	2	1	10
0	1	1	28	0	0	0	1	1	1	...	1	0	3	3	0
0	0	1	24	1	0	0	0	0	1	...	1	1	4	0	0
0	0	1	23	0	0	0	1	1	1	...	1	0	2	0	0

Dataset


HeartDiseaseorAttack
HighBP
HighChol
CholCheck
BMI
Smoker
Stroke
Diabetes
PhysActivity
Fruits
Veggies
HvyAlcoholConsump
AnyHealthcare
NoDocbcCost
GenHlth
MentHlth
PhysHlth
DiffWalk
Sex
Age
Education
Income
1.0
0.8
0.6
0.4
0.2

Model

```
# 'HeartDiseaseorAttack' as target variable (y)
features = data.drop('HeartDiseaseorAttack', axis=1)
target = data['HeartDiseaseorAttack']

# Split
X_train, X_test, y_train, y_test = train_test_split(features, target, test_size=0.2, random_state=42)

# Normalize non-binary
columns_to_normalize = X_train.columns[X_train.nunique() > 2]
normalizer = MinMaxScaler()
X_train_norm = X_train.copy()  # Make a copy to preserve original DataFrame
X_train_norm[columns_to_normalize] = normalizer.fit_transform(X_train[columns_to_normalize])
X_test_norm = X_test.copy()
X_test_norm[columns_to_normalize] = normalizer.transform(X_test[columns_to_normalize])

# Undersampling
undersampler = RandomUnderSampler(random_state=42)
X_train_resampled, y_train_resampled = undersampler.fit_resample(X_train_norm, y_train)
```

Model

specifies the minimum number of samples required to split an internal node in the decision trees

helps control the growth of the tree by setting a minimum number of samples that each leaf node must have

Number of features considered at each split

number of different combinations of hyperparameters

5-fold cross-validation (5 subsets)

controls the number of CPU cores used for the computation

```python
# Initialize Random Forest model
base_rf = RandomForestClassifier(random_state=42)

# Define hyperparameter search space
param_distributions = {
    'n_estimators': [100, 200, 300],
    'max_depth': [5, 10, 15, None],
    'min_samples_split': [2, 5, 10],
    'min_samples_leaf': [1, 2, 4],
    'max_features': ['sqrt', 'log2']
}

# Perform randomized search for hyperparameter tuning
random_search = RandomizedSearchCV(
    base_rf,
    param_distributions=param_distributions,
    n_iter=20,
    cv=5,
    random_state=42,
    n_jobs=-1,
    scoring='recall'  # Using recall as the evaluation metric
)

# Fit model on resampled data
random_search.fit(X_train_resampled, y_train_resampled)

# Get the best model
final_model = random_search.best_estimator_
```

Performance Metrics

Dependent Variable *(253,680)*

- *Heart disease True = 23,893 (high imbalance)*

Without sampling

Classification Report:

	precision	recall	f1-score	support
0	0.91	0.98	0.95	45968
1	0.39	0.10	0.16	4768
accuracy			0.90	50736
macro avg	0.65	0.54	0.55	50736
weighted avg	0.86	0.90	0.87	50736

Oversampling

Classification Report:

	precision	recall	f1-score	support
0	0.92	0.96	0.94	45968
1	0.36	0.22	0.27	4768
accuracy			0.89	50736
macro avg	0.64	0.59	0.61	50736
weighted avg	0.87	0.89	0.88	50736

SMOTE

Classification Report:

	precision	recall	f1-score	support
0	0.92	0.96	0.94	45968
1	0.39	0.23	0.29	4768
accuracy			0.89	50736
macro avg	0.66	0.60	0.62	50736
weighted avg	0.87	0.89	0.88	50736

Undersampling

Classification Report:

	precision	recall	f1-score	support
0	0.97	0.72	0.83	45968
1	0.23	0.82	0.36	4768
accuracy			0.73	50736
macro avg	0.60	0.77	0.60	50736
weighted avg	0.91	0.73	0.79	50736

Performance Metrics

Undersampling

```
Confusion Matrix:
[[33210 12758]
 [  858  3910]]

Best Hyperparameters:
{'n_estimators': 200, 'min_samples_split': 5, 'min_samples_leaf': 2, 'max_features': 'sqrt', 'max_depth': 10}
```

```
X_train_resampled
```

	HighBP	HighChol	CholCheck	BMI	Smoker	Stroke	Diabetes	PhysActivity	Fruits	Veggies	...	AnyHealthcare	NoDocbcCost	GenHlth	MentHlth	Physl
143879	0	1	1	0.232558	1	0	0.0	1	1	1	...	1	0	0.25	0.000000	0.00(
211252	1	0	1	0.174419	0	0	1.0	1	0	1	...	1	0	0.25	0.000000	0.33
232980	0	0	1	0.116279	1	0	0.5	1	1	1	...	1	1	0.25	0.833333	0.00(
25411	0	0	1	0.186047	1	0	0.0	1	1	1	...	1	0	0.50	0.000000	0.00(
251205	1	0	1	0.104651	1	0	0.0	1	1	1	...	1	0	0.00	0.000000	0.00(
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
87313	1	0	1	0.220930	1	0	0.0	1	1	1	...	1	0	0.75	0.000000	0.06(
69092	1	1	1	0.174419	0	0	0.0	0	1	1	...	1	0	0.50	0.000000	0.06(
130523	0	1	1	0.162791	1	0	0.0	1	0	1	...	1	0	0.25	0.000000	0.03
85305	1	1	1	0.209302	1	0	1.0	1	0	1	...	1	0	0.75	0.066667	0.33
213458	1	1	1	0.337209	1	0	1.0	0	1	1	...	1	0	0.75	0.000000	0.23

38250 rows × 21 columns

Oversampling

```
Confusion Matrix:
[[44081  1887]
 [ 3709  1059]]

Best Hyperparameters:
{'n_estimators': 200, 'min_samples_split': 2, 'min_samples_leaf': 1, 'max_features': 'log2', 'max_depth': None}
```

```
X_train_resampled
```

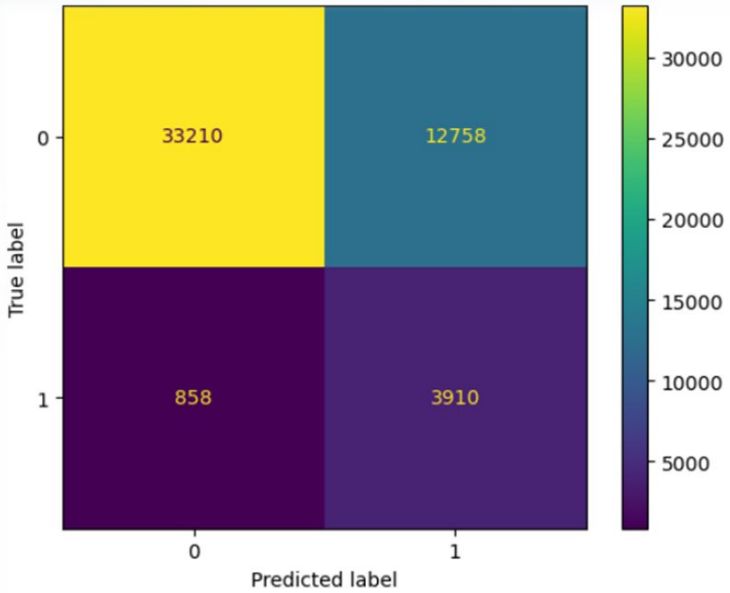
	HighBP	HighChol	CholCheck	BMI	Smoker	Stroke	Diabetes	PhysActivity	Fruits	Veggies	...	AnyHealthcare	NoDocbcCost	GenHlth	MentHlth	Physl
0	0	1	1	0.093023	1	0	0.0	1	1	1	...	1	0	0.25	0.0	0.00(
1	0	0	1	0.255814	0	0	0.0	1	0	1	...	1	0	0.50	0.0	0.00(
2	1	1	1	0.139535	0	0	1.0	1	1	1	...	1	0	0.25	0.0	0.16
3	0	1	1	0.174419	0	0	0.0	1	1	1	...	1	0	0.00	0.0	0.00(
4	0	1	1	0.139535	0	0	0.0	1	1	1	...	1	0	0.50	0.0	0.00(
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
367633	1	1	1	0.220930	0	0	0.0	1	1	1	...	1	0	0.75	0.0	0.33
367634	0	0	1	0.162791	0	0	0.0	1	1	1	...	1	0	0.25	0.0	0.00(
367635	1	1	1	0.255814	1	1	1.0	1	0	1	...	1	0	0.75	0.0	0.46
367636	0	1	1	0.220930	1	0	0.0	0	0	0	...	1	0	1.00	1.0	1.00(
367637	1	1	1	0.313953	1	0	1.0	1	0	1	...	1	0	0.75	0.0	1.00(

367638 rows × 21 columns

Confusion Matrix

Dependent Variable

- *Heart disease*



Risk Indicator Application

Pickle Files:

Input Health Variables

High Blood Pressure: Adults who have been told they have high blood pressure by a doctor, nurse, or other health professional

1

High Cholesterol

1

Cholesterol Check

1

BMI

25 − +

Smoker

1

Stroke

1

Diabetes

2

Physical Activity

0

Fruits

0

Sex (0=Female, 1=Male)

0

Age Group (1-13)

7

1 13

Education Level (1-6)

3

1 6

Income Level (1-8)

4

1 8

Confirm

Estimation:


```
{
  "prediction" : "Low Risk"
  "probability" : 0.14909608673464403
  "top_factors" : "            feature  contribution
                  Age              Age      0.101585
                  GenHlth      GenHlth      0.092340
                  Income        Income      0.015402
                  Education  Education      0.005520
                  BMI              BMI      0.003917
                  Smoker        Smoker      0.000000
                  CholCheck  CholCheck      0.000000
                  HighChol    HighChol      0.000000
                  HighBP        HighBP      0.000000
                  Stroke        Stroke      0.000000"
}
```