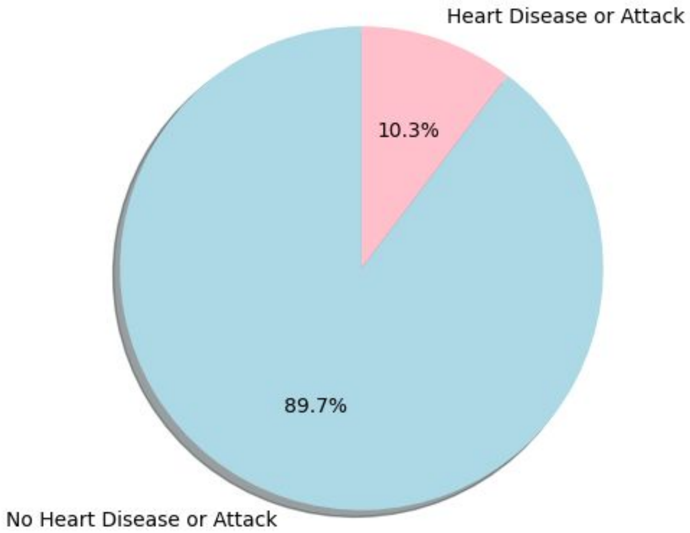
Application of ML techniques in the medical field, prediction of heart diseases

Elisa Ancarani

MSc of Artificial Intelligence, Unibo

The Dataset

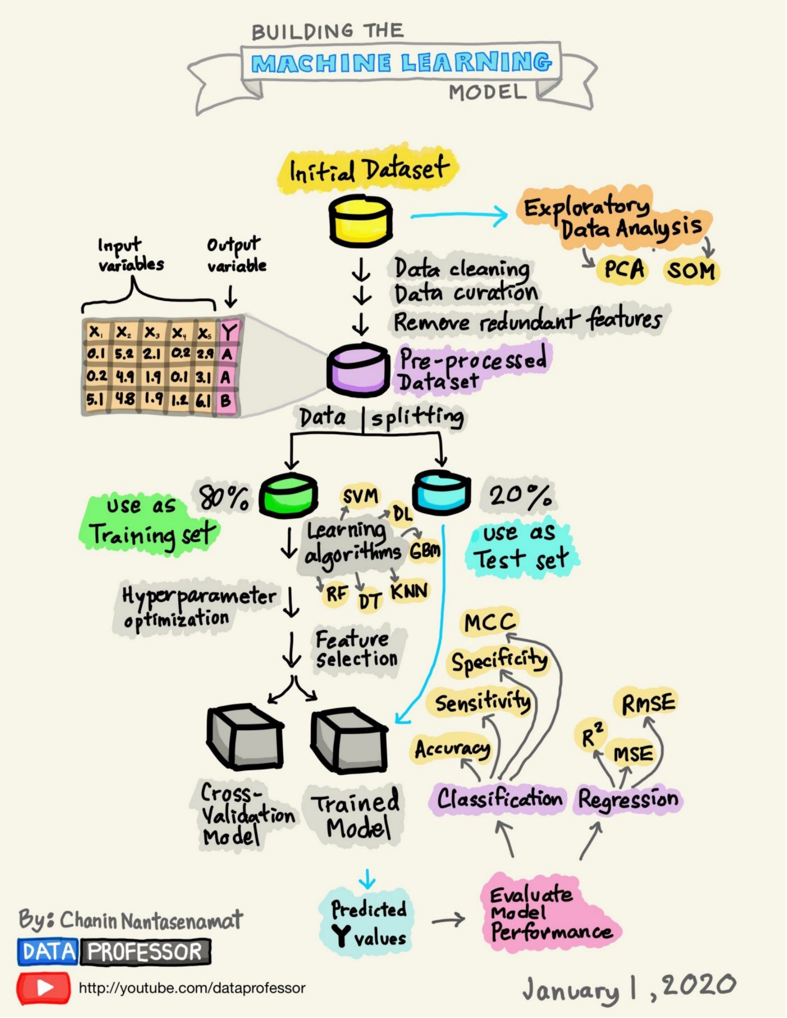
- 22 features, 253680 records
- Target Feature (binary): Heart Disease or Attack
- **Main Challenge**: Heavily Imbalanced!!



Link to the dataset here

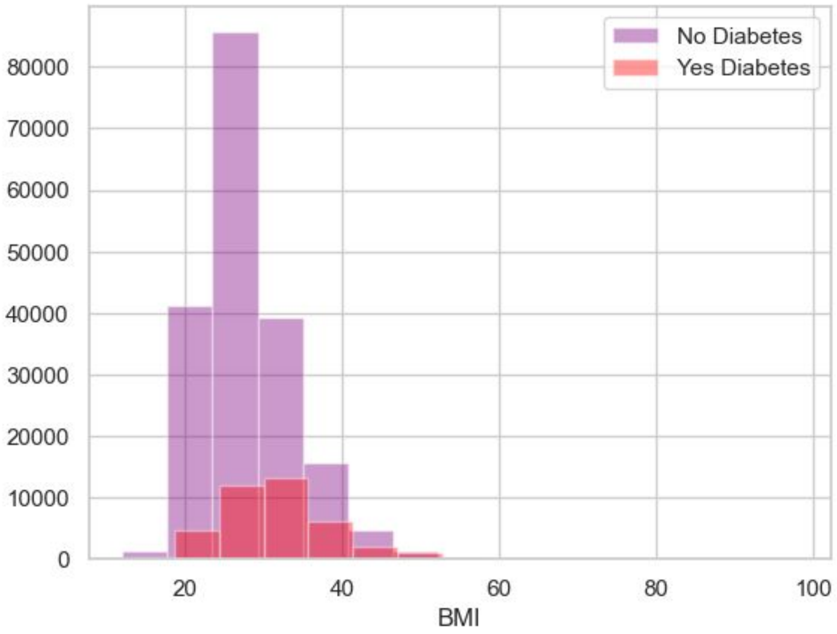
Overall Pipeline

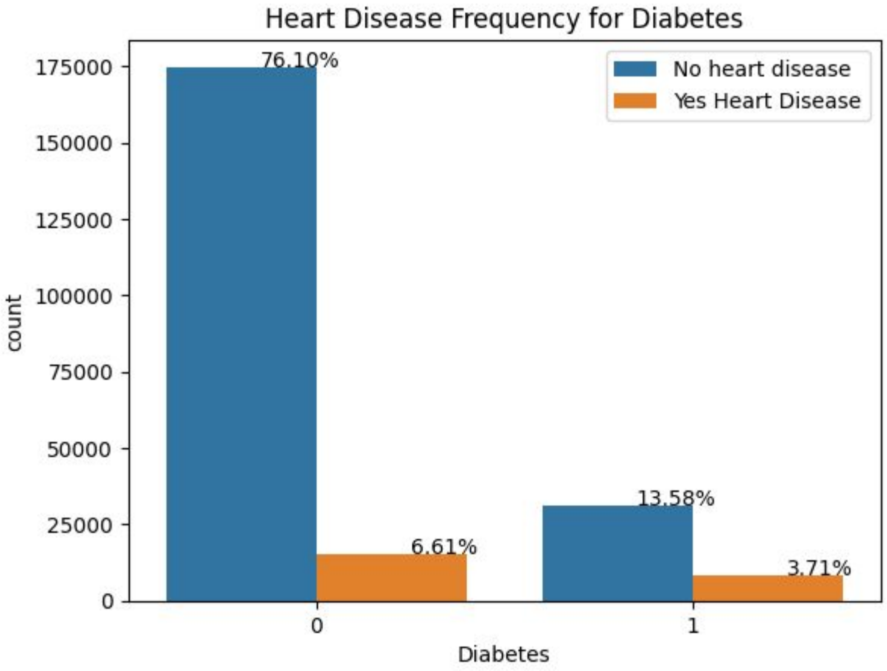
- Exploratory Data Analysis (EDA)
- Pre processing
  - data cleaning
  - outlier detection
  - data transformation
  - duplicates
- Handling Imbalance Problem
- Application of ML Algorithm
  - KNN, Decision Tree, Random Forest, XGCBoost
- Evaluation of the models (F1 Score, Accuracy, Precision)

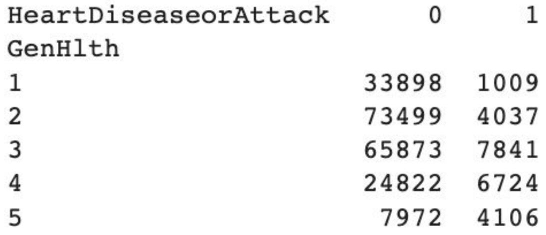


X_1	X_2	X_3	X_4	X_5	Y
0.1	5.2	2.1	0.2	2.9	A
0.2	4.9	1.9	0.1	3.1	A
5.1	4.8	1.9	1.2	6.1	B

EDA



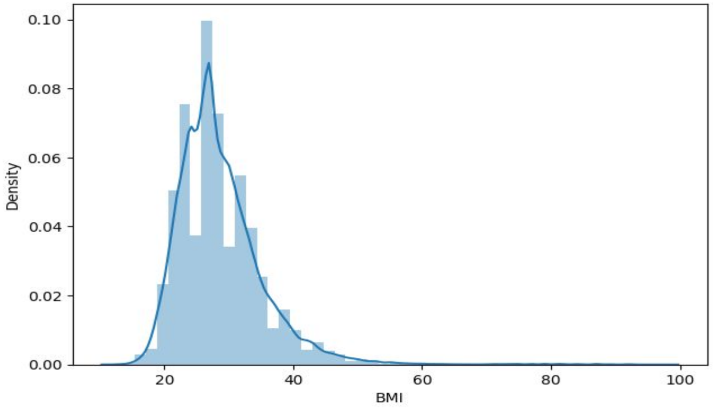


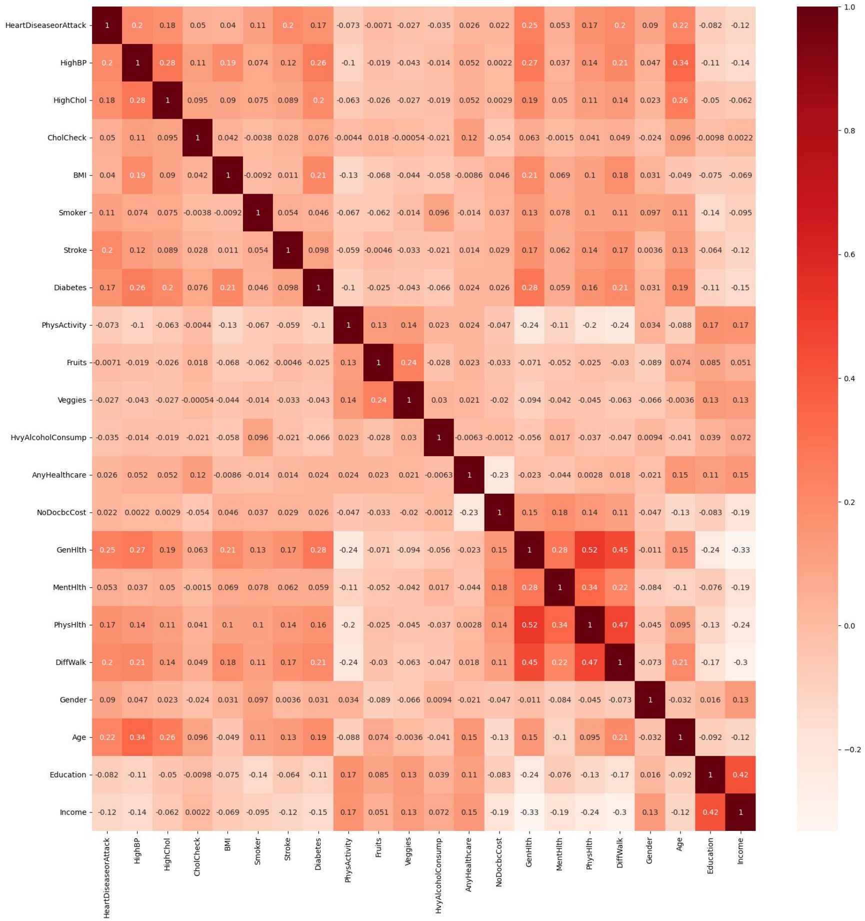


```
HeartDiseaseorAttack      0     1
GenHlth
1                     33898  1009
2                     73499  4037
3                     65873  7841
4                     24822  6724
5                      7972  4106
```

Pre Processing

- Not particularly high correlated features, features are all kept
- Removing outliers would mean, in our case, removing relevant data.
- Drop duplicates values
- Features from float64 to int32





	HeartDiseaseorAttack	HighBP	HighChol	CholCheck	BMI	Smoker	Stroke	Diabetes	PhysActivity	Fruits	Veggies	HvyAlcoholConsump	AnyHealthcare	NoDocbcCost	GenHlth	MentHlth	PhysHlth	DiffWalk	Gender	Age	Education	Income
HeartDiseaseorAttack	1	0.2	0.18	0.05	0.04	0.11	0.2	0.17	-0.073	-0.0071	-0.027	-0.035	0.026	0.022	0.25	0.053	0.17	0.2	0.09	0.22	-0.082	-0.12
HighBP	0.2	1	0.28	0.11	0.19	0.074	0.12	0.26	-0.1	-0.019	-0.043	-0.014	0.052	0.0022	0.27	0.037	0.14	0.21	0.047	0.34	-0.11	-0.14
HighChol	0.18	0.28	1	0.095	0.09	0.075	0.089	0.2	-0.063	-0.026	-0.027	-0.019	0.052	0.0029	0.19	0.05	0.11	0.14	0.023	0.26	-0.05	-0.062
CholCheck	0.05	0.11	0.095	1	0.042	-0.0038	0.028	0.076	-0.0044	0.018	-0.00054	-0.021	0.12	-0.054	0.063	-0.0015	0.041	0.049	-0.024	0.096	-0.0098	0.0022
BMI	0.04	0.19	0.09	0.042	1	-0.0092	0.011	0.21	-0.13	-0.068	-0.044	-0.058	-0.0086	0.046	0.21	0.069	0.1	0.18	0.031	-0.049	-0.075	-0.069
Smoker	0.11	0.074	0.075	-0.0038	-0.0092	1	0.054	0.046	-0.067	-0.062	-0.014	0.096	-0.014	0.037	0.13	0.078	0.1	0.11	0.097	0.11	-0.14	-0.095
Stroke	0.2	0.12	0.089	0.028	0.011	0.054	1	0.098	-0.059	-0.0046	-0.033	-0.021	0.014	0.029	0.17	0.062	0.14	0.17	0.0036	0.13	-0.064	-0.12
Diabetes	0.17	0.26	0.2	0.076	0.21	0.046	0.098	1	-0.1	-0.025	-0.043	-0.066	0.024	0.026	0.28	0.059	0.16	0.21	0.031	0.19	-0.11	-0.15
PhysActivity	-0.073	-0.1	-0.063	-0.0044	-0.13	-0.067	-0.059	-0.1	1	0.13	0.14	0.023	0.024	-0.047	-0.24	-0.11	-0.2	-0.24	0.034	-0.088	0.17	0.17
Fruits	-0.0071	-0.019	-0.026	0.018	-0.068	-0.062	-0.0046	-0.025	0.13	1	0.24	-0.028	0.023	-0.033	-0.071	-0.052	-0.025	-0.03	-0.089	0.074	0.085	0.051
Veggies	-0.027	-0.043	-0.027	-0.00054	-0.044	-0.014	-0.033	-0.043	0.14	0.24	1	0.03	0.021	-0.02	-0.094	-0.042	-0.045	-0.063	-0.066	-0.0036	0.13	0.13
HvyAlcoholConsump	-0.035	-0.014	-0.019	-0.021	-0.058	0.096	-0.021	-0.066	0.023	-0.028	0.03	1	-0.0063	-0.0012	-0.056	0.017	-0.037	-0.047	0.0094	-0.041	0.039	0.072
AnyHealthcare	0.026	0.052	0.052	0.12	-0.0086	-0.014	0.014	0.024	0.024	0.023	0.021	-0.0063	1	-0.23	-0.023	-0.044	0.0028	0.018	-0.021	0.15	0.11	0.15
NoDocbcCost	0.022	0.0022	0.0029	-0.054	0.046	0.037	0.029	0.026	-0.047	-0.033	-0.02	-0.0012	-0.23	1	0.15	0.18	0.14	0.11	-0.047	-0.13	-0.083	-0.19
GenHlth	0.25	0.27	0.19	0.063	0.21	0.13	0.17	0.28	-0.24	-0.071	-0.094	-0.056	-0.023	0.15	1	0.28	0.52	0.45	-0.011	0.15	-0.24	-0.33
MentHlth	0.053	0.037	0.05	-0.0015	0.069	0.078	0.062	0.059	-0.11	-0.052	-0.042	0.017	-0.044	0.18	0.28	1	0.34	0.22	-0.084	-0.1	-0.076	-0.19
PhysHlth	0.17	0.14	0.11	0.041	0.1	0.1	0.14	0.16	-0.2	-0.025	-0.045	-0.037	0.0028	0.14	0.52	0.34	1	0.47	-0.045	0.095	-0.13	-0.24
DiffWalk	0.2	0.21	0.14	0.049	0.18	0.11	0.17	0.21	-0.24	-0.03	-0.063	-0.047	0.018	0.11	0.45	0.22	0.47	1	-0.073	0.21	-0.17	-0.3
Gender	0.09	0.047	0.023	-0.024	0.031	0.097	0.0036	0.031	0.034	-0.089	-0.066	0.0094	-0.021	-0.047	-0.011	-0.084	-0.045	-0.073	1	-0.032	0.016	0.13
Age	0.22	0.34	0.26	0.096	-0.049	0.11	0.13	0.19	-0.088	0.074	-0.0036	-0.041	0.15	-0.13	0.15	-0.1	0.095	0.21	-0.032	1	-0.092	-0.12
Education	-0.082	-0.11	-0.05	-0.0098	-0.075	-0.14	-0.064	-0.11	0.17	0.085	0.13	0.039	0.11	-0.083	-0.24	-0.076	-0.13	-0.17	0.016	-0.092	1	0.42
Income	-0.12	-0.14	-0.062	0.0022	-0.069	-0.095	-0.12	-0.15	0.17	0.051	0.13	0.072	0.15	-0.19	-0.33	-0.19	-0.24	-0.3	0.13	-0.12	0.42	1

Sampling Method

Methods:

- Random Oversampling, Adasyn
- SMOTE (5 versions)
  - standard 🙂
  - Hybrid: SMOTE + Tomek Links 🙂
  - Hybrid: SMOTEENN
    - Last two points are implemented using two different resampling strategies
- Random Undersampling, NearMiss
- Hybrid: Random Oversampling + Random Oversampling
- Hybrid: SMOTE + Random Undersampling

Models

- K-Nearest Neighbour
- Decision Tree
- Random Forest
- XGBoost - Classifier

Metrics

- Models are evaluated focusing on true positives (sick patients) w.r.t to true negatives (healthy patients).
- Evaluation considers a trade off between accuracy and f1-score
- Also Precision has an important role in the choice of the best model

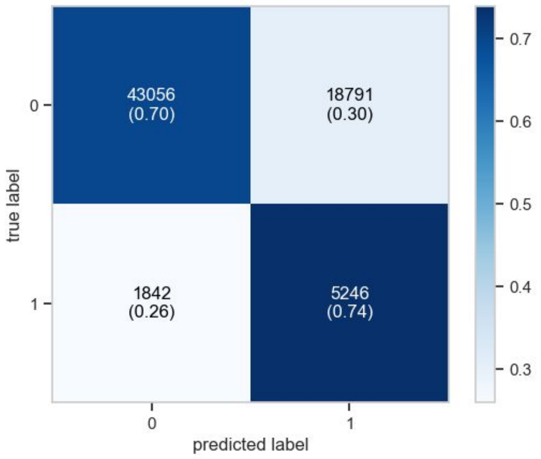
The Dataset

- Dataset Division: 70% to train the model, 30% to test it
- All models are trained with Cross Validation with cv=5 to help mitigating overfitting issue.
  - Except for KNN -> main issue is that it was really slow and the results were worse compared to other models 🐌

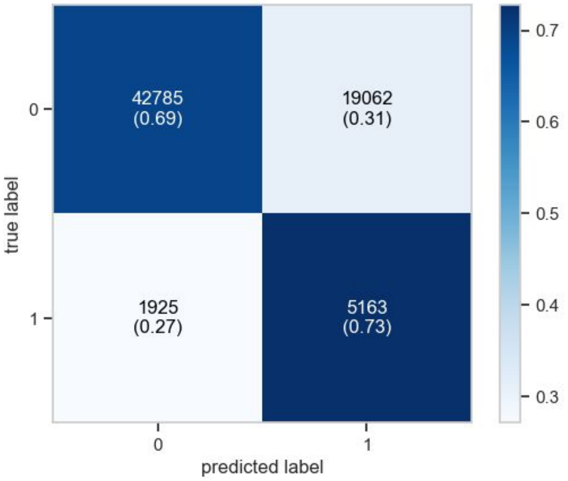
Results

Model	Accuracy	F1-Score	Precision
Random Forest	70%	(0) 81 (1) 34	(0) 96% (1) 22%
XGBC	70%	(0) 80 (1) 33	(0) 96% (1) 21%

Random Forest



XGBClassifier



Comments

- High Accuracy doesn’t mean necessarily good results
    - an high accuracy was given also when the number of true positives was very low, those cases were not taken into account
- With some sampling strategies such as random oversampling and random oversampling I obtained apparently good results, those strategies have two main drawbacks:
        - ROS - randomly duplicates samples in the minority class → model is not able to generalise well
        - RUS - randomly removes samples from the majority class → looses lot of info
- SMOOTEN with edited nearest neighbour is not that bad, but it was not as good as the other two versions (lower accuracy and f1-score).
- Precision is not as high as hoped but the number of samples classified but the number of false positives samples is much higher compared to the number of true positives.

Thanks for the attention!