RMIT UNIVERSITY

MACHINE LEARNING

COSC2753 - Sem A, 2024

Individual Assignment 01: Diabetes classification

PRSENTED BY: LUU QUOC NHAT (S3924942)

PRESENTED TO: DR. NGUYEN THIEN BAO

I. Problem statement

Diabetes, a metabolic disorder characterized by increased blood sugar levels, presents a significant medical threat to millions worldwide. Early detection and intervention are imperative for mitigating its impact, yet a large portion of the population remains undiagnosed. Utilizing a comprehensive dataset of health indicators, this study aims to develop a simulated predictive model to assess the likelihood of an individual being diabetic.

II. Dataset Overview

The provided training and testing datasets consist of 202,944 and 50,736 rows, respectively. Both have 25 columns, in which 23 features can be used for training the models, except for the identifier "Id" and the independent variable "Status." The target is a binary variable with the value of 1 indicating pre-diabetes and diabetes, and 0 indicates not. As depicted in Table 1, the features are grouped into different aspects to analyze the dataset more informatively.

Aspect	Features
Health perception	HighBP, HighChol, BMI, GenHlth, MentHlth, PhysHlth
Lifestyle indicators	Smoker, PhysActivity, Fruits, Veggies, HvyAlcoholConsump
Demographic characteristics	Sex, Age, Education, Income
Medical history	Stroke, HeartDiseaseOrAttack, ExtraMedTest, AnyHealthcare, NoDocbcCost, ExtraAlcoholTest, CholCheck, DiffWalk

Table 01. Group of features in different aspects

III. Explotary Data Analysis (EDA)

1. Summary Statistics (Appendix A):

Upon analyzing the summary statistics of the training dataset, several key observations and insights were drawn regarding the indicators associated with diabetes risk:

- **Health Perception:** The analysis suggests that a significant portion of the population is dealing with high blood pressure and cholesterol while also reporting extremely poor perceptions of their general, mental, and physical health.
- **Lifestyle Indicators:** Most individuals exhibit healthy lifestyle indicators such as regular physical activity and high consumption of fruits and vegetables, while a relatively small portion engages in smoking and heavy alcohol consumption.
- **Demographic Characteristics:** Demographic characteristics show a balanced gender distribution, with the population having attained relatively high levels of education and income.
- **Medical History and Healthcare Access:** The majority of the population has access to healthcare and has undergone cholesterol checks, indicating good awareness and access to medical services. However, the upper and lower bound values for ExtraMedTest and ExtraAlcoholTest warrant further investigation.

2. Handle missing and duplicate data (Appendix B):

The absence of missing data and the removal of 208 duplicate rows signify the dataset's reliability and integrity. Removing such duplicate rows reduces inflated weights and biases towards repeated observations and somehow reduces unessential computations.

3. Relationships between variables (Appendix C):

The heatmap analysis revealed strong positive correlations between ExtraMedTest & ExtraAlcoholTest and Status, while AnyHealthcare, Sex, Fruits, and NoDocbccost showed minimal correlations. Variables like HighBP, HighChol, BMI, etc., exhibited moderate correlations with Status, indicating their potential relevance in predicting diabetes risk.

4. Data Distribution (Appendix D):

This section visualizes target and categorical independent variable distributions with bar charts and numerical variable distributions with histograms. There's a significant class imbalance in the target variable, highlighting the need for careful model training and evaluation. Yeo-Johnson transformation is applied to normalize highly skewed numerical features like BMI, MentHlth, and PhysHlth.

5. Handle outliers (Appendix E):

Winsorization, a robust and reliable transformation technique, is employed to treat the detected outliers. By replacing the extreme data points, particularly instances whose values are out of the specified lower and upper bound (e.g., 0.05 and 0.95 quantities respectively), by less extreme values, this approach mitigates the influence of extreme data points.

IV. Model Proposal:

Diabetes prediction is a critical task that can benefit from various machine learning models. Considering the dataset's features and the objective of predicting the onset of diabetes, the following three types of models are suitable: Logistic Regression, Decision Trees, and Random Forests.

1.Logistic Regression models:

Logistic Regressions are statistical models inherently used for binary classification problems. They predicts the interpretable probability that a given instance belongs to a particular class (e.g., diabetes or no diabetes) by modeling the linear relationships between independent variables.

- Strengths:

- **Interpretability:** Logistic regression provides coefficients that indicate the magnitude of the impact of each feature on its probabilistic output [1].
- **Linear Separation:** they are efficient when the association between the independent variables and the logarithm of the outcome's odds demonstrates a near-linear relationship. [1].

- Weaknesses:

- **Limited complexity:** Logistic Regression models assume a linear relationship between the independent variables, which may not capture complex interactions or hidden patterns of the dataset.
- **Sensitivity to outliers:** outliers in the dataset can heavily influence the coefficients present in the hypothesis. Since the nature of these statistical models is to attempt to capture the pattern of Fall data instances even if they are noise and outliers, that makes the model not general towards unseen data.

- Suitability:

Logistic regression is suitable for this task due to its simplicity and interpretability. This is particularly helpful in this medical classification context, in which the conclusion drawn should be explainable. Additionally, although there exists but there are not many correlated pairs of independent variables in the given dataset, indicating that logistic regression is still an appropriate option.

2. Decision Tree models:

Decision trees serve as a non-parametric approach in supervised learning for classification purposes. They iteratively divide the dataset into subsets according to the most influential attribute, forming a structure resembling a tree.

- Strengths:

- **Capturing Complex Relationships:** Decision trees can represent intricate connections among features, capturing non-linear patterns effectively.
- **Interpretability:** The hierarchical structure of decision trees makes them easy to interpret and visualize, facilitating clear communication of findings.
- **Robustness:** Decision trees exhibit resilience towards outliers and missing data, rendering them appropriate for handling noisy datasets. [2].

- Weaknesses:

- **Overfitting:** Decision trees are prone to overfitting, especially when the tree grows deep.
- **Instability:** Small variations in the data can lead to a completely different tree structure, making decision trees unstable [3].

- Suitability: Decision trees are suitable for this task because they are capable of handling both numerical and categorical features, and multiple data types. They are also adept at capturing complex interactions between features.

3. Random Forest models:

Random Forest is a technique in ensemble learning that constructs numerous decision trees and integrates their forecasts to enhance reliability and resilience.

- **Strengths:** offer several advantages beyond individual decision trees, including:

- **Ensemble Learning:** By aggregating multiple decision trees, Random Forests effectively mitigate overfitting, thereby enhancing predictive accuracy.
- **Robustness:** Random Forest models excel in handling noisy or erroneous data, making them well-suited for real-world healthcare applications. Additionally, this robustness allows for the extraction of feature importances through statistical data analysis [4].

- Weaknesses:

- **Complexity:** Random Forests can be more challenging to interpret compared to standalone decision trees due to the intricacies of the combined model, requiring additional effort to understand and explain [5].
- **Computationally Demanding:** Training numerous decision trees within a Random Forest ensemble can be computationally intensive, especially with large datasets.

- Suitability: The given dataset exhibits a strong imbalanced distribution, with a disproportionate number of instances representing either diabetic or non-diabetic cases, requiring robust modeling techniques to prevent bias towards the majority class. Additionally, the dataset might contain multicollinear features, where certain attributes are highly correlated, potentially affecting the performance of linear models like logistic regression but allowing tree-based methods like Random Forest to capture complex interactions more effectively.

V. Model Implementation:

1.Logistic Regression models:

The optimal outcome of logistic regression models is acquired by polynomial features. Among all combinations of hyperparameters, the most optimal logistic regression performs well on degree (2), penalty ('l2'), lambda (101.5), max_iter (300), class_weight ('balanced') with the metrics of accuracy (0.90), recall (0.89), precision (0.81), and f1-score (0.84). Specifically:

- **degree:** determines the degree of polynomial features to be created. Higher degrees allow the model to capture more complex relationships but may also lead to overfitting.
- **penalty**: specifies the norm used in the penalization. Sample penalty terms include 'l1' (Lasso) and 'l2' (Ridge). Particularly, L1 regularization promotes sparsity by shrinking some coefficients to zero, while L2 regularization shrinks all coefficients towards zero.
- **lambda:** is the regularization parameter used to control the strength of regularization. Specifically, smaller values indicate stronger regularization.
- **max_iter:** The maximum number of iterations taken for the solvers to converge. Increasing it may lead to a better fit, but it also increases computation time.
- **class_weight:** 'balanced' adjusts weights inversely proportional to class frequencies, which is useful for imbalanced datasets

2. Decision Tree models:

The fined-tune decision tree slightly overperforms the baseline and post-pruned ones. With hyperparameters max_depth (252), min_samples_split (2), min_samples_leaf (1), criterion ('entropy'), and class_weight ('balanced'), the fined-tuned model shows reliable results with accuracy (0.93), recall (0.90), precision (0.89), and f1-score (0.90).

- **max_depth:** determines the maximum depth of the tree. Deeper trees can capture intricate relationships but are susceptible to overfitting..
- **min_samples_split:** determines the minimum number of samples needed to split an internal node. Greater values for min_samples_split hinder premature splitting, which may mitigate overfitting..
- **min_samples_leaf:** specifies the minimum number of samples required to be at a leaf node. Higher values result in simpler trees with fewer nodes.
- **Criterion:** This is the function of measuring the quality of a split. While 'gini' measures impurity using the Gini index, 'entropy' measures impurity using information gain.
- **class_weight:** "balanced" adjusts the weight with priority to the minority class, which is particularly helpful when working on an imbalanced dataset.
- **ccp_alpha:** is the complexity parameter used for Minimal Cost-Complexity Pruning. Higher values increase the amount of regularization applied to the tree. The most optimal value of ccp_alpha is the one that minimizes the performance gap between train and validation sets.

3. Random Forest:

The fine-tuned Random Forest model excels across multiple evaluation metrics, showcasing its robustness and reliability in classification tasks. With criterion ('entropy'), n_estimators (300), min_samples_split (2), and class_weight ('sub_samples'), it achieves accuracy (0.95), recall (0.92), precision (0.91), f1-score (0.91).

- **n_estimators:** determines the number of trees within the forest. Higher estimators may enhance performance across various metrics but also extends computation time.

- **max_depth:** controls the maximum depth of each tree in the forest. This hyperparameter prevents overfitting by minimizing the likelihood of overfitting individual trees.
- **min_samples_split:** specifies the minimum number of samples required to split an internal node, which is applied to all trees in the forest.
- **class_weight:** "balanced_subsample" balances the class weights for each bootstrap sample during training.

VI. Model Evaluation & Utilmate judgement:

Model	Hyper parameters	Accuracy	Recall	Precision	F1-score
Baseline Logistic Regression	max_iter=1300 class_weight="balanced"	0.88	0.87	0.79	0.82
Fine-tuned Polynomial Logistic Regression	degree=2 penalty="l1" lambda=10^-1,5 max_iter=300 class_weight="balanced"	0.89	0.88	0.81	0.84
	degree=2 penalty="l2" lambda=10^-1,5 max_iter=300 class_weight="balanced"	0.90	0.89	0.81	084
Baseline Decision Tree	class_weight="balanced"	0.93	0.89	0.88	0.89
Fine-tuned Decision Tree	max_depth: 252, min_samples_split: 2, min_samples_leaf: 1, criterion: "entropy", class_weight="balanced"	0.93	0.90	0.89	0.90
Post-pruned Decision Tree	class_weight='balanced' criterion="entropy" ccp_alpha=0.025	0.93	0.91	0.86	0.88
Baseline Random Forest	n_estimators=100, class_weight='balanced_subsample'	0.94	0.92	0.88	0.90
Fine-tuned Random Forest	criterion: "entropy" n_estimators: 300 min_samples_split: 2 class_weight="balanced_subsample"	0.95	0.91	0.92	0.92

Table 02. Model performance comparison across evaluation metrics.

As evaluating the diabetes classification models, key metrics such as accuracy, recall, precision, and F1-score were employed. Accuracy measures overall correct classification, while recall focuses on correctly identifying positive instances. Precision assesses the model's ability to avoid false positives, crucial in medical contexts where misclassifications can have severe consequences. More importantly, the F1-score combines precision and recall, valuable in imbalanced datasets.

Given the medical importance of timely diabetes detection, precision and f1-score emerges as a critical metric. Misidentifying diabetic patients can lead to delayed treatment, while false positives may trigger unnecessary interventions. The fine-tuned Random Forest model stands out, achieving high scores across all metrics: 95% accuracy, 91% recall, 92% precision, and an F1-score of 92%. Its robustness and effectiveness in handling complex data make it the optimal choice for efficiently predicting diabetic patients and facilitating timely interventions to improve patient outcomes.

In conclusion, our study underscores the potential of machine learning in predicting diabetes onset, offering valuable insights into disease risk stratification and personalized healthcare delivery. Continued research efforts aimed at refining predictive models and translating findings into clinical practice are essential for advancing the field of predictive analytics in healthcare.

Reference

[1] Nusinovici, S. et al. (2020) 'Logistic regression was as good as machine learning for predicting major chronic diseases', Journal of Clinical Epidemiology, 122, pp. 56–69. doi:10.1016/j.jclinepi.2020.03.002.

[2] Azad, C. et al. (2021) 'Prediction model using smote, genetic algorithm and decision tree (PMSGD) for classification of diabetes mellitus', Multimedia Systems, 28(4), pp. 1289–1307. doi:10.1007/s00530-021-00817-2

[3] R.-H. Li and G. G. Belford, "Instability of decision tree classification algorithms," Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, Jul. 2002. doi:10.1145/775047.775131

[4] Md. Z. Alam, M. S. Rahman, and M. S. Rahman, "A random forest based predictor for medical data classification using feature ranking," Informatics in Medicine Unlocked, vol. 15, p. 100180, 2019. doi:10.1016/j.imu.2019.100180

[5] Mervin, L.H. et al. (2021) 'Probabilistic random forest improves bioactivity predictions close to the classification threshold by taking into account experimental uncertainty', Journal of Cheminformatics, 13(1). doi:10.1186/s13321-021-00539-7.

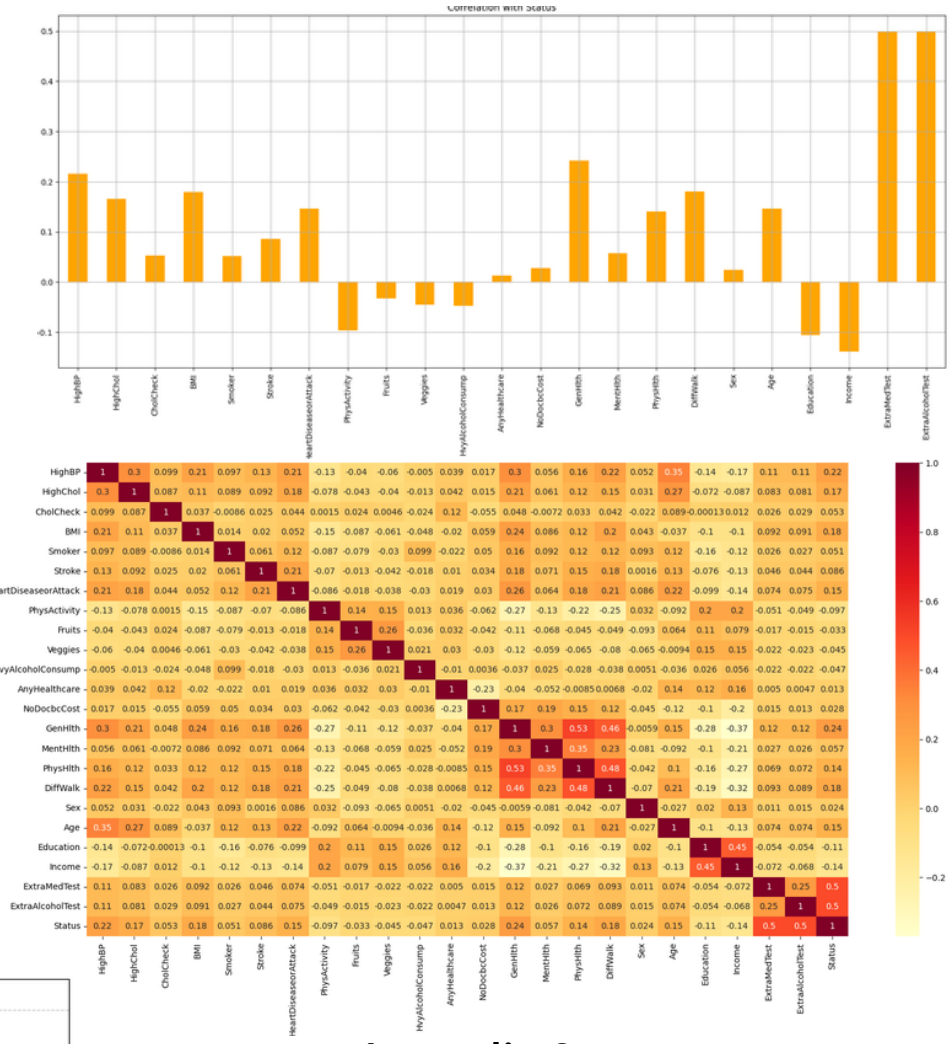
Appendix

	count	mean	std	min	25%	50%	75%	max
HighBP	202944.0	0.428700	0.494891	0.0	0.0	0.0	1.0	1.0
HighChol	202944.0	0.424344	0.494244	0.0	0.0	0.0	1.0	1.0
CholCheck	202944.0	0.962655	0.189607	0.0	1.0	1.0	1.0	1.0
BMI	202944.0	28.379824	6.612738	12.0	24.0	27.0	31.0	98.0
Smoker	202944.0	0.442634	0.496700	0.0	0.0	0.0	1.0	1.0
Stroke	202944.0	0.040844	0.197929	0.0	0.0	0.0	0.0	1.0
HeartDiseaseorAttack	202944.0	0.094391	0.292372	0.0	0.0	0.0	0.0	1.0
PhysActivity	202944.0	0.756302	0.429313	0.0	1.0	1.0	1.0	1.0
Fruits	202944.0	0.635372	0.481327	0.0	0.0	1.0	1.0	1.0
Veggies	202944.0	0.811519	0.391096	0.0	1.0	1.0	1.0	1.0
HvyAlcoholConsump	202944.0	0.055912	0.229752	0.0	0.0	0.0	0.0	1.0
AnyHealthcare	202944.0	0.951543	0.214730	0.0	1.0	1.0	1.0	1.0
NoDocbcCost	202944.0	0.083693	0.276928	0.0	0.0	0.0	0.0	1.0
GenHlth	202944.0	2.514024	1.070370	1.0	2.0	2.0	3.0	5.0
MentHlth	202944.0	3.196971	7.427247	0.0	0.0	0.0	2.0	30.0
PhysHlth	202944.0	4.256455	8.736665	0.0	0.0	0.0	3.0	30.0
DiffWalk	202944.0	0.168707	0.374494	0.0	0.0	0.0	0.0	1.0
Sex	202944.0	0.439545	0.496333	0.0	0.0	0.0	1.0	1.0
Age	202944.0	8.037449	3.051568	1.0	6.0	8.0	10.0	13.0
Education	202944.0	5.050245	0.985601	1.0	4.0	5.0	6.0	6.0
Income	202944.0	6.055641	2.070140	1.0	5.0	7.0	8.0	8.0
ExtraMedTest	202944.0	-7.408660	75.993743	-199.0	-55.0	0.0	40.0	199.0
ExtraAlcoholTest	202944.0	-7.560041	75.927137	-199.0	-55.0	0.0	40.0	199.0
Status	202944.0	0.175571	0.380455	0.0	0.0	0.0	0.0	1.0

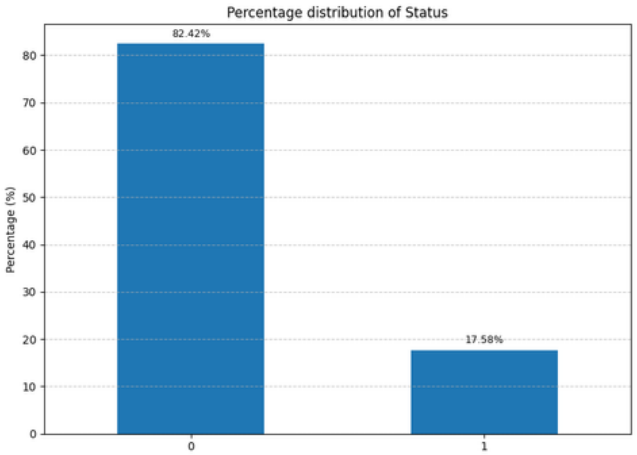
Appendix A

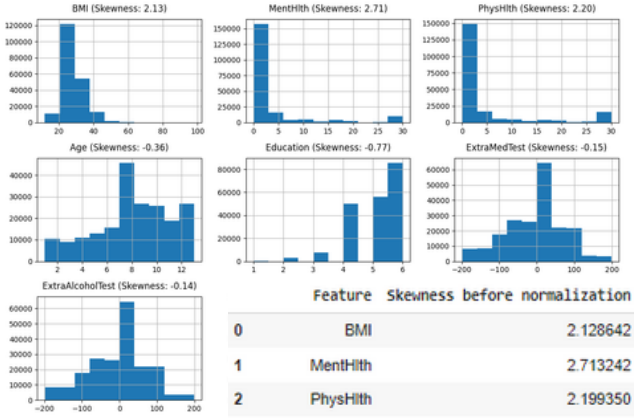
```
Number of duplicate rows found: 208
Shape of the training dataset before removing duplicates:  (202944, 24)
Shape of the training dataset after removing duplicates:  (202736, 24)
```

Appendix B



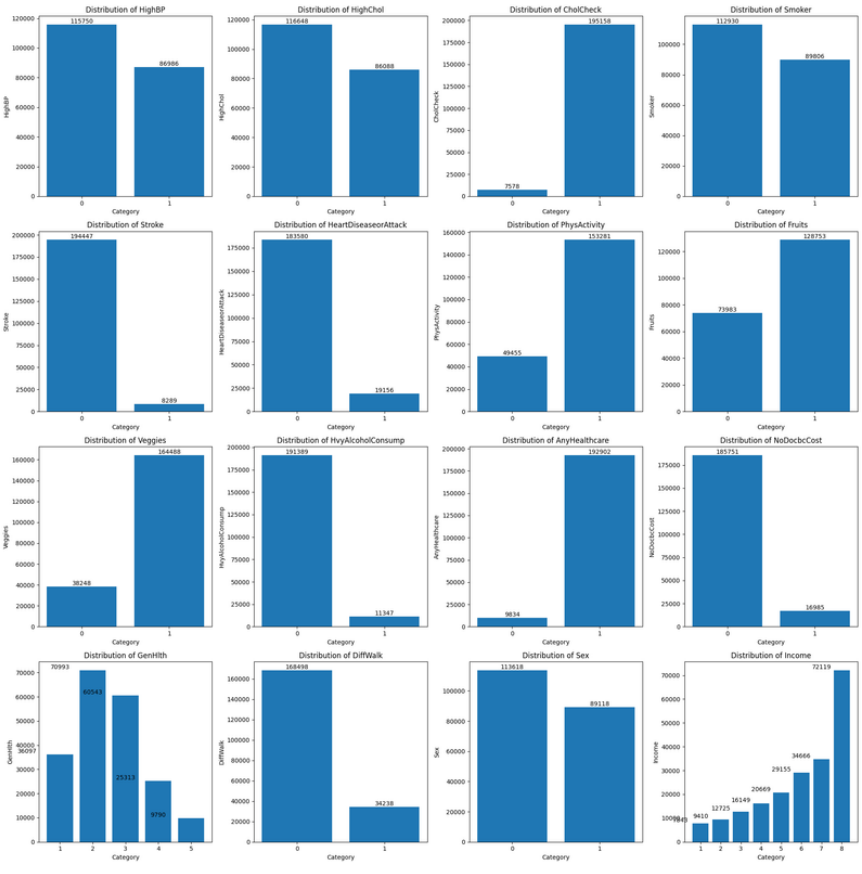
Appendix C

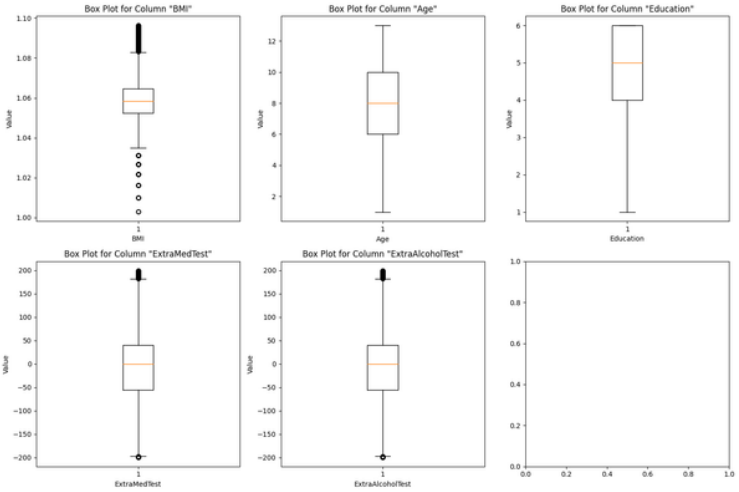




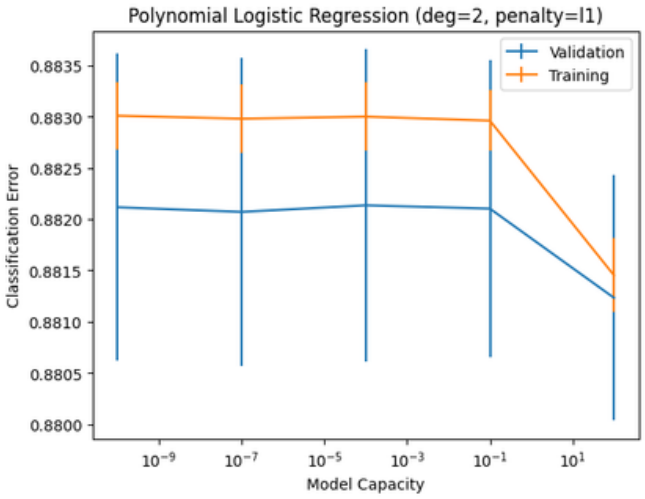
	Feature	Skewness before normalization
0	BMI	2.128642
1	MentHlth	2.713242
2	PhysHlth	2.199350
3	Age	-0.360907
4	Education	-0.774172
5	ExtraMedTest	-0.145355
6	ExtraAlcoholTest	-0.141193

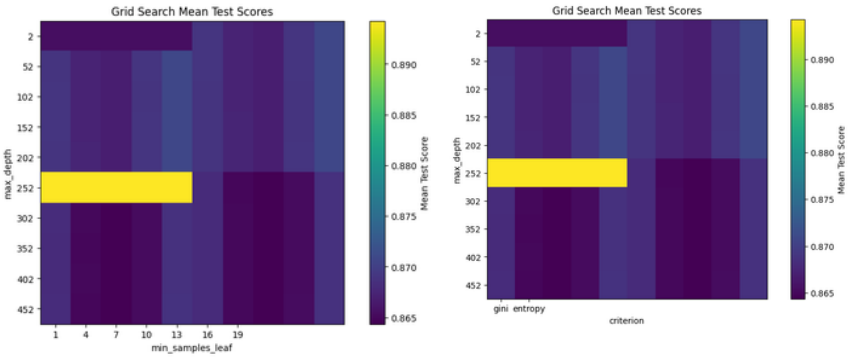
Appendix D

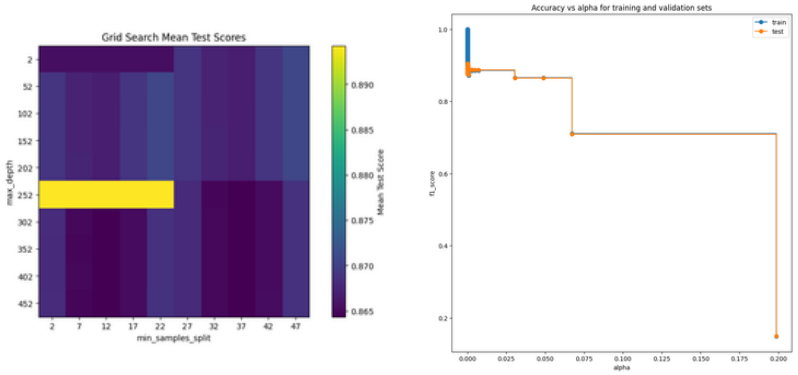




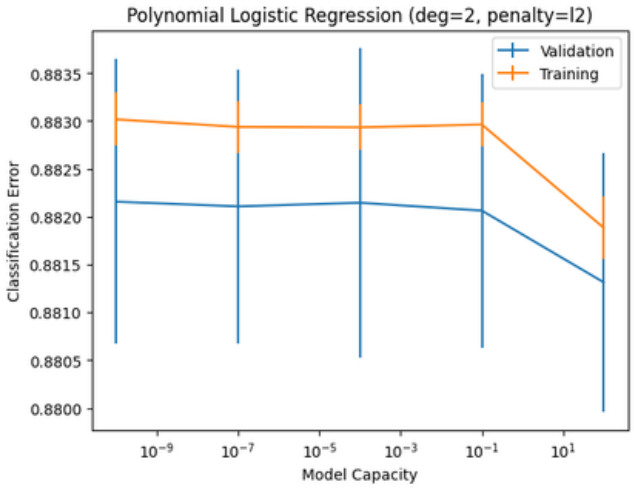
Appendix E



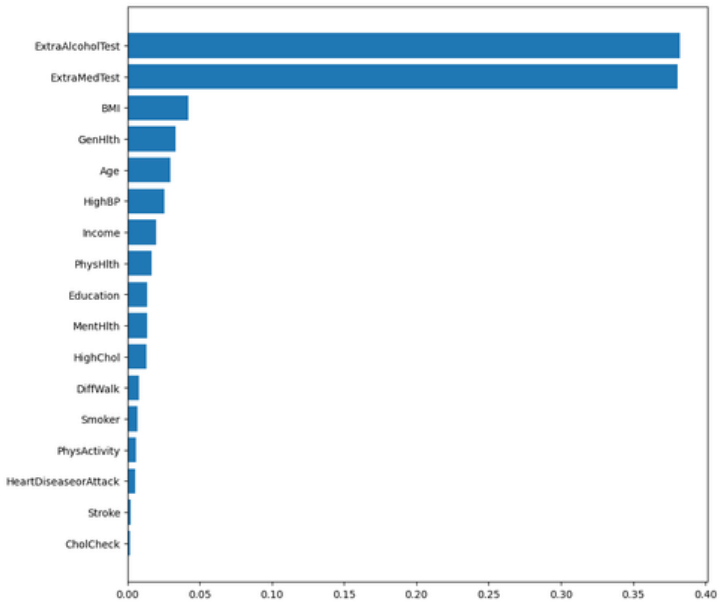


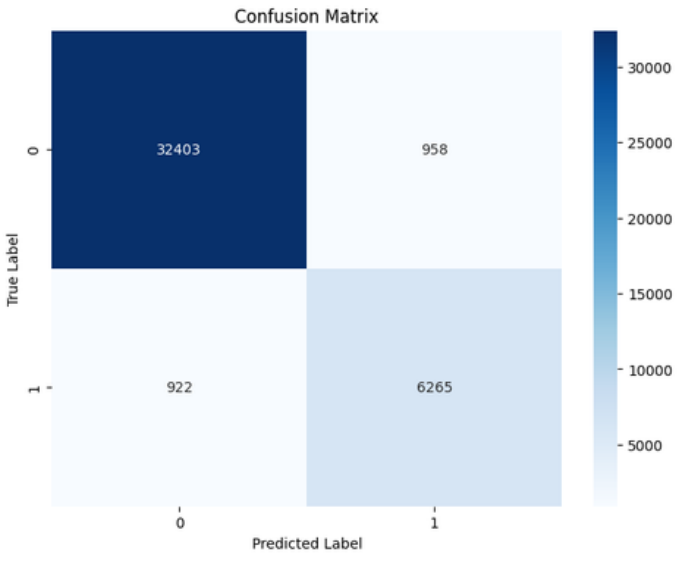


Appendix G



Appendix F





Appendix H