Chronic Diseases Prediction using Machine Learning Techniques

Mariam Kili Bechir

10-12-2023

Outline

- Concept note and implementation plan:
  - Background
  - Objectives
  - SDG Relation
- Data
  - Data Collection
  - Exploratory Data Analysis (EDA) and Feature Engineering
- Model Selection and Training
  - Model Evaluation and Hyperparameter Tuning
  - Model Refinement and Testing
- Results
- Deployment
- Future Work

Concept note and implementation plan

Background

Overview:

- Enhance chronic disease prediction using machine learning techniques
- developing predictive models for common chronic diseases such as diabetes, heart diseases, and Parkinson's.
- contribute to SDGs, and proactively manage health conditions.

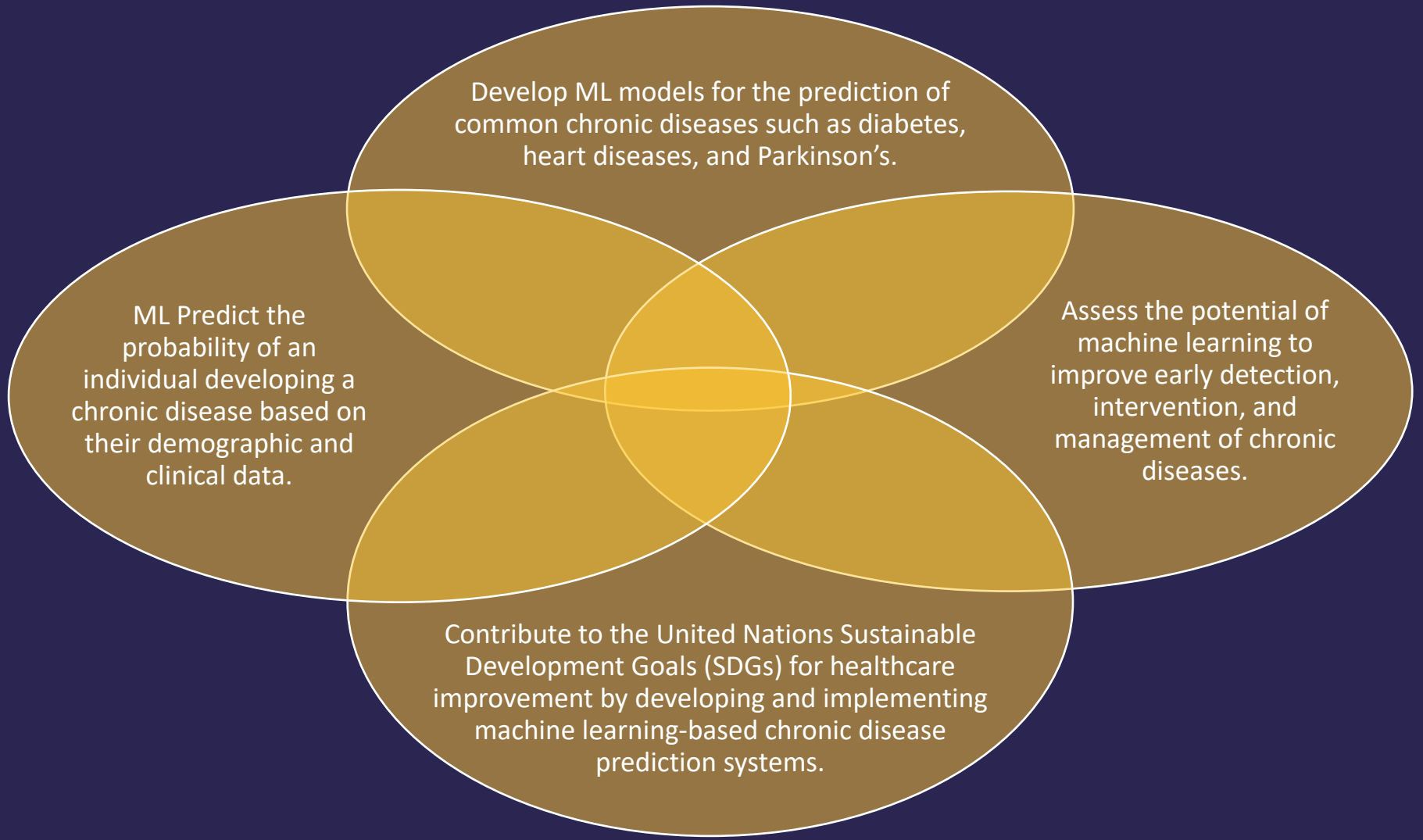
brief background

- Chronic diseases pose a global health threat, especially in regions with limited access to advanced healthcare
- leveraging AI and Machine Learning for early detection

Importance of the problem being solved

- Identifying individuals at risk early
- effective intervention
- promoting well-being and reducing global health disparities

Objectives


Develop ML models for the prediction of common chronic diseases such as diabetes, heart diseases, and Parkinson's.
ML Predict the probability of an individual developing a chronic disease based on their demographic and clinical data.
Assess the potential of machine learning to improve early detection, intervention, and management of chronic diseases.
Contribute to the United Nations Sustainable Development Goals (SDGs) for healthcare improvement by developing and implementing machine learning-based chronic disease prediction systems.

SDG Relation

Predict diseases that cause a health challenge

Ensuring healthy lives for all

improved global health healthcare systems

Access of good quality health services

Data

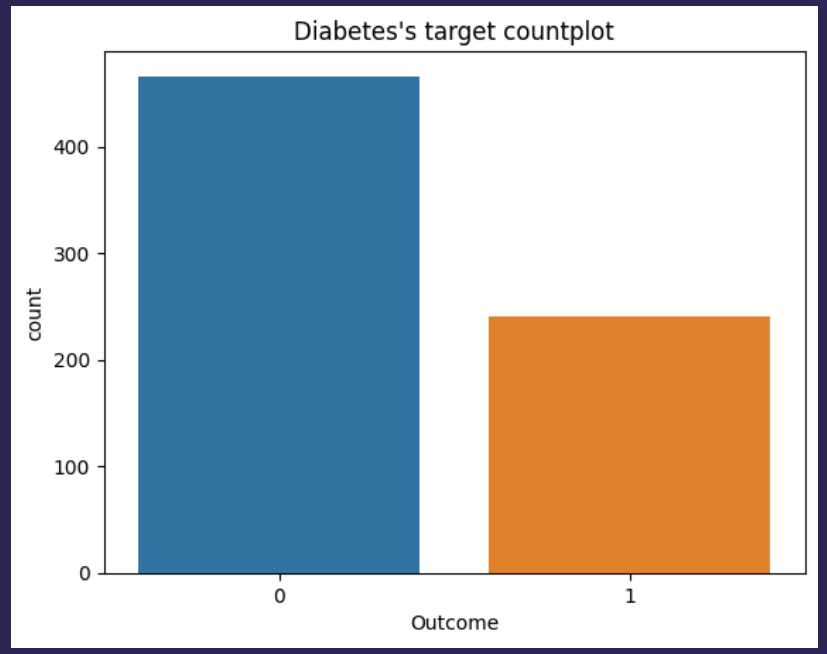
Data Collection

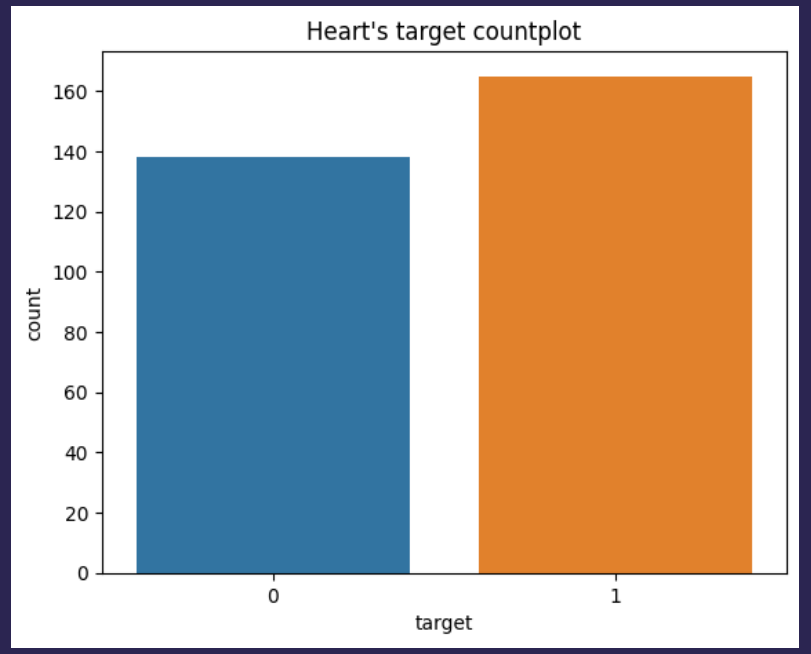
- Source of the datasets: 3 datasets was collected from Kaggle
- Preprocessing steps during data collection: matplotlib, seaborn, pandas profiling
- Data cleaning: no missing values, duplicated values removed

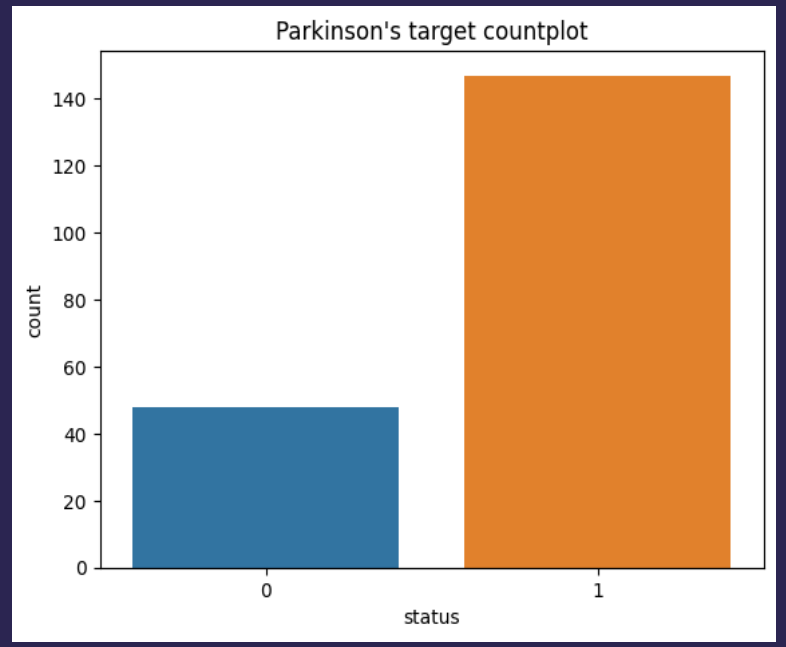
Diabetes: 768 rows × 9 columns

heart: 303 rows × 14 columns

Parkinson: 195 rows × 24 columns

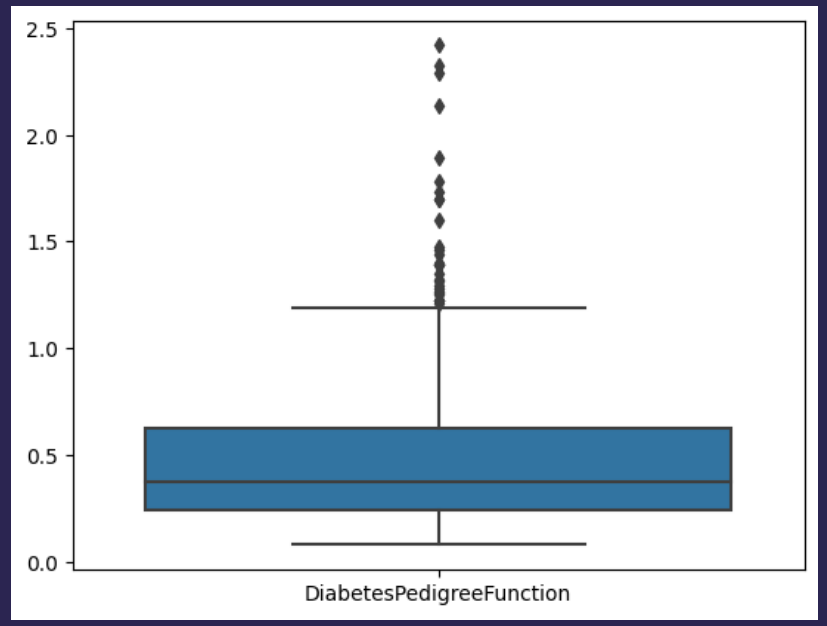


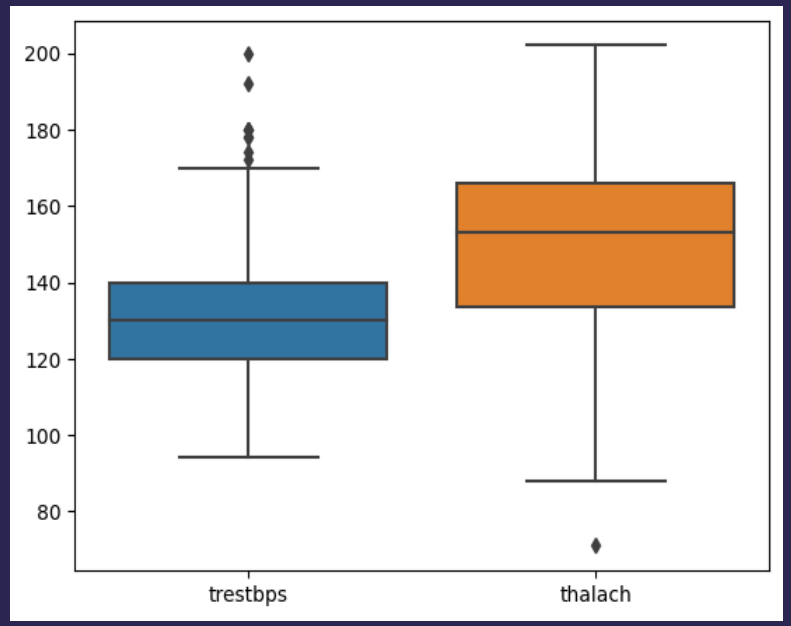


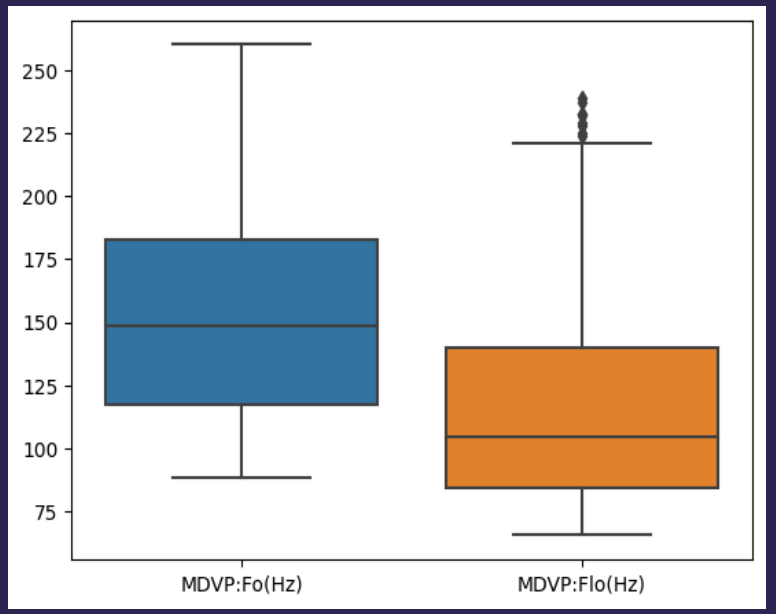


Exploratory Data Analysis (EDA) and Feature Engineering

- Handing Outliers
- Rationale behind feature engineering decisions: Improve the accuracies of the models
- Normalization techniques used: MinMaxscaler(), StandardScaler()







Model

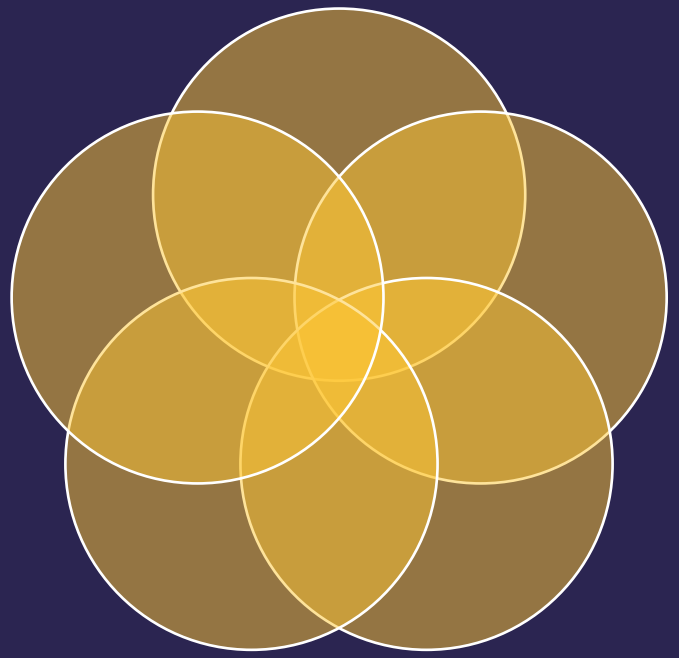
Model Selection and Training

Logistic Regression:

- simple and computationally efficient. robust to overfitting
- May not perform well with highly non-linear data.

XGBoost :

- better predictive performance, robust to outliers. It's an efficient and scalable algorithm. it can prone to overfitting, especially with insufficient regularization and requires tuning of hyperparameters.



SVM:

- effective in high-dimensional spaces, versatile due to different kernel functions, resistant to overfitting
- computationally height, with large datasets, model parameters (e.g., the choice of the kernel) can be challenging to interpret.

Random Forest:

- Robust and less prone to overfitting compared to individual decision trees. It can handle a large number of features and provides feature importance scores. The Weaknesses are Lack of interpretability compared to decision trees and the computation is sometimes more expensive.

k-Nearest Neighbors (KNN)

- non-parametric and flexible algorithm, intuitive and easy . It can capture complex decision boundaries. However, it's computationally expensive during prediction, especially with large datasets and also sensitive to irrelevant or redundant features.

Model Evaluation and Hyperparameter Tuning

- Use of hyperparameterTuning Gridsearch techniquewith cv=5 for each model
- Result of hyperparameter Tuning on Heart dataset

	model	Best score	best_parameters
0	logistic_regression	0.885245	C= 0.23, max_iter= 100
1	random_forest	0.8852459	max_depth=2, max_leaf_nodes= 6
3	SVM	0.86885245	C= 100, gamma= 1, kernel= 'linear'
4	xg_boost	0.797619	max_depth=9, max_leaf_nodes=3
5	KNN	0.702381	metric= 'manhattan', 'n_neighbors': 7

Model Evaluation and Hyperparameter Tuning

- Use of hyperparameterTuning Gridsearch techniquewith cv=5 for each model
- Result of hyperparameter Tuning on Diabetes dataset

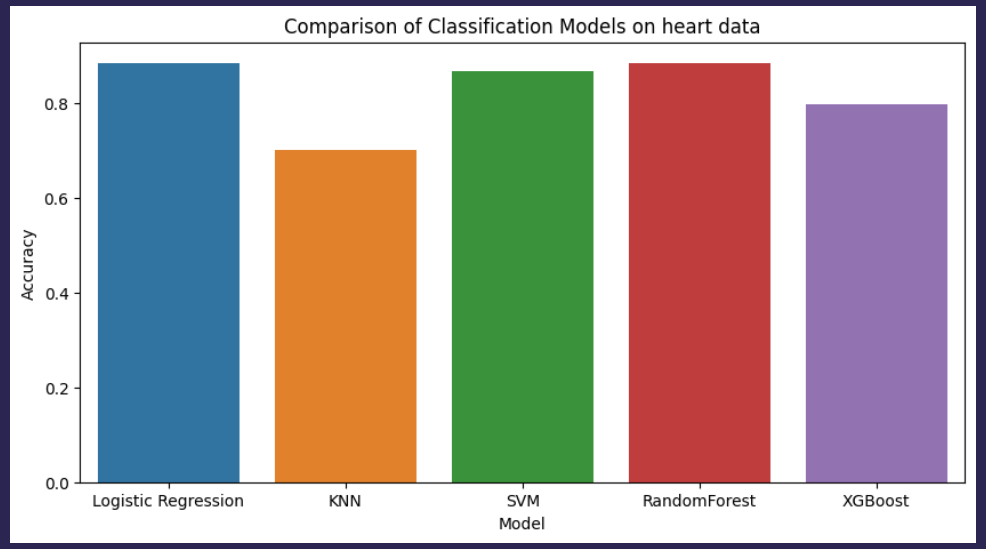
	model	best_score	best_params
0	svm	0.780168	{'C': 10, 'kernel': 'linear'}
1	knn	0.762255	{'n_neighbors': 10}
2	random_forest	0.780155	{'max_depth': 11, 'max_features': 'log2', 'n_es...
3	xg_boost	0.792207	{'max_depth': 10, 'n_estimators'=15)
4	logistic_regression	0.781781	{'C': 5, 'n_jobs': -1}

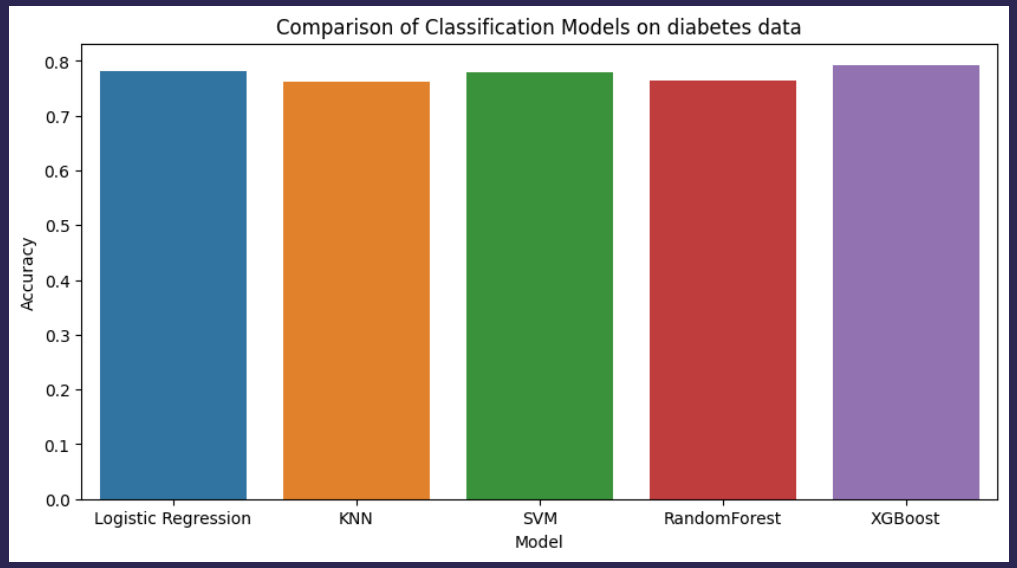
Model Evaluation and Hyperparameter Tuning

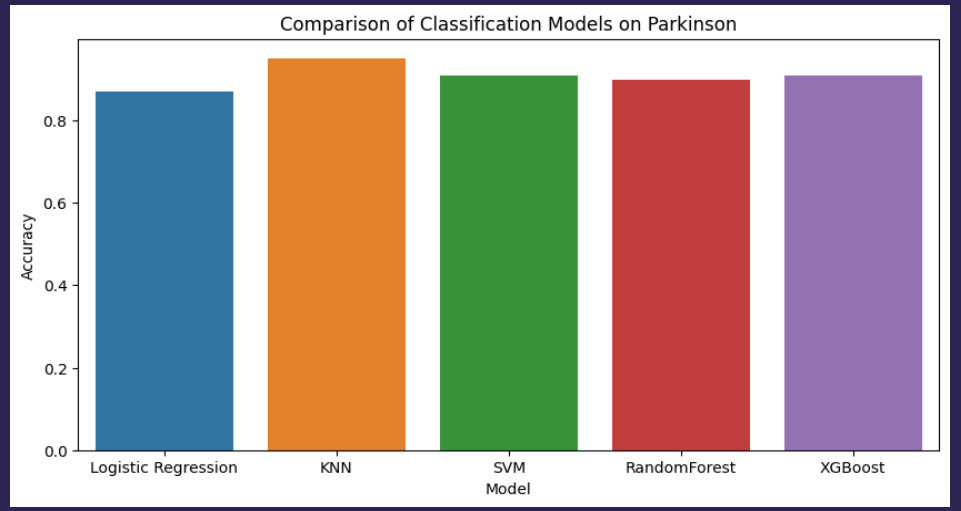
- Use of hyperparameterTuning Gridsearch techniquewith cv=5 for each model
- Result of hyperparameter Tuning on Parkinson dataset

	model	best_score	best_params
0	svm	0.910081	{'C': 1, 'gamma': 1, 'kernel': 'rbf'}
1	knn	0.897581	{'metric': 'manhattan', 'n_neighbors': 11
2	random_forest	0.878024	{'max_depth': 10, 'max_leaf_nodes': 9
3	xg_boost	0.884476	{'max_depth': 6, 'max_leaf_nodes': 3
4	logistic_regression	0.833266	{'C': 0.012742749857031334, 'max_iter': 100

Model Evaluation and Hyperparameter Tuning

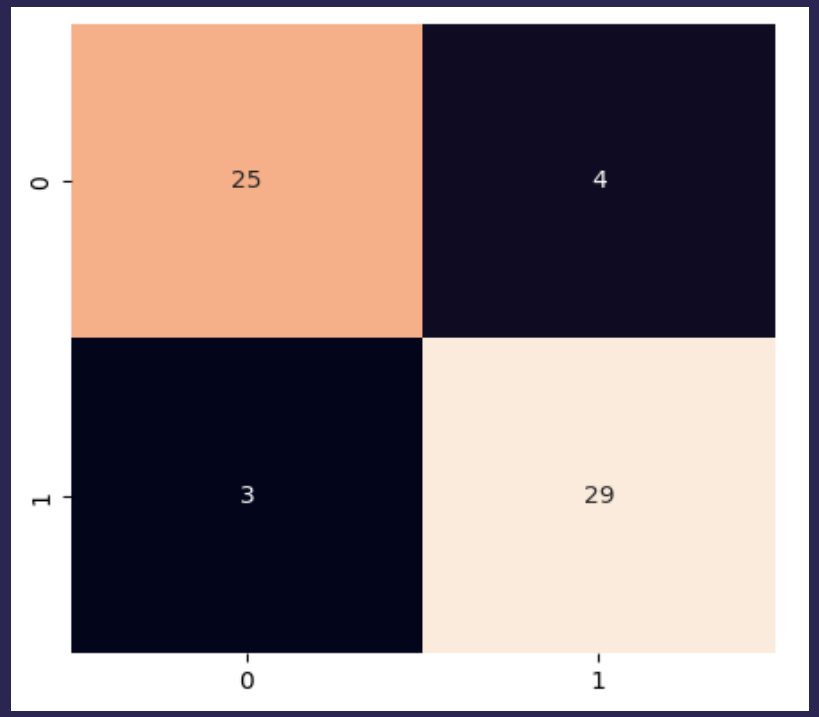




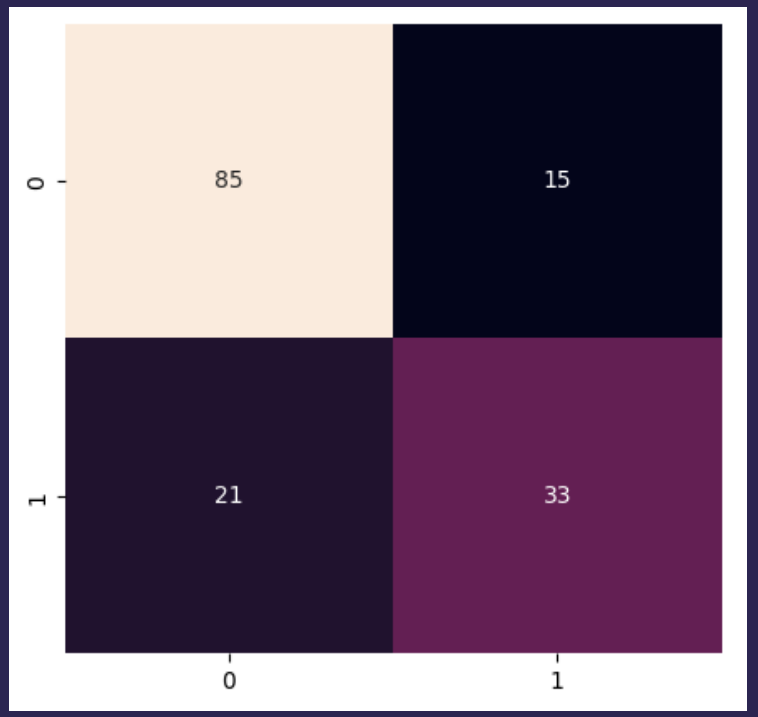


Model Refinement and Testing

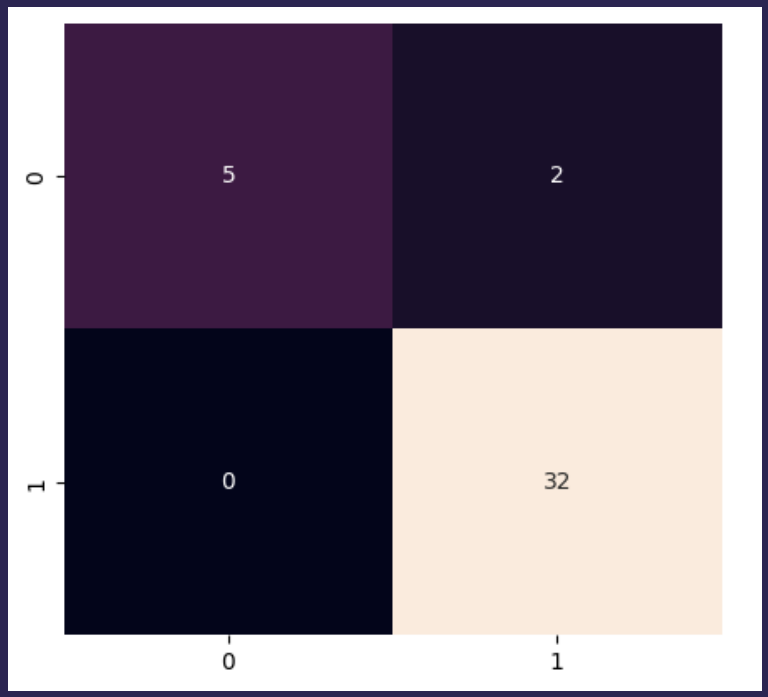
- Techniques used for model improvement: Hyperparameter Tuning, standardization and MinMaxScaler feature normalization techniques are used, Recursive Feature Elimination with Cross-Validation (RFECV) feature selection technique for LR on heart data and For diabetes dataset, we used feature_importances_ of XGBoost
- Model metrics:



1-confusion_matrix with Logistic Regression for heart dataset



2-confusion matrix with RandomForest for diabetes dataset



confusion matrix with KNN for Parkinson dataset

Results

Evaluation Results

Disease type	Selected Model	Scores	
		Training score	Test Score
Heart disease	Logistic Regression	86%	89%
Diabetes	XGBoost	97%	79%
Parkinson	KNN	95%	95%

Evaluation Results

Classificatio_report for LogistciRegression

	precision	recall	f1-score	support
0	0.89	0.86	0.88	29
1	0.88	0.91	0.89	32
accuracy			0.89	61
macro avg	0.89	0.88	0.88	61
weighted avg	0.89	0.89	0.89	61

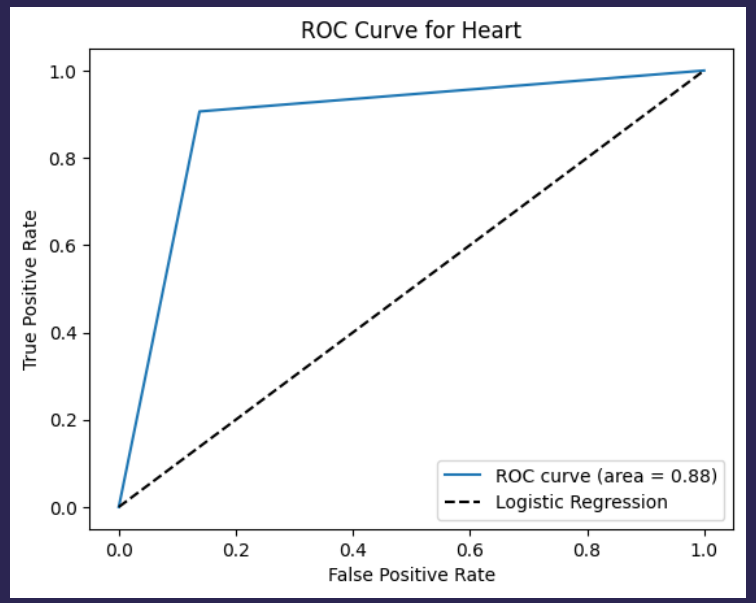
Classificatio_report for XGBoost

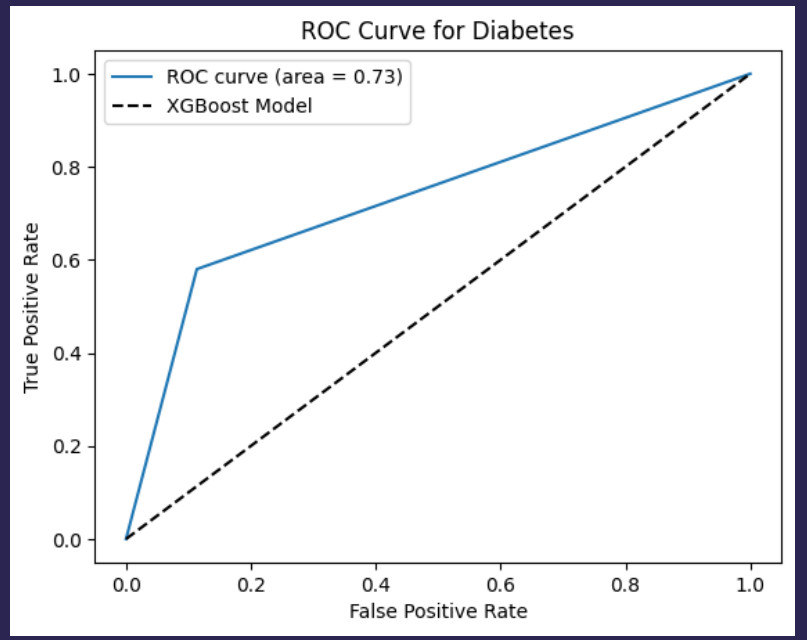
	precision	recall	f1-score	support
0	0.81	0.89	0.85	100
1	0.75	0.61	0.67	54
accuracy			0.79	154
macro avg	0.78	0.75	0.76	154
weighted avg	0.79	0.79	0.79	154

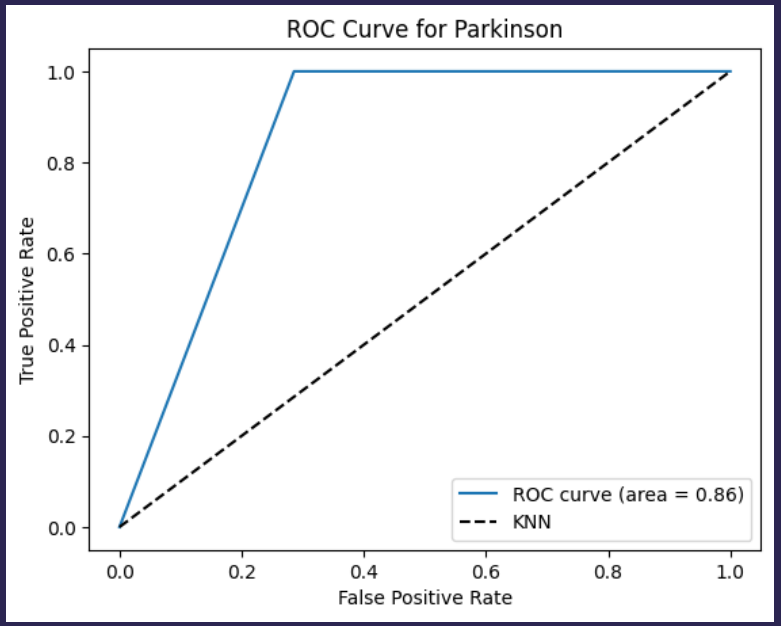
Classificatio_report for KNN

	precision	recall	f1-score	support
0	1.00	0.71	0.83	7
1	0.94	1.00	0.97	32
accuracy			0.95	39
macro avg	0.97	0.86	0.90	39
weighted avg	0.95	0.95	0.95	39

Evaluation Results

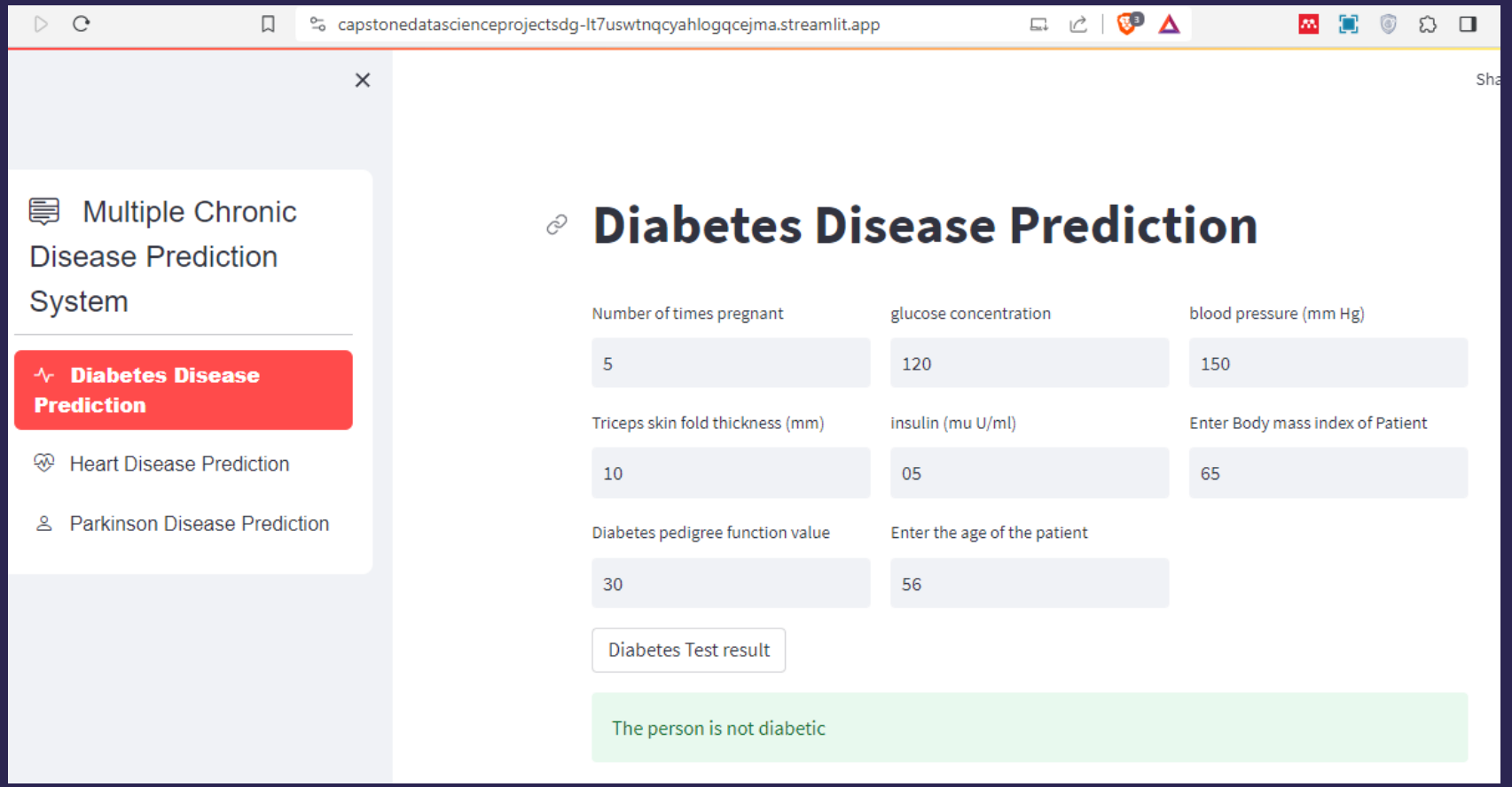






Deployment

- Libraries used: Model deployed on Streamlit cloud: https://capstonedatascienceprojectsdg-lt7uswtnqcyahlogqcejma.streamlit.app/



Future Work

- The use of this app to predict chronic diseases
- This application can be a more large application, we can add build models for more diseases like Malaria, Kidney, covid etc
- This means developping the same app on large datasets to predict more diseases
- We can also add different types of diseases such as infectious diseases for prediction

References

- 1-https://en.wikipedia.org/wiki/Sustainable_Development_Goal_3
- 2-https://www.un.org/sustainabledevelopment/health/
- 3-https://www.google.com/url?sa=i&url=https%3A%2F%2Fwww.dimins.com%2Fblog%2F2022%2F06%2F13%2Fbig-data-healthcare%2F&psig=AOvVaw1a3o-FR8Ewbof_NTEvPkou&ust=1702123866255000&source=images&cd=vfe&opi=89978449&ved=0CBEQjRxqFwoTCIjihdrn_4IDFQAAAAAdAAAAABAI
- 4-https://unicsoft.com/wp-content/uploads/2022/07/ML_Model_1140.png

5-https://www.google.com/url?sa=i&url=https%3A%2F%2Fresults.rathinamcollege.com%2F&psig=AOvVaw0nA46MaHLD0vNYfCccwfaC&ust=1702141705422000&source=images&cd=vfe&opi=89978449&ved=0CBEQjRxqFwoTCJDTgpKrgIMDFQAAAAAdAAAAABAYμ

Thank you!

ISTANBUL INTERNATIONAL CENTER FOR PRIVATE SECTOR IN DEVELOPMENT

SDG AI LAB

UN TECHNOLOGY

UNITED NATIONS TECHNOLOGY BANK FOR LEAST DEVELOPED COUNTRIES