

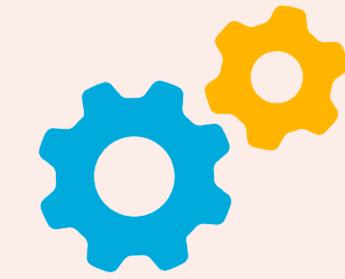
SUGAR, SPICE, AND EVERYTHING NICE: PREDICTING DIABETES RISK USING ENSEMBLE TECHNIQUES AND EXPLAINABLE ARTIFICIAL INTELLIGENCE

Applied Physics 157 Final Project
Andrea Rose V. Franco



DIABETES

The only method of preventing diabetes complications is to identify and treat the disease early (Vijayan et al. 2015)



MACHINE LEARNING

- Utilized to **forecast diseases**
- Comparison of classifiers:
 - Early-stage risks of diabetes and **ensemble techniques** were found with **higher accuracy**.

OBJECTIVES

- Compare three ensemble techniques: Bagging, Random Forest, and XGBoost in making diabetes risk predictions
- Explain the model predictions using an Explainable AI approach (SHAP value)

METHODOLOGY

Data collection and pre-processing

Training the machine learning models

Compare model metrics

Feature Importance

Explainable AI for analysis

DIABETES DATA

The screenshot shows the UC Irvine Machine Learning Repository page for the "Early Stage Diabetes Risk Prediction" dataset. The page has a header with navigation links for Datasets, Contribute Dataset, and About Us, along with a search bar and login link. The main content area features a blue header for the dataset, followed by a summary section, dataset characteristics, and dataset information. On the right side, there are download options (CSV, ZIP, IMPORT IN PYTHON), citation statistics (1 citations, 34496 views), DOI (10.24432/C5VG8H), and a license section. Below the main content, there are two tables showing variable names and their descriptions.

Dataset Characteristics:

- Multivariate
- Categorical, Integer

Subject Area: Computer Science

Associated Tasks: Classification

Instances: 520

Features: 16

Dataset Information:

Additional Information: This has been collected using direct questionnaires from the patients of Sylhet Diabetes Hospital in Sylhet, Bangladesh and approved by a doctor.

Has Missing Values? Yes

Introductory Paper:

[Likelihood Prediction of Diabetes at Early Stage Using Data Mining Techniques](#)
By M. M. F. Islam, Rahatara Ferdousi, Sadikur Rahman, Humayra Yasmin Bushra. 2019
Published in Computer Vision and Machine Intelligence in Medical Image Analysis

Variable Name

age	Variable Name
irritability	delayed_healing
gender	partial_paresis
polyuria	sudden_weight_loss
polydipsia	muscle_stiffness
weakness	alopecia
polyphagia	obesity
genital_thrush	
visual_blurring	
itching	

- 520 patients
- 16 Features
- Target variable:
Positive or negative

Variable Name	Variable Name
age	irritability
gender	delayed_healing
polyuria	partial_paresis
polydipsia	sudden_weight_loss
weakness	muscle_stiffness
polyphagia	alopecia
genital_thrush	obesity
visual_blurring	
itching	

Link: <https://archive.ics.uci.edu/dataset/529/early+stage+diabetes+risk+prediction+dataset>

PRE-PROCESSING

Original dataset

	age	polyuria	polydipsia	sudden_weight_loss	weakness	polyphagia	genital_thrush	visual_blurring	itching	irritability	delayed_healing	target
0	40	No	Yes		No	Yes	No	No	No	Yes	No	Yes
1	58	No	No		No	Yes	No	No	Yes	No	No	No
2	41	Yes	No		No	Yes	Yes	No	No	Yes	No	Yes
3	45	No	No		Yes	Yes	Yes	Yes	No	Yes	No	Yes
4	60	Yes	Yes		Yes	Yes	Yes	No	Yes	Yes	Yes	Yes
...
515	39	Yes	Yes		Yes	No	Yes	No	No	Yes	No	Yes
516	48	Yes	Yes		Yes	Yes	Yes	No	No	Yes	Yes	Yes
517	58	Yes	Yes		Yes	Yes	Yes	No	Yes	No	No	No
518	32	No	No		No	Yes	No	No	Yes	Yes	No	Yes
519	42	No	No		No	No	No	No	No	No	No	No

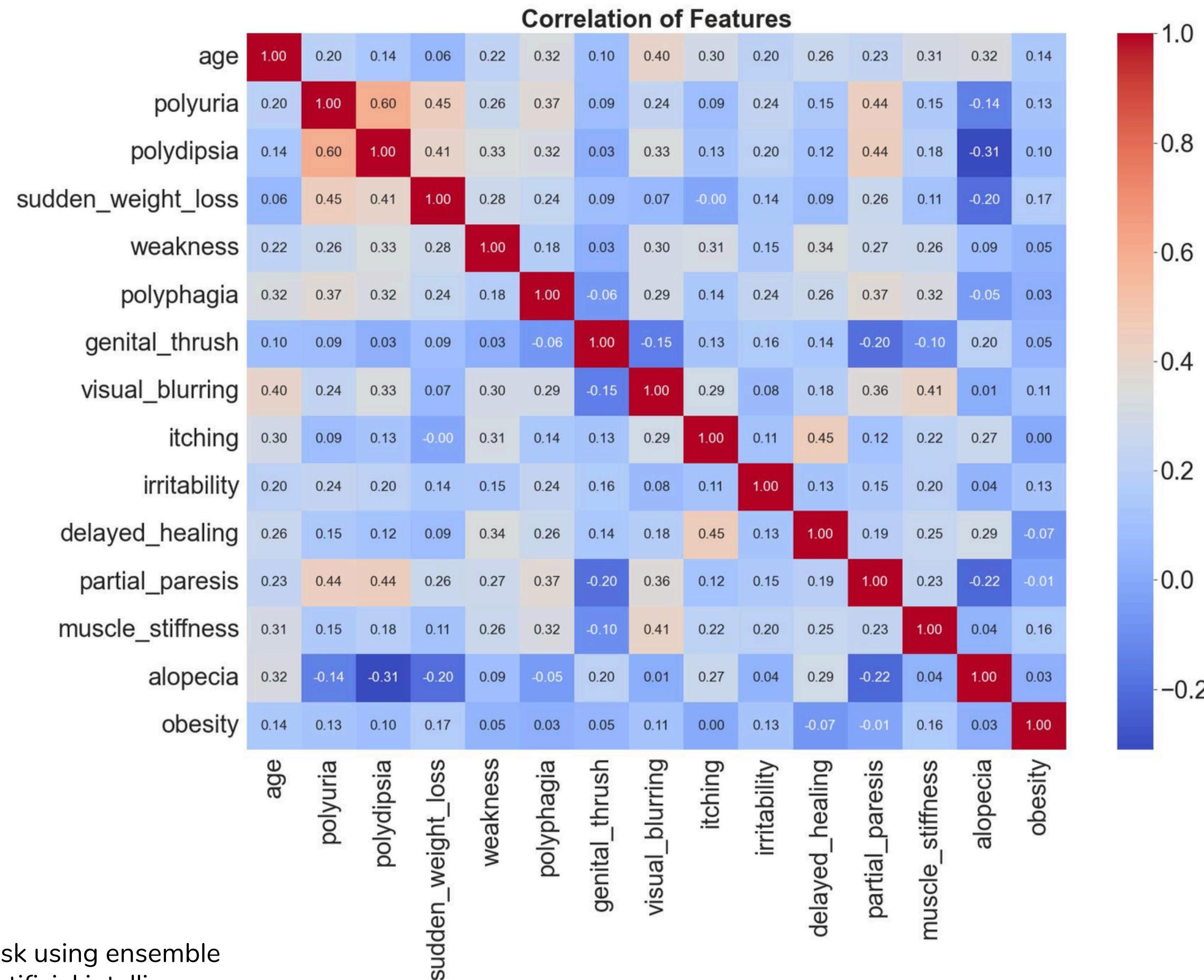
520 rows × 16 columns

Dataset with 0 or 1 for binary features

	age	polyuria	polydipsia	sudden_weight_loss	weakness	polyphagia	genital_thrush	visual_blurring	itching	irritability	delayed_healing	target
0	40	0	1		0	1	0	0	0	1	0	1
1	58	0	0		0	1	0	0	1	0	0	0
2	41	1	0		0	1	1	0	0	1	0	1
3	45	0	0		1	1	1	1	0	1	0	1
4	60	1	1		1	1	1	0	1	1	1	1
...
515	39	1	1		1	0	1	0	0	1	0	1
516	48	1	1		1	1	1	0	0	1	1	1
517	58	1	1		1	1	1	0	1	0	0	0
518	32	0	0		0	1	0	0	1	1	0	1
519	42	0	0		0	0	0	0	0	0	0	0

520 rows × 16 columns

FEATURE CORRELATION



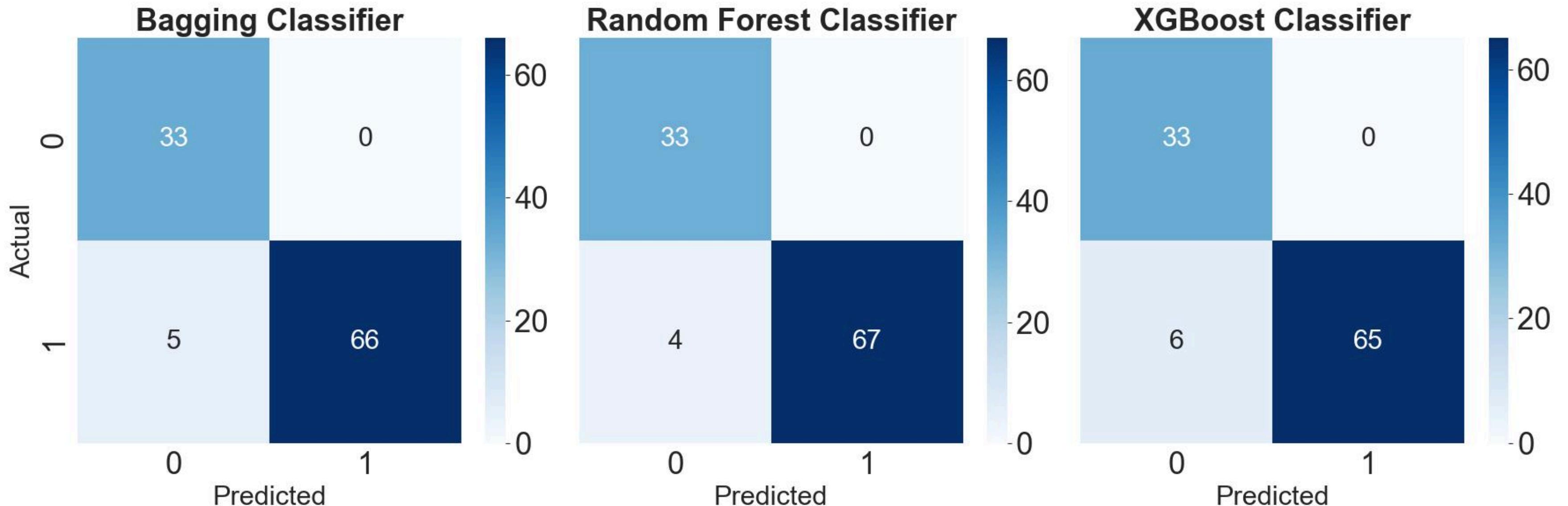
MODEL TRAINING

- 80% Training, 20% Testing
- BaggingClassifier, RandomForestClassifier, XGBClassifier
 - default parameters
- number of trees: 1000

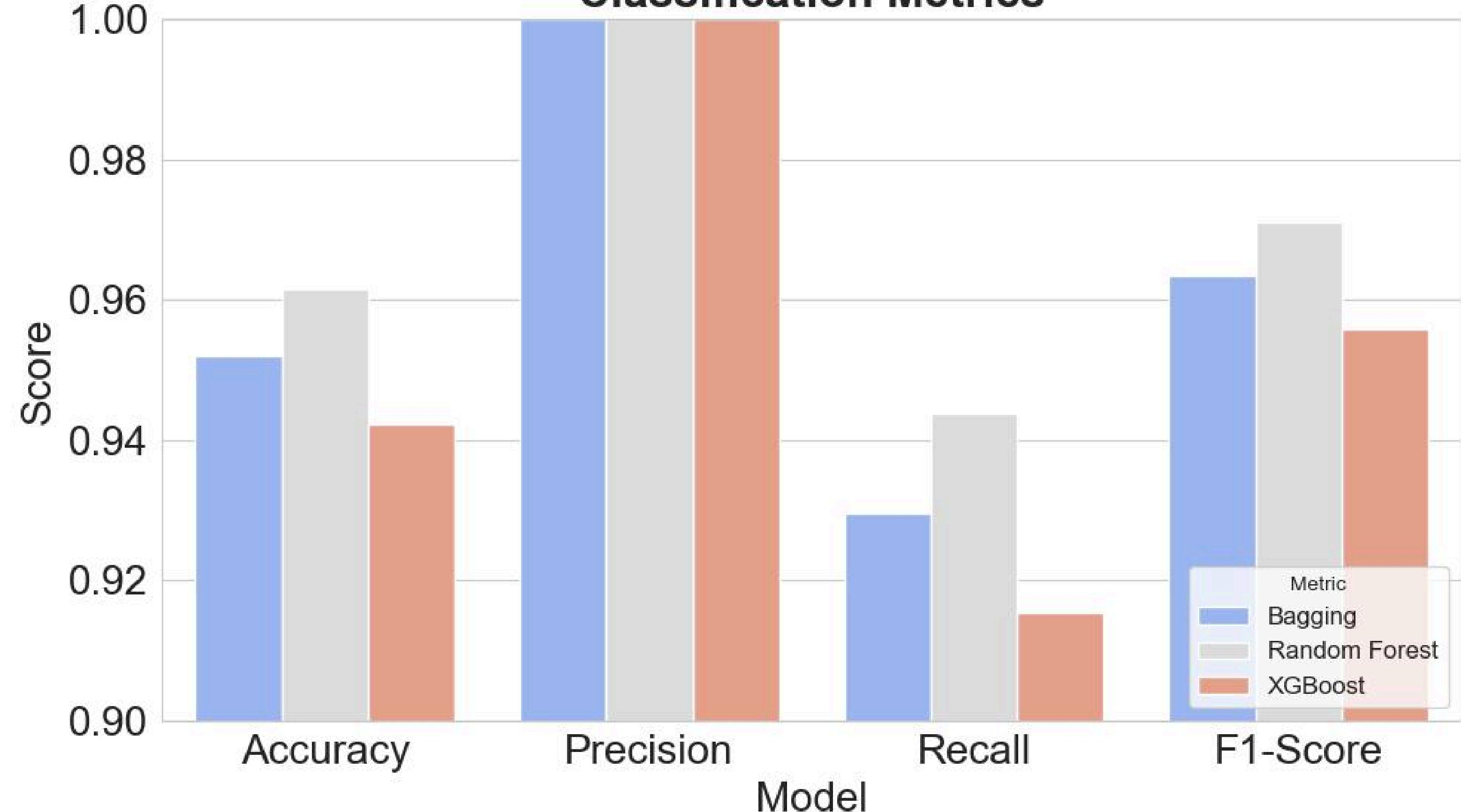
VALIDATION

	Bagging	Random Forest	XGBoost
5-fold Cross Validation Score	0.9568	0.9688	0.9592
Out-of-bag score	0.9615	0.9688	-

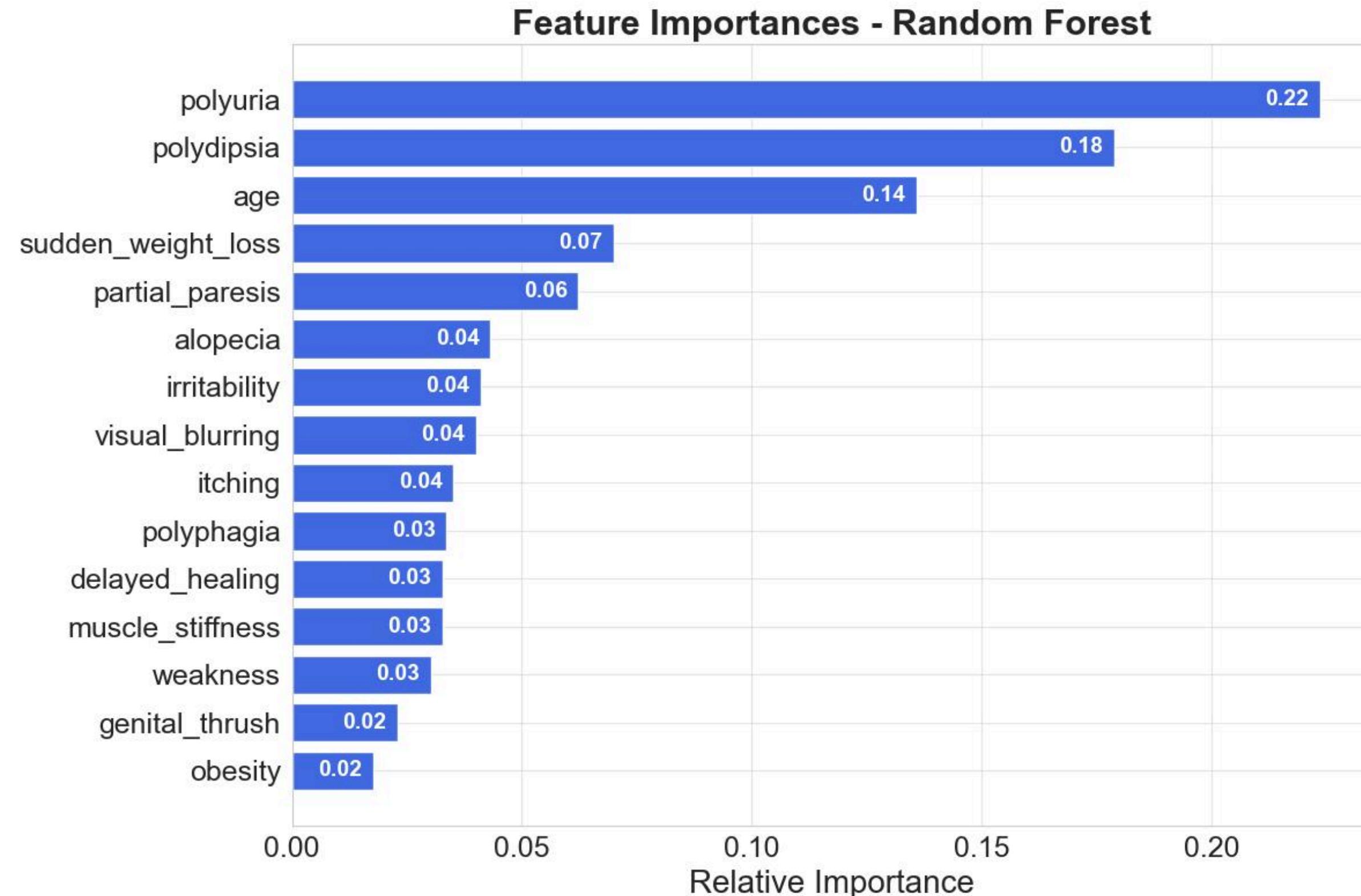
CONFUSION MATRIX



Classification Metrics



FEATURE IMPORTANCE



SHAPLEY ADDITIVE VALUES (SHAP)

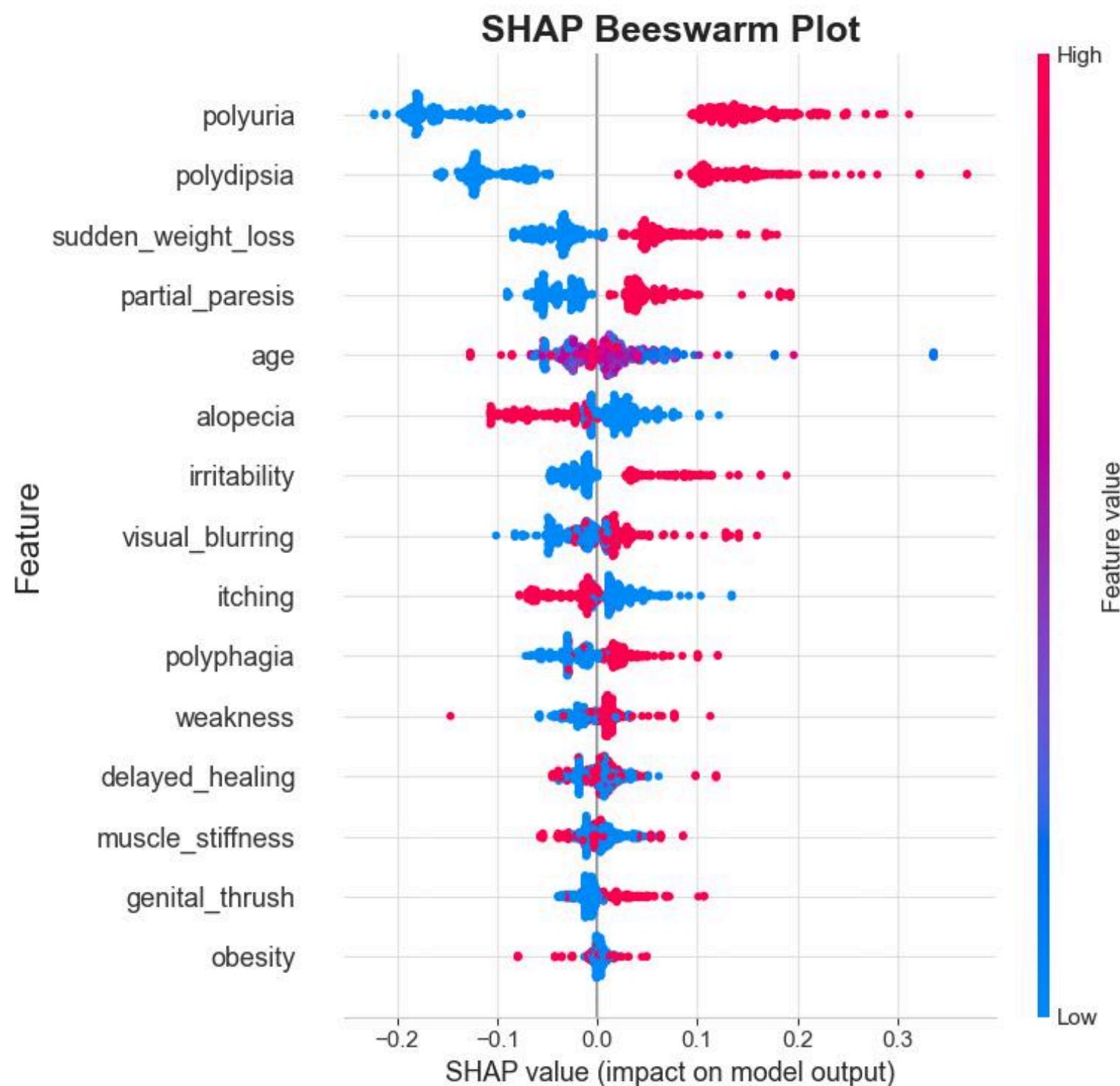
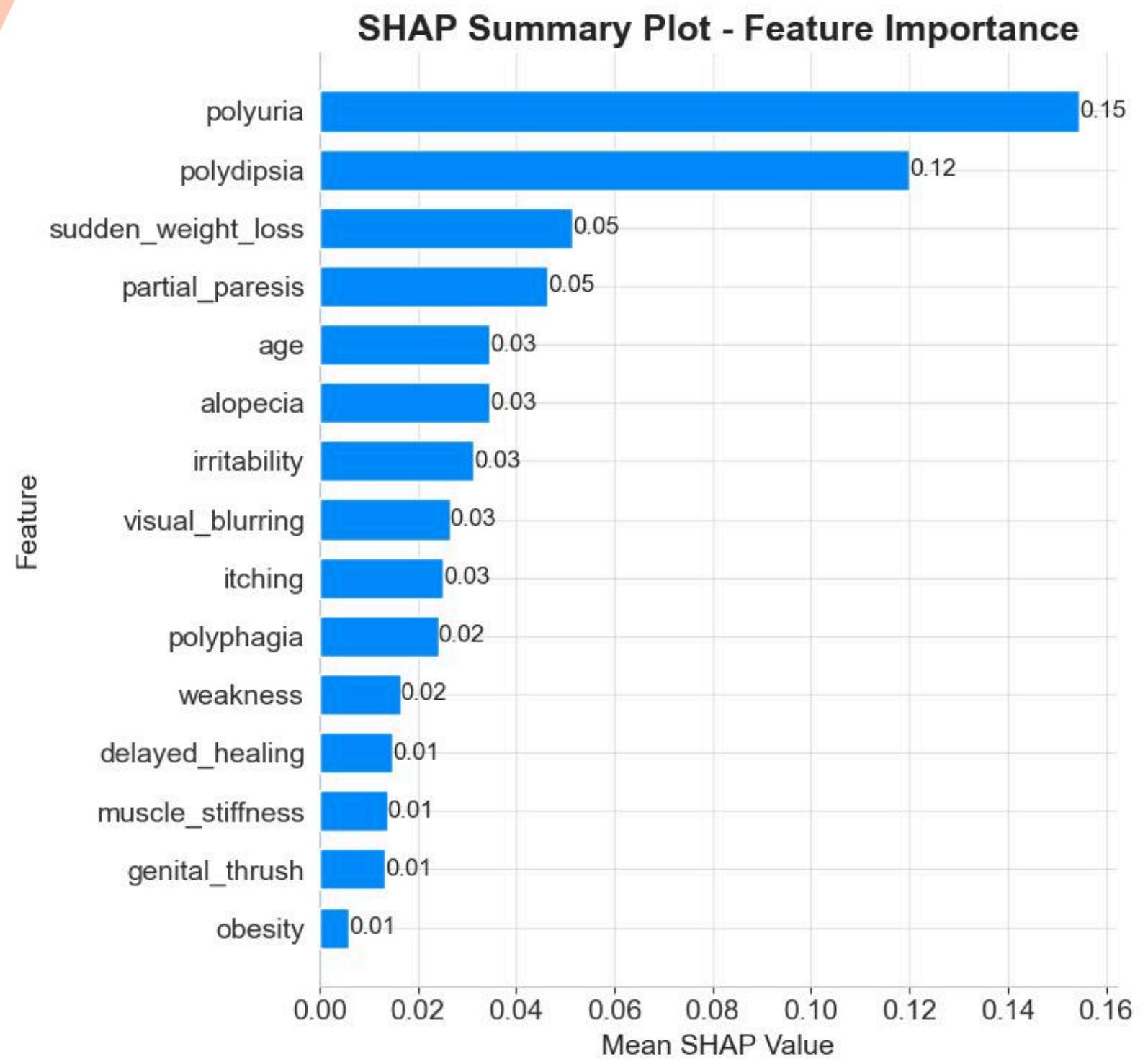
- an approach to **explain prediction output** generated by machine learning models
- Shapley value:

$$\phi_j = \sum_{\substack{S \subseteq F \setminus \{j\} \\ \text{all features except } j}} \frac{|S|! (|F| - |S| - 1)!}{|F|!} [f(S \cup \{j\}) - f(S)]$$

marginal contribution of feature j
weight

- Positive SHAP value = positive association with diabetes

SHAP SUMMARY PLOTS



SUMMARY

- Compared three ensemble machine learning techniques
- Random Forest has highest accuracy
- Used SHAP to explain random forest predictions

LIMITATIONS

- Basis for model parameters
- Cross-validation parameters

FUTURE WORK

- Increase number of features
- Hypertune parameters of models



THANK YOU!

Reference:

- Vijayan V.V., Anjali C. Prediction and diagnosis of diabetes mellitus—A machine learning approach; Proceedings of the 2015 IEEE Recent Advances in Intelligent Computational Systems (RAICS); Kerala, India. 10–12 December 2015; pp. 122–127