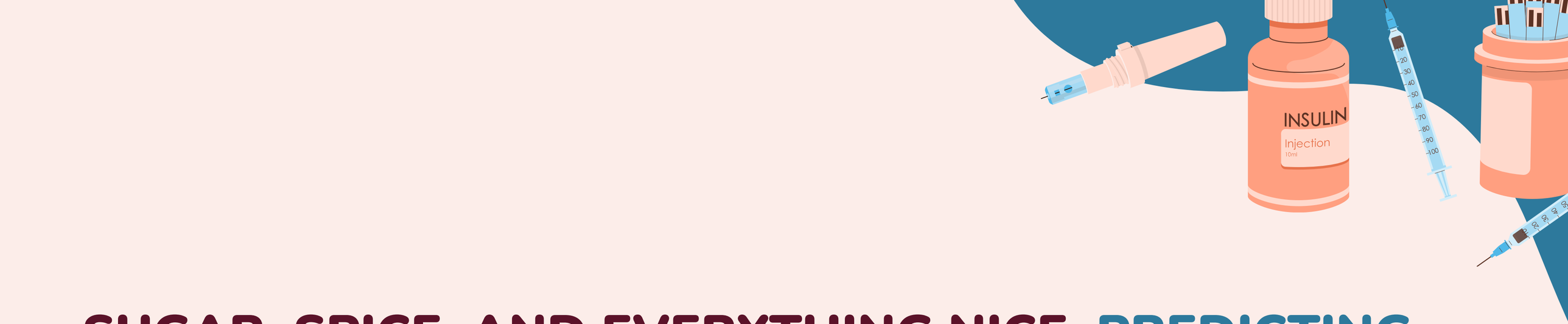


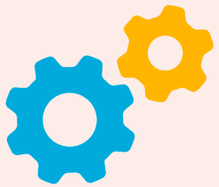
SUGAR, SPICE, AND EVERYTHING NICE: PREDICTING DIABETES RISK USING ENSEMBLE TECHNIQUES AND EXPLAINABLE ARTIFICIAL INTELLIGENCE

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 (Vijavan et al. 2015)

Franco: Predicting diabetes risk using ensemble techniques and explainable artificial intelligence



MACHINE LEARNING

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

Franco: Predicting diabetes risk using ensemble techniques and explainable artificial intelligence

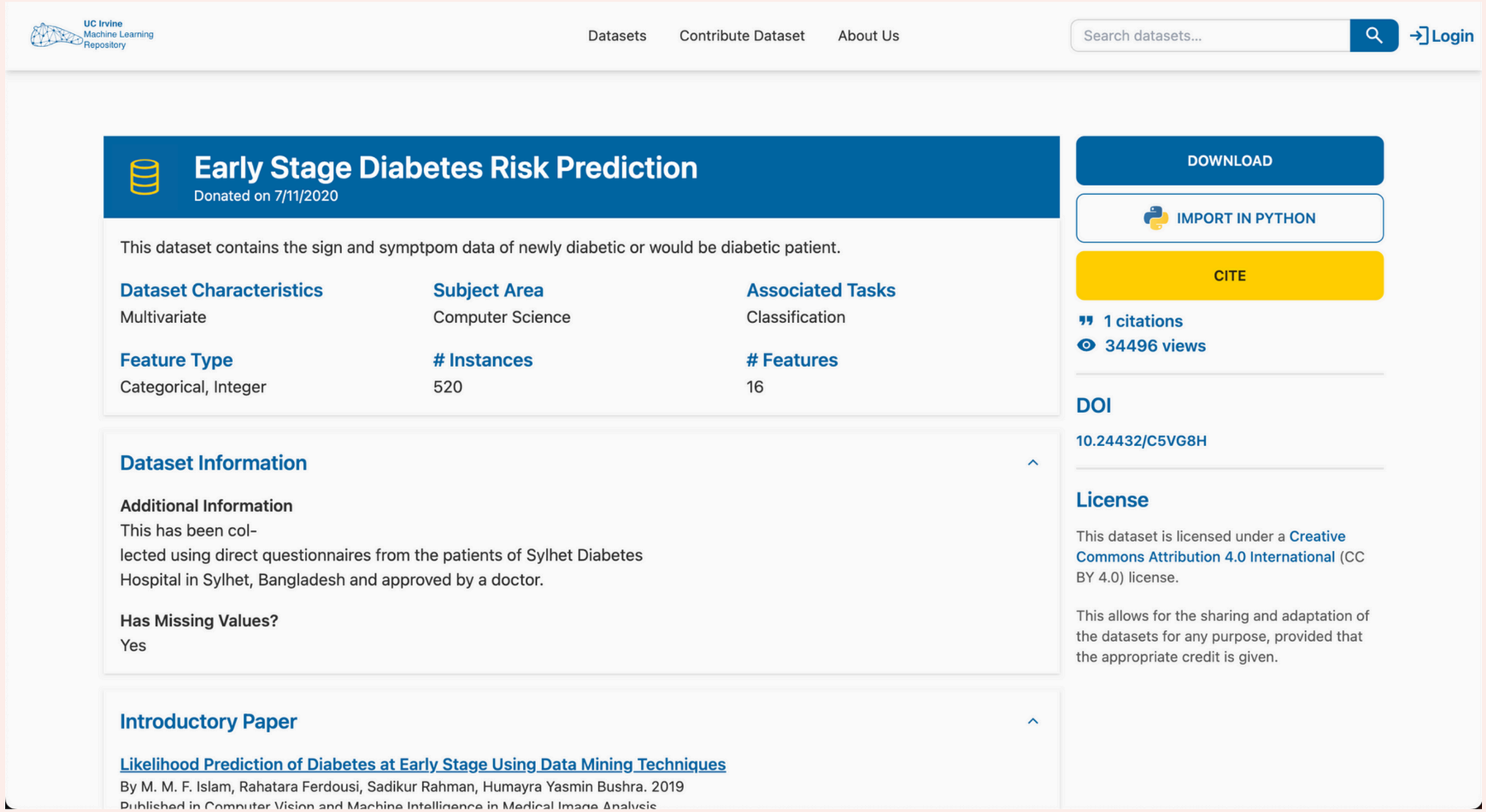
OBJECTIVES

- 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



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

Variable Name
age
gender
polyuria
polydipsia
sudden_weight_loss
weakness
polyphagia
genital_thrush
visual_blurring
itching

Variable Name
irritability
delayed_healing
partial_paresis
muscle_stiffness
alopecia
obesity

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

Franco: Predicting diabetes risk using ensemble techniques and explainable artificial intelligence

PRE-PROCESSING

Original dataset

	age	polyuria	polydipsia	sudden_weight_loss	weakness	polyphagia	genital_thrush	visual_blurring	itching	irritability	delayed_healing	
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	

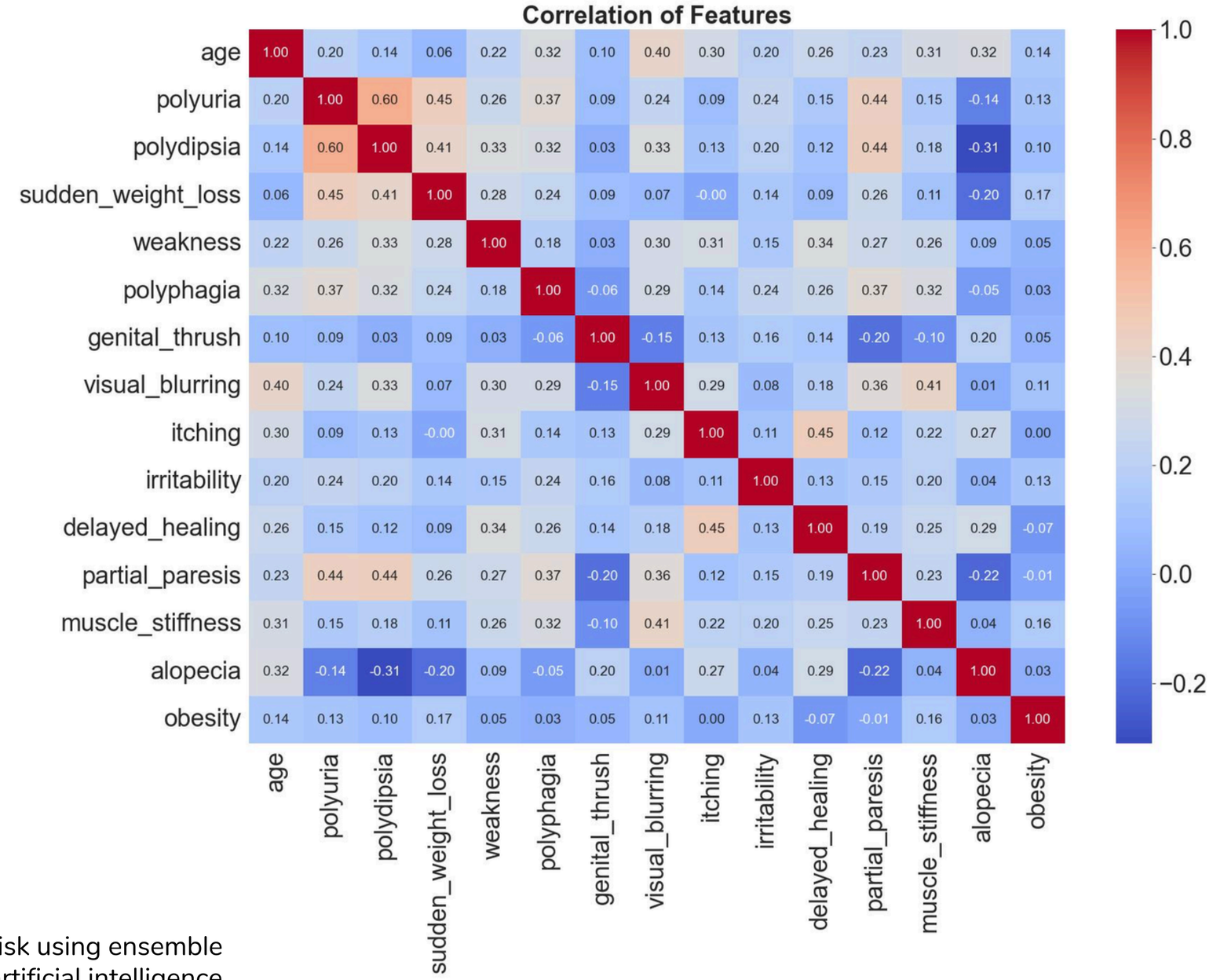
520 rows × 16 columns

Dataset with 0 or 1 for binary features

	age	polyuria	polydipsia	sudden_weight_loss	weakness	polyphagia	genital_thrush	visual_blurring	itching	irritability	delayed_healing
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



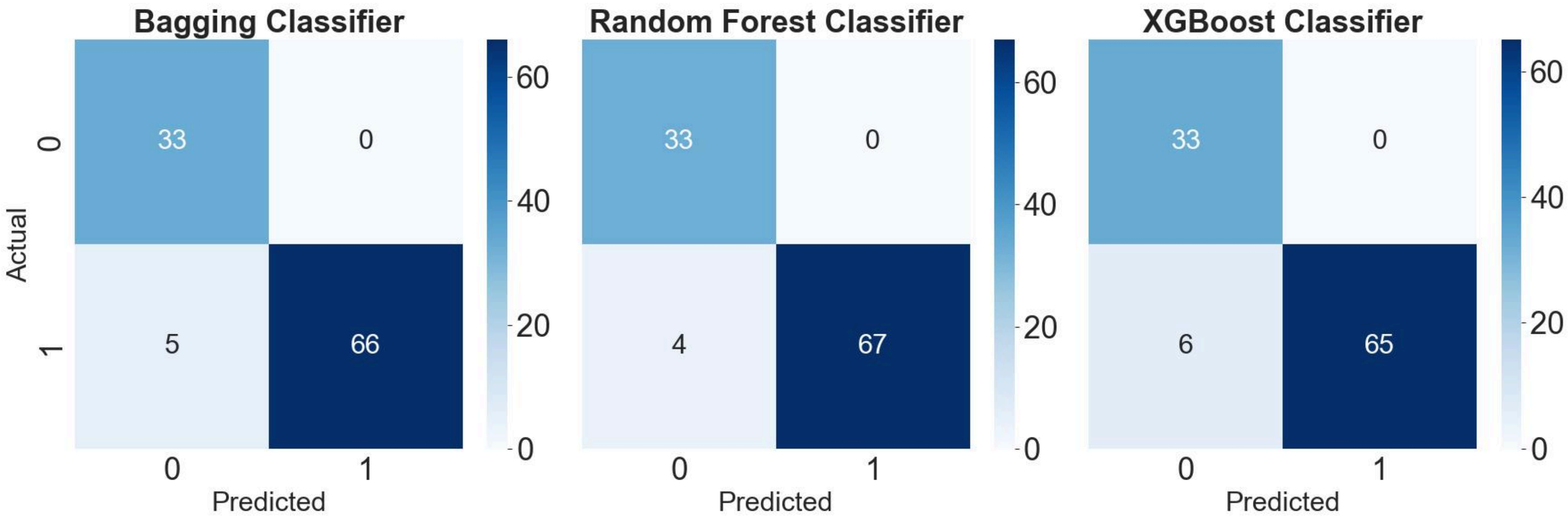
MODEL TRAINING

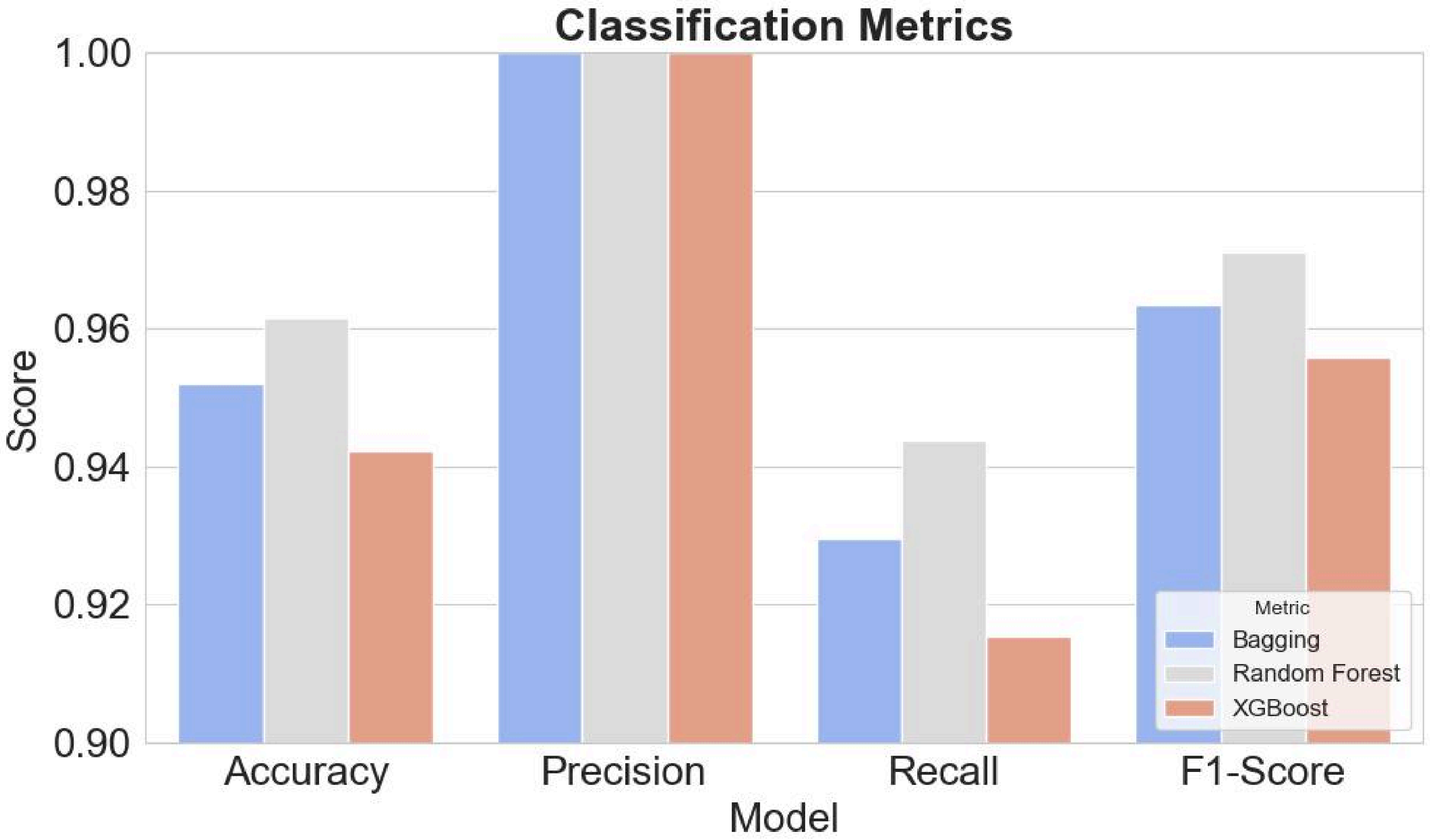
- 80% Training, 20% Testing
- BaggingClassifier, RandomForestClassifier, XGBClassifier
  - default parameters
- number of trees: 1000

VALIDATION

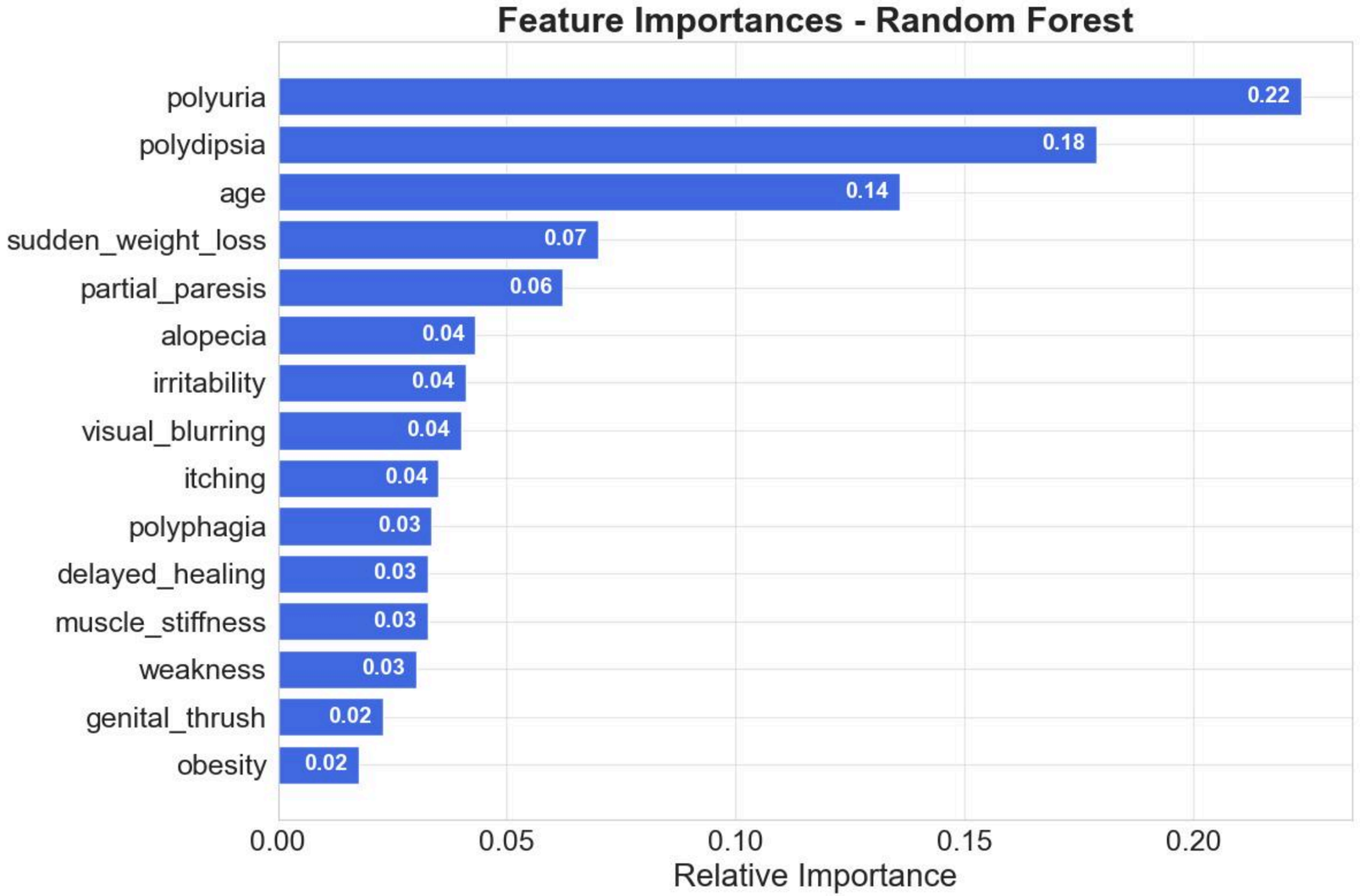
	Bagging	Random Forest	XGBoost
5-fold Cross Validation Score	0.9568	0.9688	0.9592
Out-of-bag score	0.9615	0.9688	-

CONFUSION MATRIX




Classification Metrics
1.00
0.98
0.96
0.94
0.92
0.90
Score
Accuracy
Precision
Recall
F1-Score
Model
Metric
Bagging
Random Forest
XGBoost

FEATURE IMPORTANCE



SHAPLEY ADDITIVE VALUES (SHAP)

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

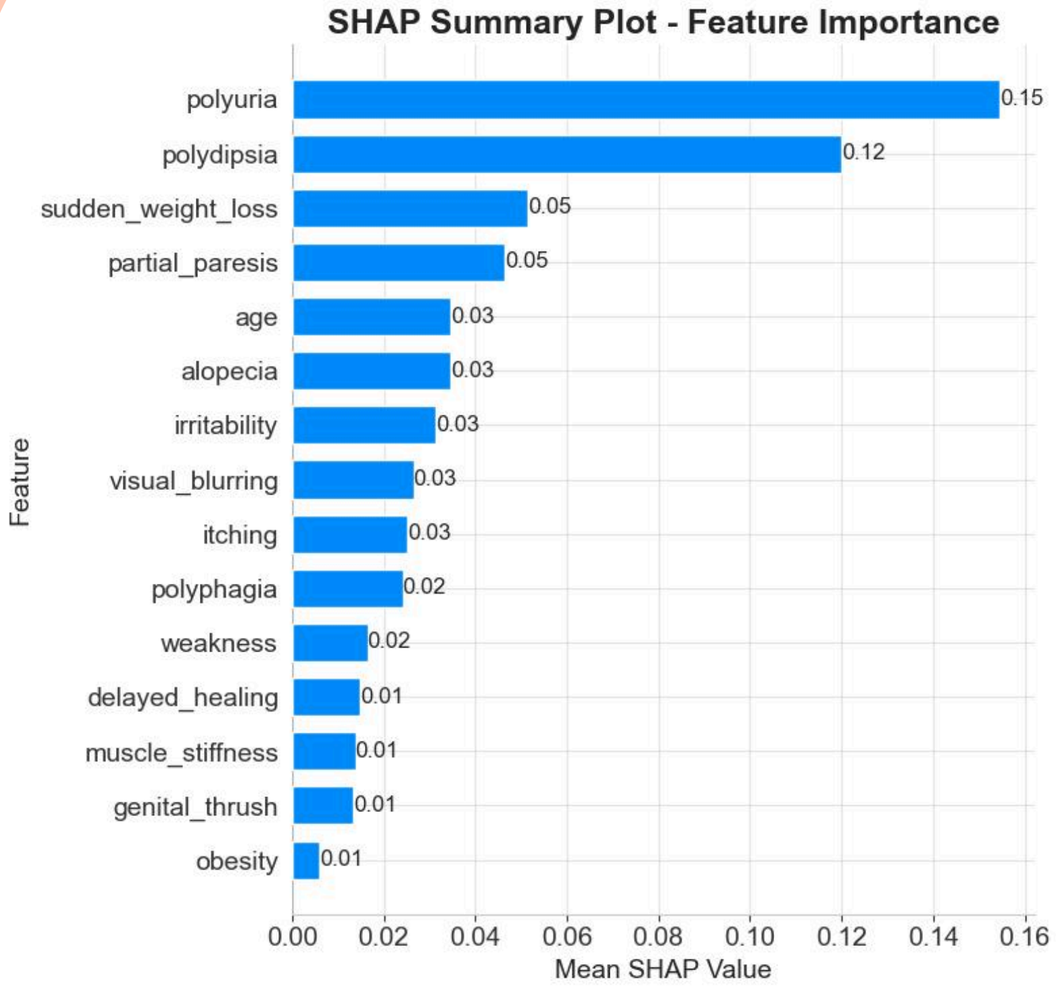
$$\emptyset_j = \sum_{S \subseteq F \setminus \{j\}} \frac{|S|!\,(|F| - |S| - 1)!}{|F|!} [f(S \cup \{j\}) - f(S)]$$

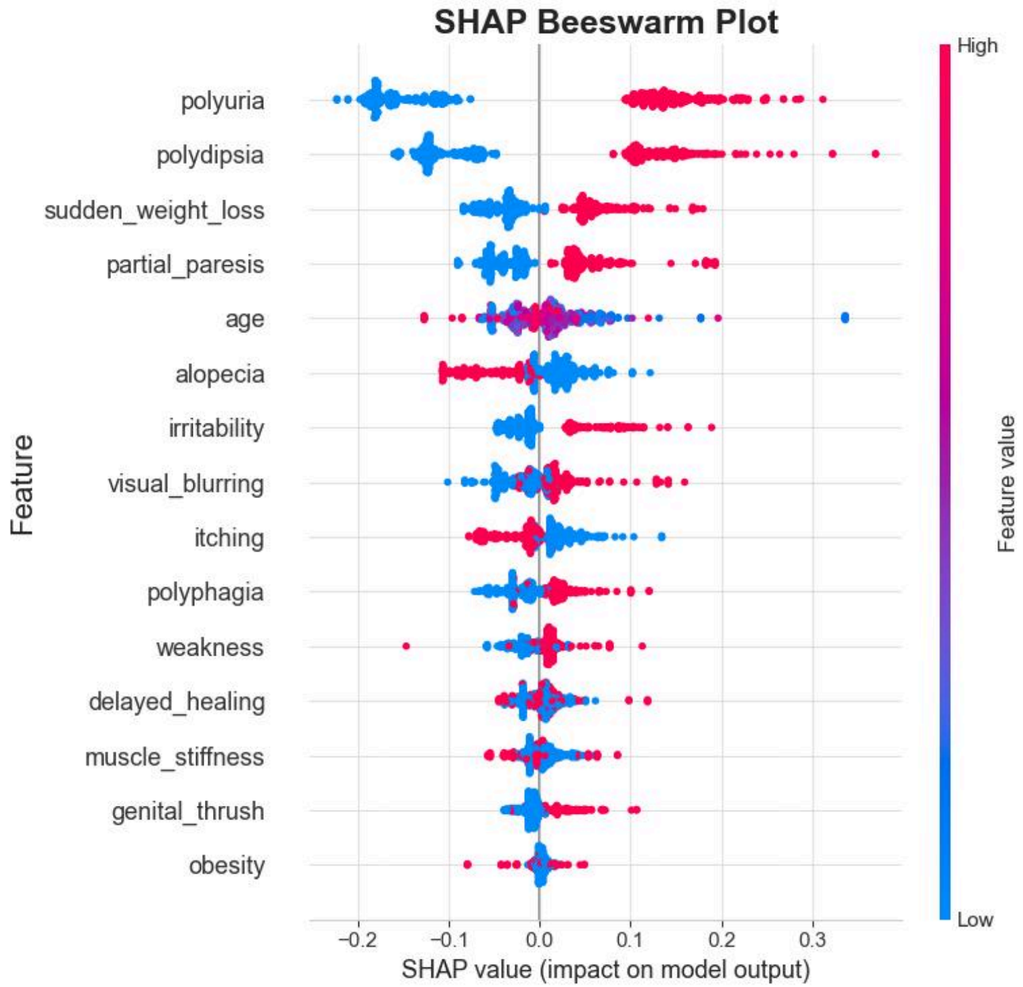
all features except j — weight — marginal contribution of feature j

- Positive SHAP value = positive association with diabetes

Franco: Predicting diabetes risk using ensemble techniques and explainable artificial intelligence

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