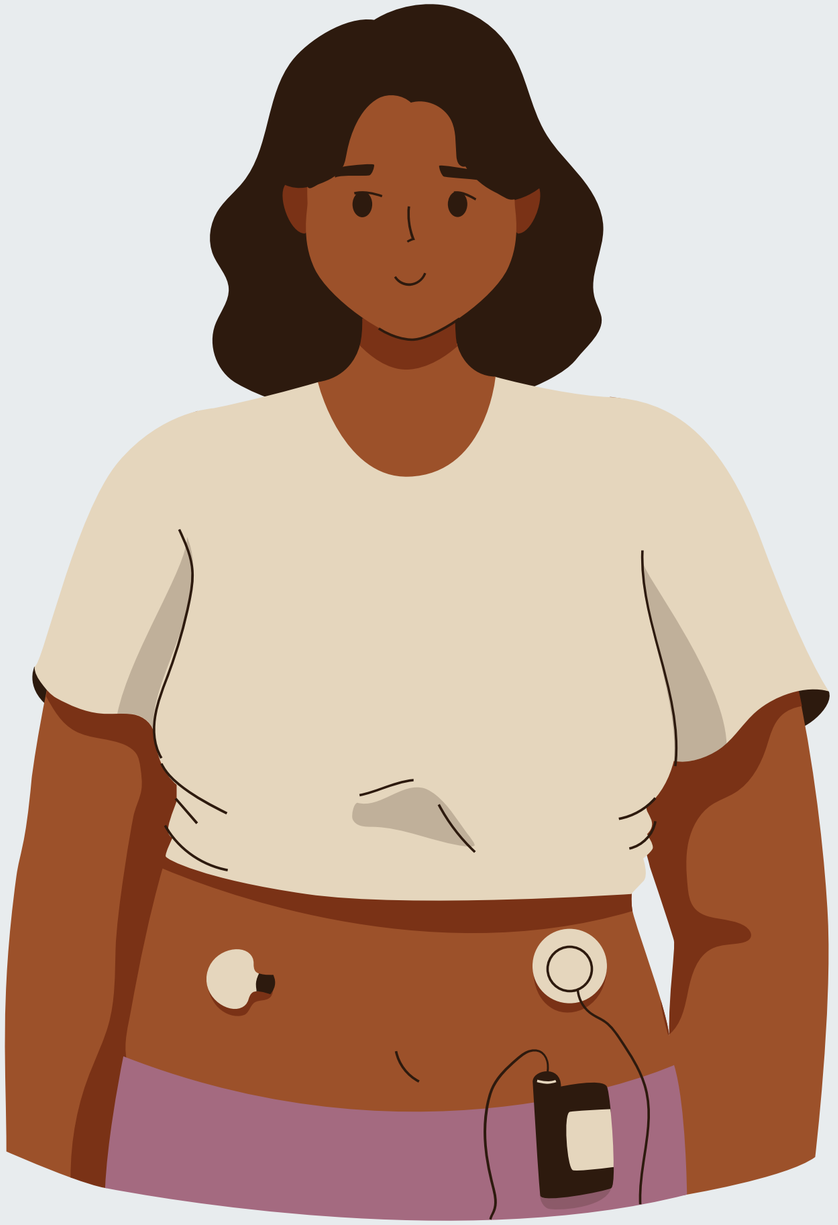
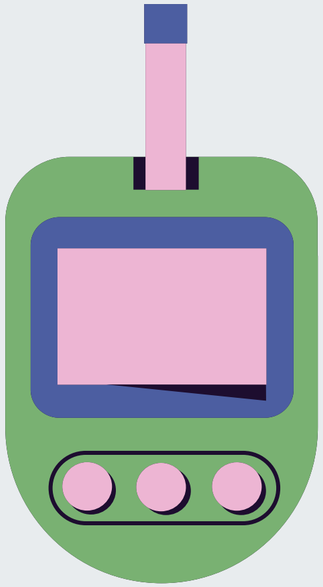
PRESENTED BY:
LESLIE LI, SATYAKI DIXIT, AND SHRESHTA PHOGAT

BUILDING RISK PREDICTION MODELS FOR DIABETES USING MACHINE LEARNING



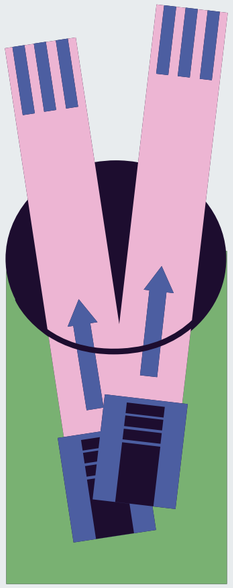
12th December 2022 | Phase 3 Evaluation and Interpretation

Introduction: Research Questions



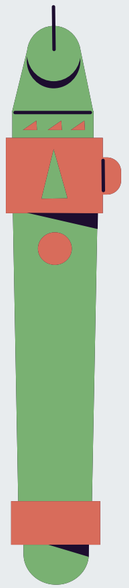
Aim 1

What risk factors are most predictive of diabetes risk?



Aim 2

What is the association among different variables?



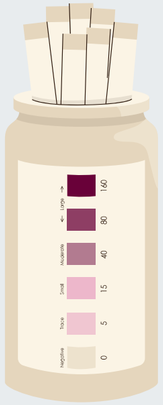
Aim3

Which ML models contribute to a more accurate prediction?

&

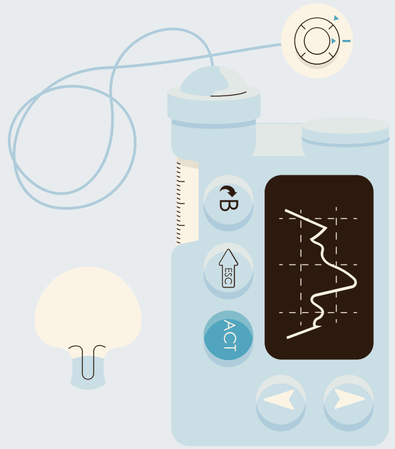
What are the optimal validation metrics to measure model performance?

Introduction: Data Source and Methodology



Data Source

- The Behavioral Risk Factor Surveillance System's survey responses in 2015.
- Health-related telephone surveys collecting state data about U.S. residents regarding their health-related risk behaviors, chronic health conditions, and use of preventive services.



Methodology

- Select essential risk factors for analysis after literature review.
- EDA with dichotomy and transformation.
- Use multivariable weighted logistic regression models to measure associations among factors.
- Apply supervised ML models and metrics.

Challenges

Issue 1

There is considerable heterogeneity in previous studies regarding machine learning techniques used, making it challenging to identify the optimal one.

Issue 2

There is a lack of transparency about the features used to train the models, which reduces their interpretability, a feature utterly relevant to the doctor.

DATA SUMMARY

Shape

- 330 features (columns)
  - 323 numerical features
  - 7 categorical features
  - 244 columns have missing values
- 441,456 survey responses (rows)
- Not balanced with a size at 541.28 MB

A Glimpse of Attributes

- High BP
- High cholesterol, cholesterol check
- BMI
- Smoke history, stoke history
- Coronary heart disease (CHD) or myocardial infarction
- Physical activity in past 30 days
- Fruit, vegetables, drinks consumption habit
- Health care coverage, doctor visit frequency, health scale
- Mental health
- Sex, age, education, income level
- Sleep/disordered breathing

Process Summary

1. Data Cleaning

- Readable format with 41 columns and 315853 rows.
- binary data set with the target variable having diabetes.
- Feature correlation with having diabetes.

2. EDA

Strong correlations occur among features, including Consume Fruit, Consume Vegetables, Physical Activity Categories, Aerobic Recommendations, and Muscle Strengthening Recommendations.

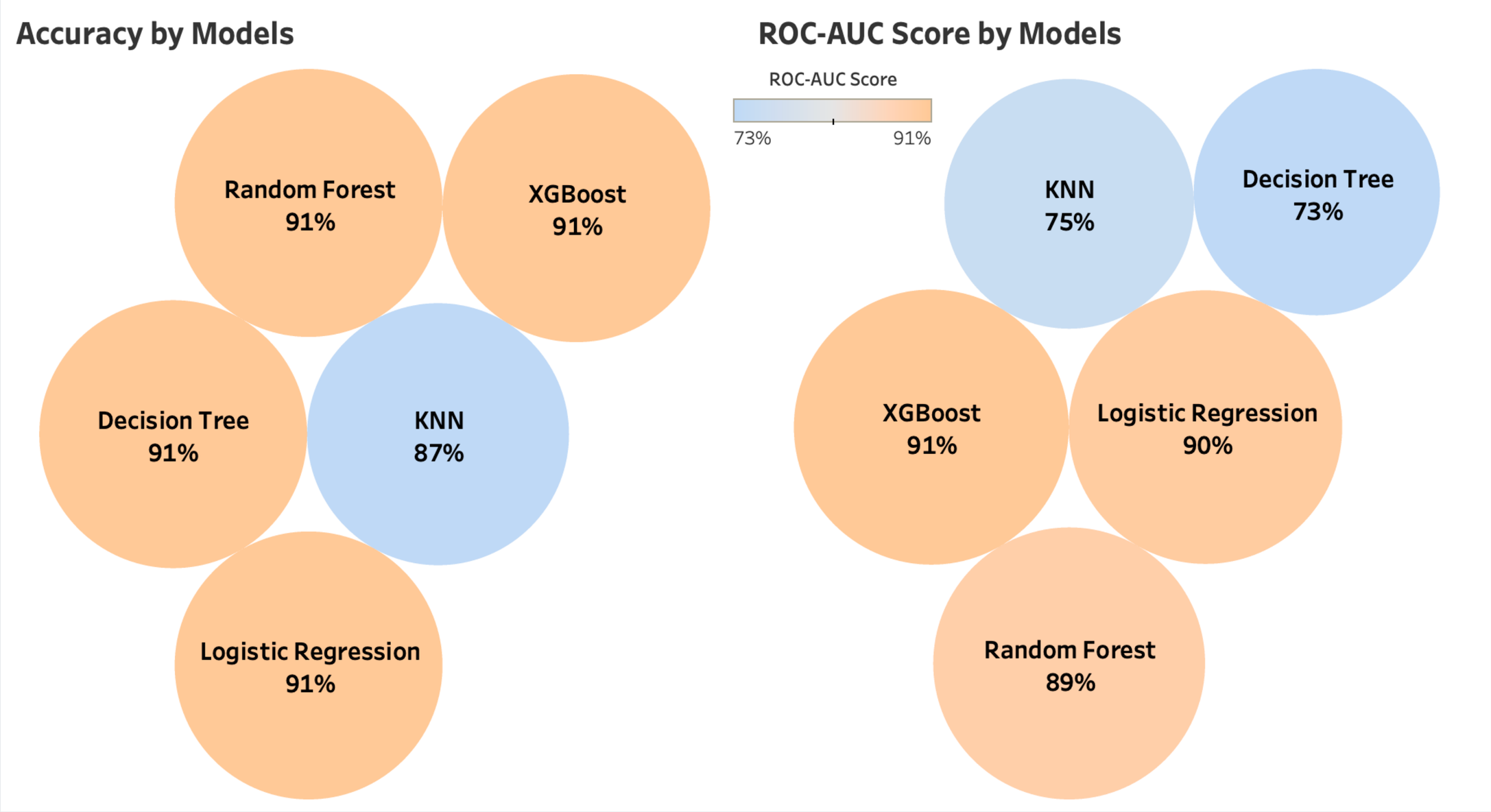
3. Model Building

Logistic Regression, KNN, Decision Tree Classifier, Random Forest Classifier, XGBoost Classifier

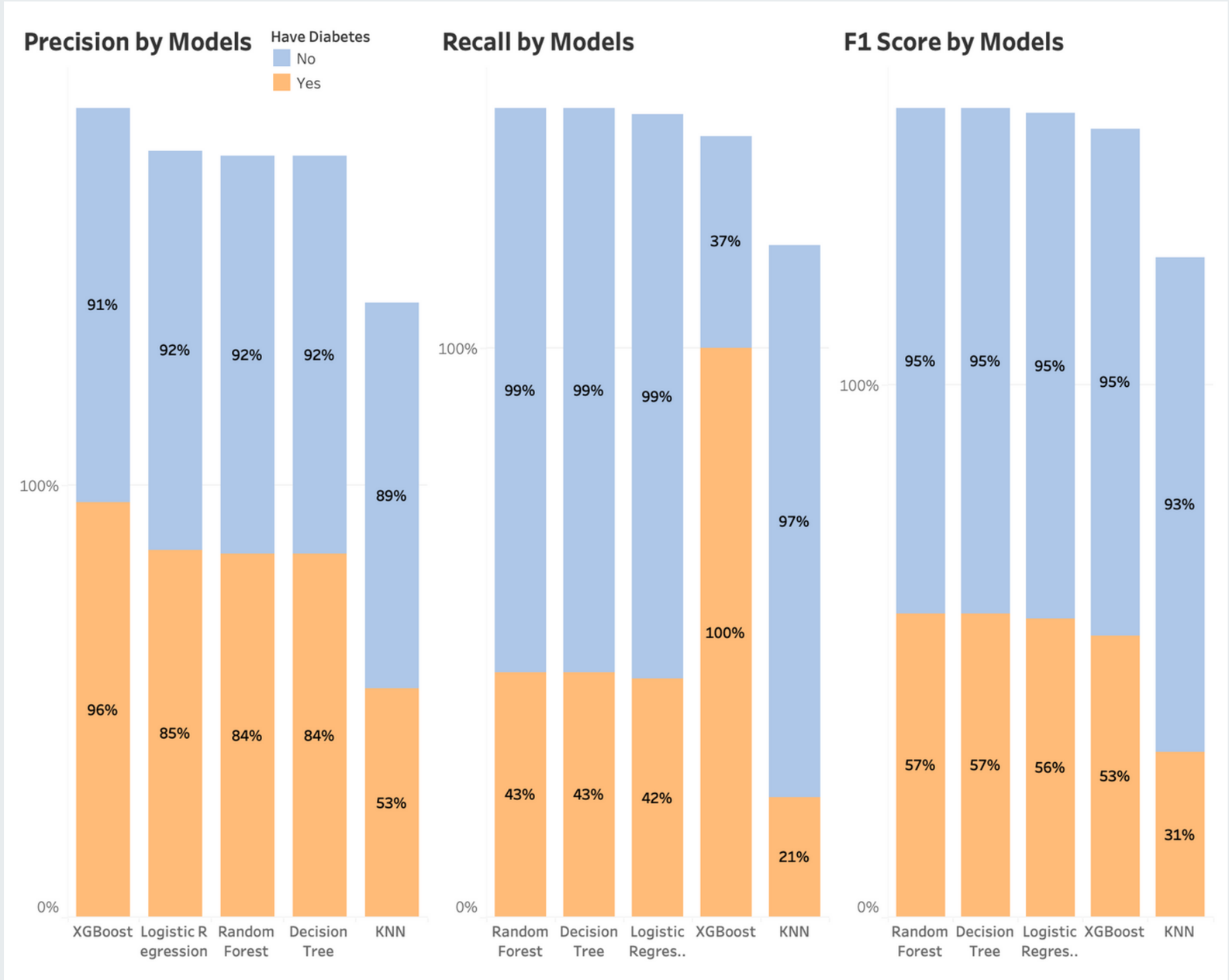
4. Outcomes

Model evaluation metrics, SHAP explanation and threshold tuning.

MODEL EVALUATION



MODEL EVALUATION 2

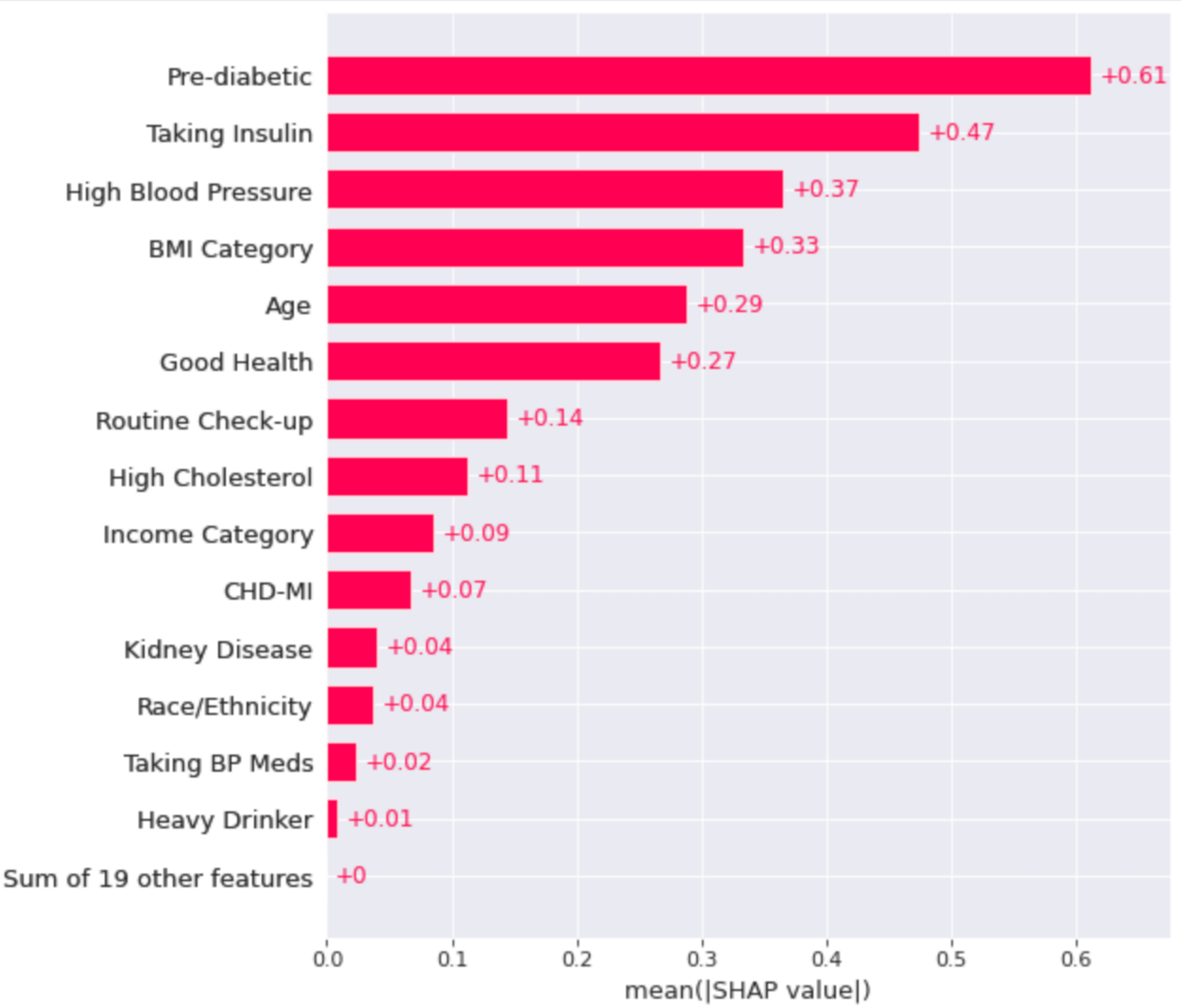


- XGBoost outperformed other models in precision, predicting most observations correctly from the total as high precision related to the low false positive rate.
- Tree-based models, including Random Forest and Decision Tree, exceeded others in recall and F1 score, followed by Logistic Regression with PCA elements. XGBoost showed an unusual trend capturing most diabetes classes than others.
- KNN lagged behind all other models in three metrics.

SHAP EXPLANATION - XGBOOST

- **SHAP (SHapley Additive exPlanations):** A game theoretic approach to explain the output of any machine learning model.
- The above explanation shows how these features contribute to the model output from the base value (the average model output over the training dataset we passed) to the model output. Features pushing the prediction higher are shown in red; conversely, those pushing the prediction lower are in blue.
- *Routine check-ups*, *high cholesterol*, *pre-diabetic condition*, and *age* pushed the prediction higher, and *BMI category*, *no blood pressure*, *taking insulin*, *good health*, and *no Coronary Heart Disease (CHD-MI)* pushed the prediction lower.

SHAP EXPLANATION - XGBOOST

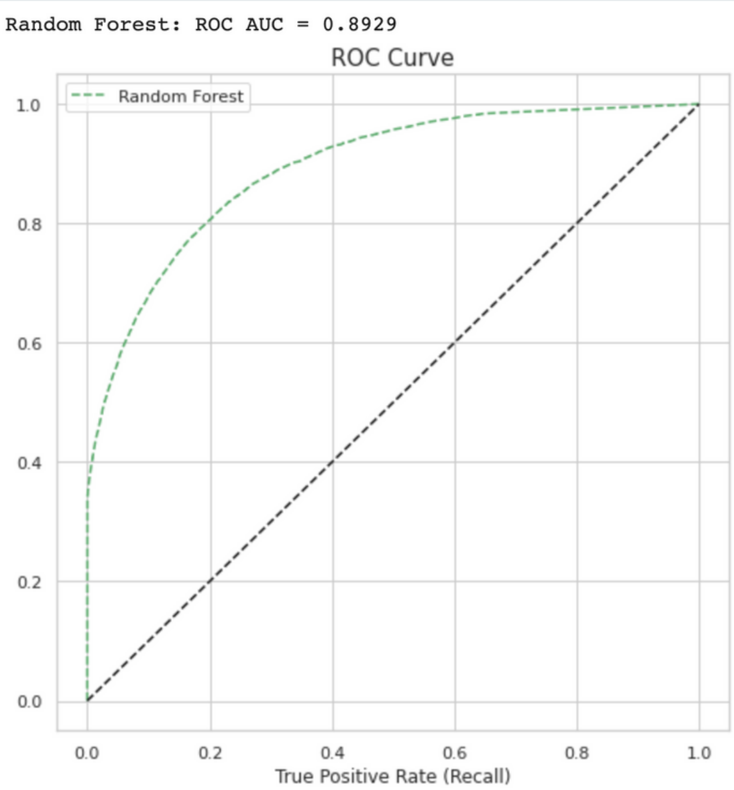


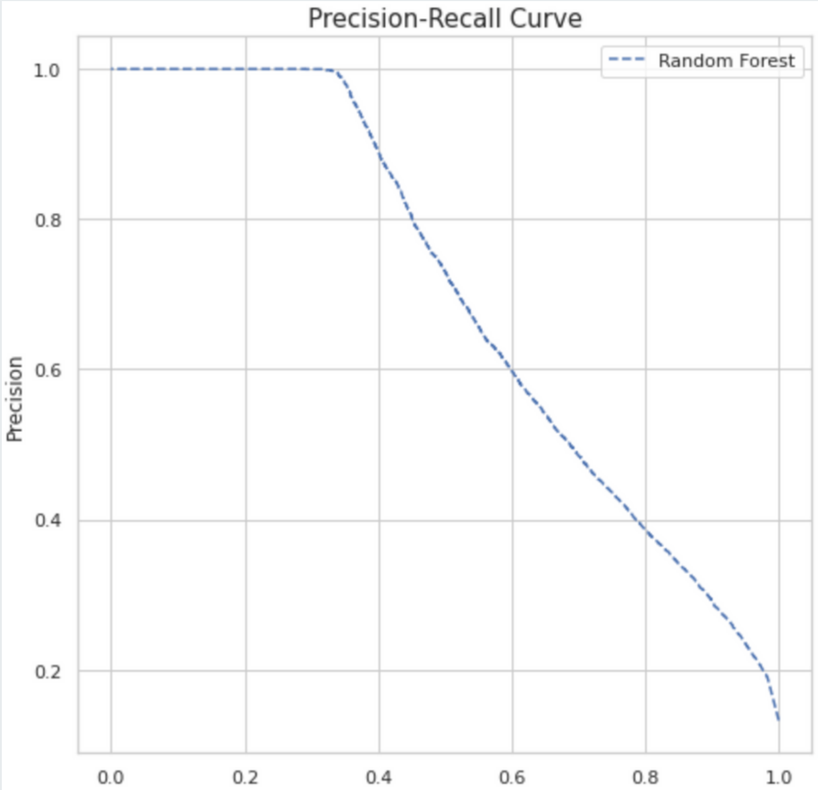
Variable Importance

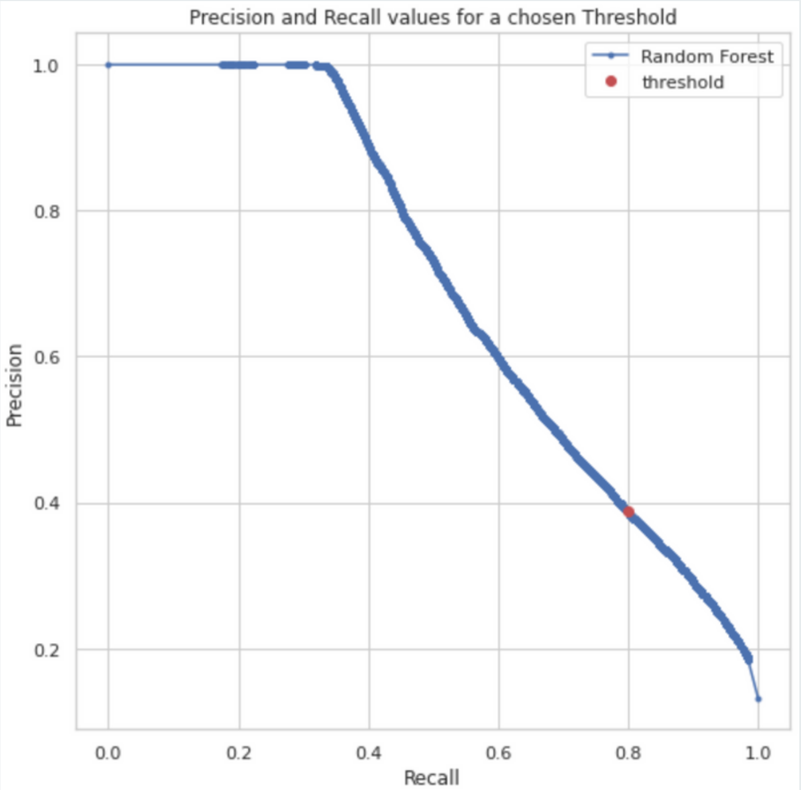
- On the left are the input variables, ranked from top to bottom by their mean absolute SHAP values for the entire datasets–i.e., the average magnitude of each variable's impact on the predicted target, having diabetes or not, across all instances.
- The mean absolute SHAP values are, on average, how much each variable impacts the precited result in the positive or negative direction.

THRESHOLD TUNING - RANDOM FOREST

Random Forest: ROC AUC = 0.8929





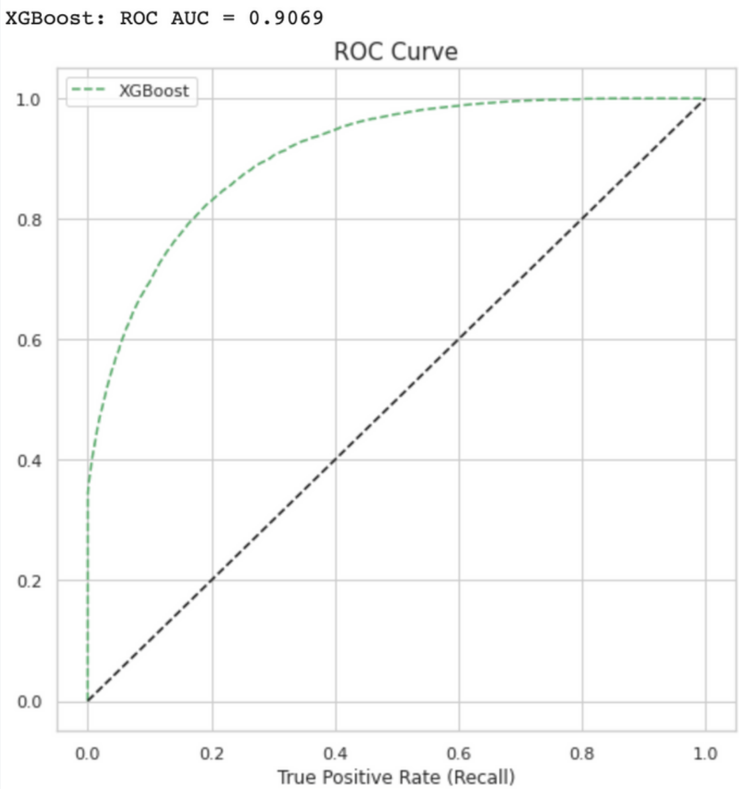


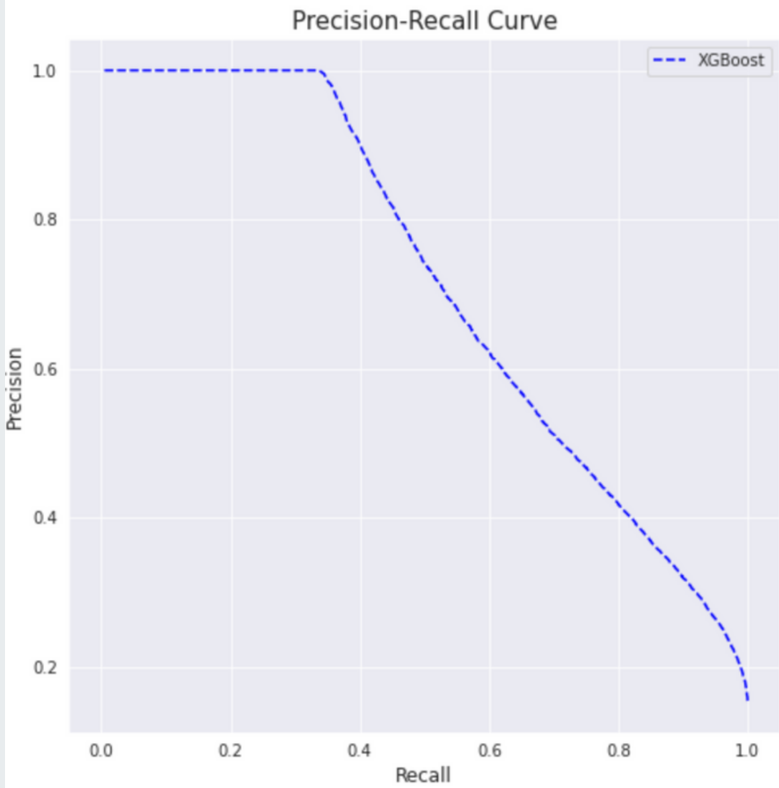
Threshold value = 0.1450

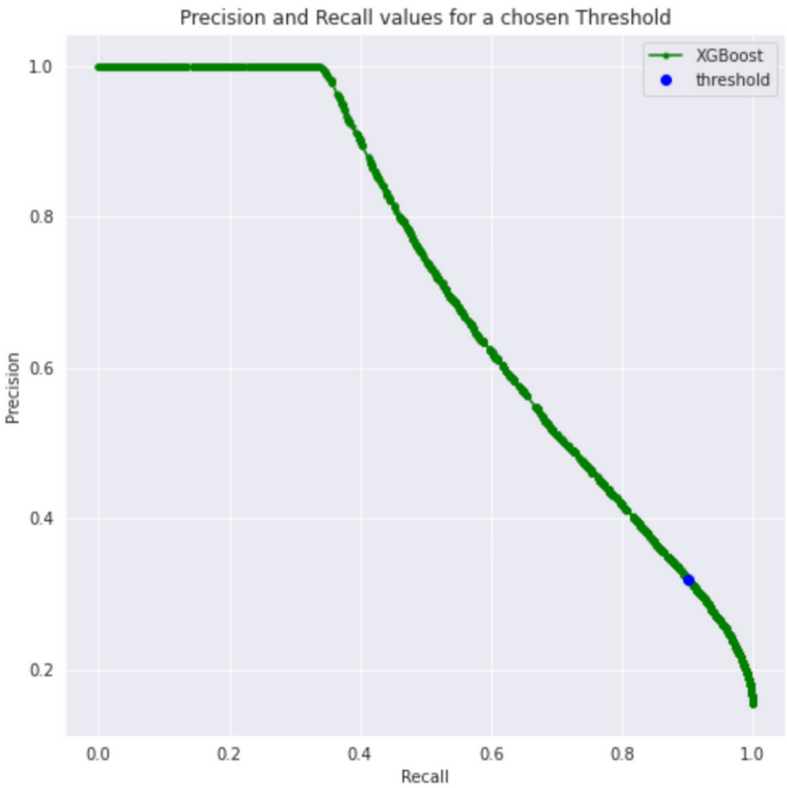
	Model Name	Precision	Recall	F1 Score	Specificity	Accuracy
0	Random Forest, optimized t	0.387895	0.798854	0.522219	0.807319	0.806197

THRESHOLD TUNING - XGBOOST

XGBoost: ROC AUC = 0.9069



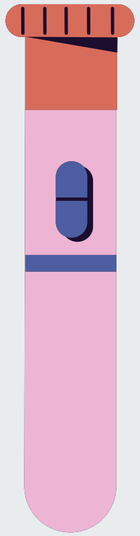




Threshold value = 0.0935

	Model Name	Precision	Recall	F1 Score	Specificity	Accuracy
0	XGBoost, optimized t	0.325564	0.868344	0.473573	0.72505	0.744048

OUTCOME SUMMARY



What risk factors are most predictive of diabetes risk?

Pre-diabetic condition, taking insulin, high blood pressure, BMI Category, age, good health evaluation, routine check-up, high cholesterol, kidney disease, and Coronary Heart Disease (CHD-MI).

What is the association among different variables?

Strong correlations occur among features, including consuming fruit & vegetables, physical activity categories, aerobic recommendations, and muscle strengthening recommendations.

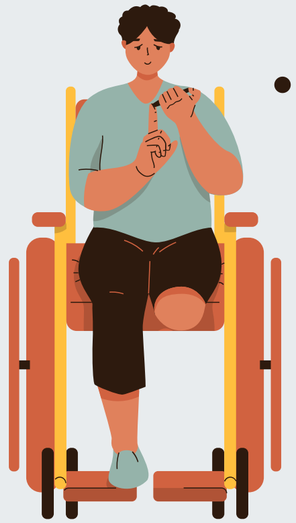
Which ML models with metrics are optimal overall?

- XGBoost Classifier achieved the highest accuracy, AUC score, and precision; a cutoff at 0.0935 led to an 87% recall.
- Random Forest led the highest recall and F1 score; a cutoff at 0.145 resulted in an 80% recall.

LIMITATIONS & FUTURE STEPS

Data Source

- An individual may have difficulty remembering events that occurred long ago or the frequency of certain behaviors. Some respondents may overreport socially desirable behaviors while underreporting behaviors they perceive to be less acceptable.
- BRFSS methodology precludes anyone from assisting respondents in completing the interview if the selected adult had difficulty participating for any reason, such as an intellectual or developmental disability.



Methodology

- Model development and validation were conducted throughout this project with only one data source. Thus, it's necessary to collect additional cases and verify the model derived in this project using other data sources.
- In future work, we can build and implement a web application for the proposed diabetic monitoring system with a proposed classification and predicting approach.
- Genetic algorithms can also be explored with the proposed prediction mechanisms for better monitoring.

REFERENCES

1. Bryanb. “XGBoost Explainability with Shap.” Kaggle, Kaggle, 11 Nov. 2020, https://www.kaggle.com/code/bryanb/xgboost-explainability-with-shap/notebook#-5.-SHapley-Additive-exPlanations.
2. Dansbecker. “Advanced Uses of Shap Values.” Kaggle, Kaggle, 9 Nov. 2021, https://www.kaggle.com/code/dansbecker/advanced-uses-of-shap-values.
3. Aidan Cooper. “Explaining Machine Learning Models: A Non-Technical Guide to Interpreting Shap Analyses.” Aidan Cooper, Aidan Cooper, 13 May 2022, https://www.aidancooper.co.uk/a-non-technical-guide-to-interpreting-shap-analyses/#:~:text=Bar%20plots,towards%20the%20predicted%20house%20prices.
4. Lama, Lara, et al. “Machine Learning for Prediction of Diabetes Risk in Middle-Aged Swedish People.” Heliyon, U.S. National Library of Medicine, 25 June 2021, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8282976/.
5. Butt, Umair Muneer, et al. “Machine Learning Based Diabetes Classification and Prediction for Healthcare Applications.” Journal of Healthcare Engineering, Hindawi, 1 Oct. 2021, https://www.hindawi.com/journals/jhe/2021/9930985/.
6. SHAP Documentation
7. Behavioral Risk Factor Surveillance System (BRFSS) Data.

THANK YOU!