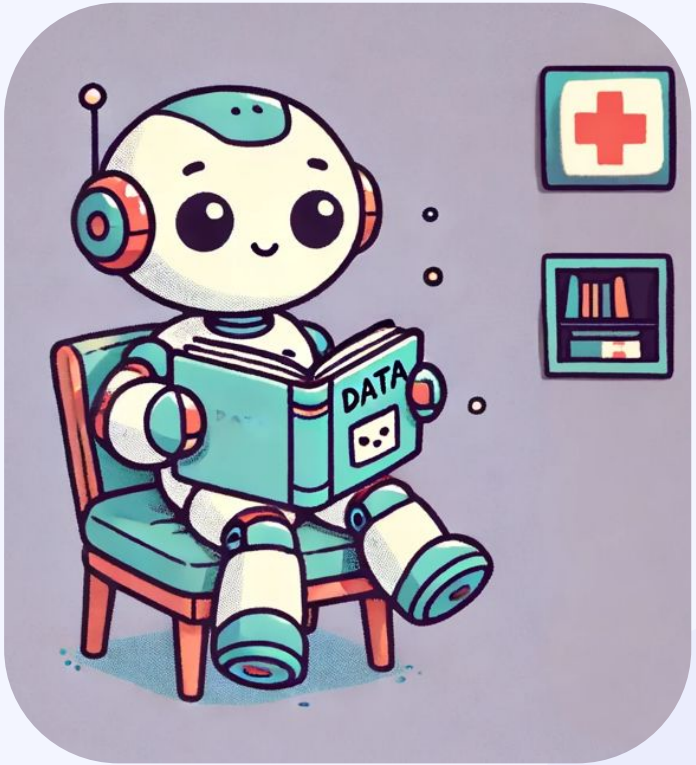
Classifiers: CDC Diabetes Health Indicators

Katherine Granados
Franklin Ledesma
Rosemary Medina-Casanova
Eros Reyes

Objective

The purpose of this mission is to develop a predictive model using machine learning techniques on the data provided by the CDC Diabetes Health Indicators survey. This project focuses on its application in the medical field, with the specific goal of identifying the most significant risk factors for predicting diabetes risk.



Data: CDC Diabetes Health Indicators

Dataset Characteristics	Associated Tasks	# Instances
Tabular, Multivariate	Classification	253,680

Subject Area	Feature Type	Features
Health and Medicine	Categorical, Integer	21

The Diabetes Health Indicators Dataset offers comprehensive healthcare data, including details on individuals' diabetes diagnosis status, alongside general health metrics like blood pressure, cholesterol levels, BMI, physical activity, and diet. Researchers leverage this data to discern patterns and correlations for better understanding and management of diabetes.

Data Selection & Prep

The dataset contains extensive health-related information collected from surveys.

Data Cleaning and Preprocessing:

- Checked for Null Values
- Normalization: Applied One Hot Encoding to normalize numerical variables like general health, mental health and physical health.

GenHlth	MentHlth	PhysHlth
5.0	18.0	15.0
3.0	0.0	0.0
5.0	30.0	30.0
2.0	0.0	0.0
2.0	3.0	0.0

GenHlth	MentHlth	PhysHlth
0.25	0.033333	0.000000
0.50	0.333333	0.000000
0.00	0.000000	0.000000
0.50	0.000000	0.333333
0.50	0.000000	0.000000

Feature Engineering & Selection

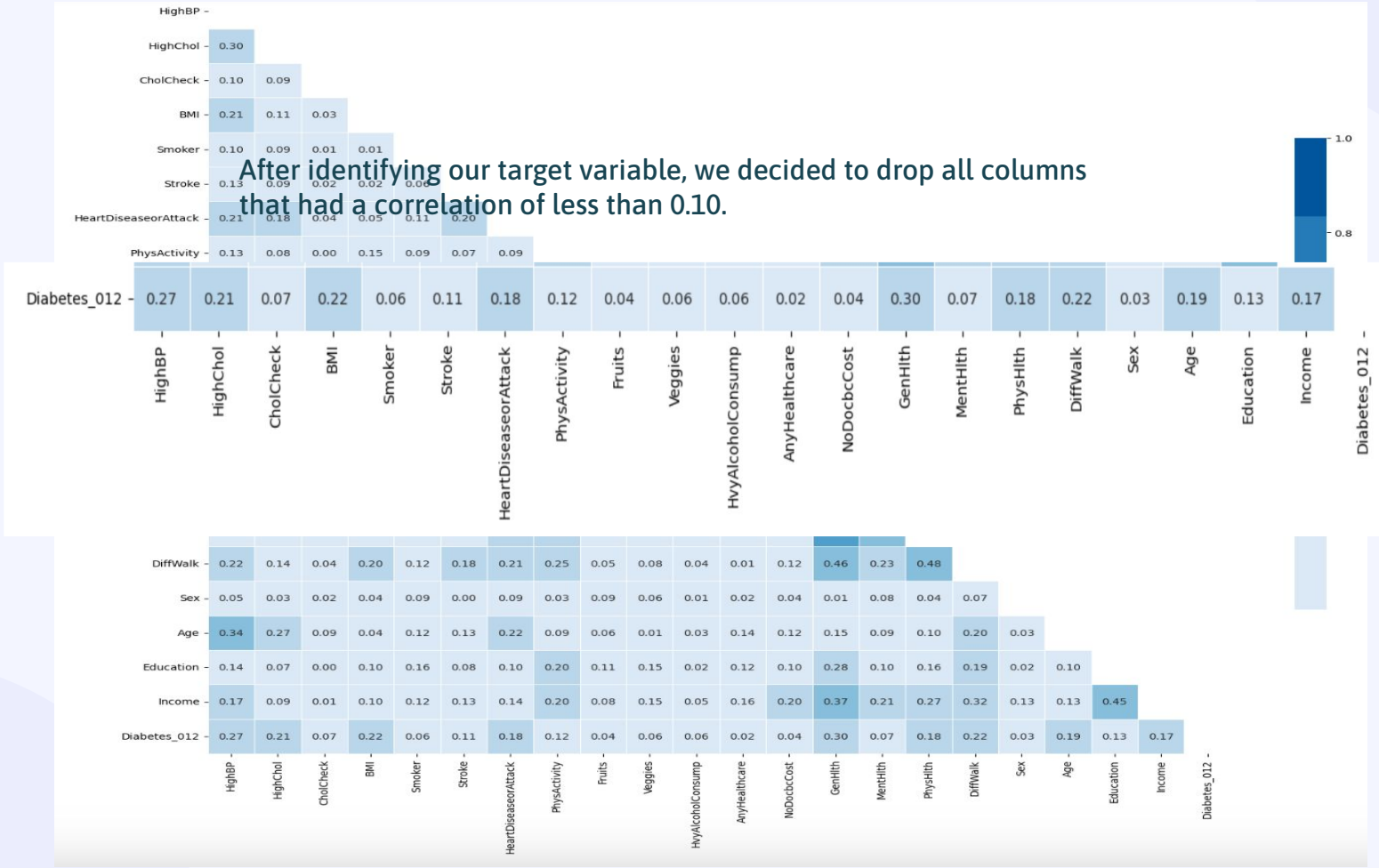
Feature Engineering:

- Feature Selection: Employed correlation analysis to select features most predictive of diabetes risk.

Challenges:

- Addressed repetitive features and ensured that selected features were informative and non-redundant. This helped us identify which features might not be useful for detecting diabetes.
- Effective feature engineering enhanced model accuracy

After identifying our target variable, we decided to drop all columns that had a correlation of less than 0.10.



Model Building & Evaluation

Models Used:

- Explored K-Nearest Neighbors (KNN), Random Forest, Gradient Boosting, Adaptive Boosting, and Bagging Classifier.
- Rationale: Bagging Classifier was selected due to its superior accuracy compared to other models, achieving an accuracy score of 84%.

Evaluation Metrics:

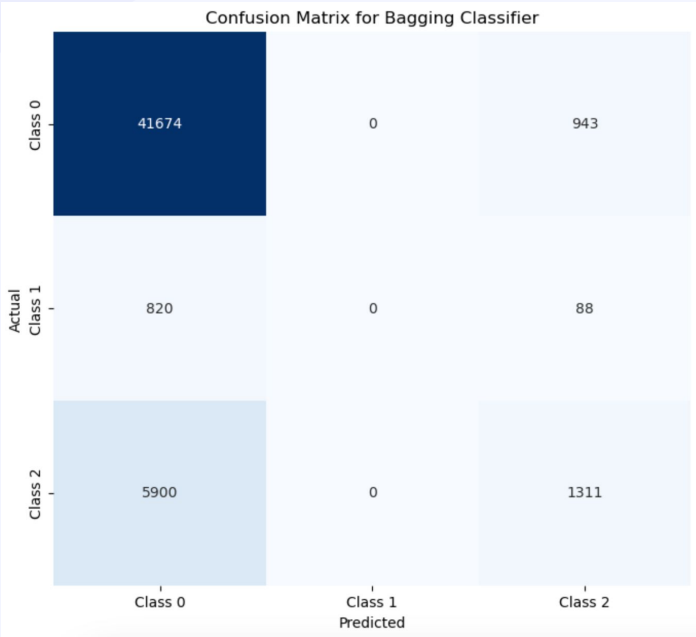
- Evaluated model performance using accuracy, precision, and recall metrics.
- Conducted cross-validation to assess model robustness and generalizability.

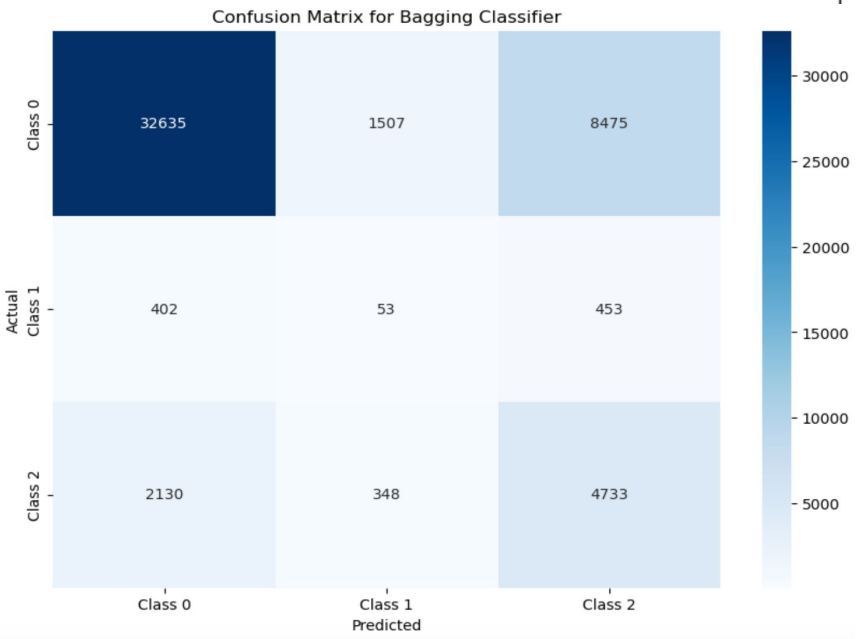
```
bagging_class = BaggingClassifier(estimator=DecisionTreeClassifier(max_depth=20),
                                  n_estimators=100,
                                  max_samples=1000)

# Train the Bagging classifier
bagging_class.fit(X_train_reduced, y_train)

# Make predictions
pred = bagging_class.predict(X_test_reduced)

# Calculate and print evaluation metrics
print("Accuracy:", accuracy_score(y_test, pred))
print("Recall:", recall_score(y_test, pred, average='weighted'))
print("Precision:", precision_score(y_test, pred, average='weighted', zero_division=0))
```

✓ 1.9s

```
Accuracy: 0.847839798170924
Recall: 0.847839798170924
Precision: 0.8040681711043033
```

Model Optimization





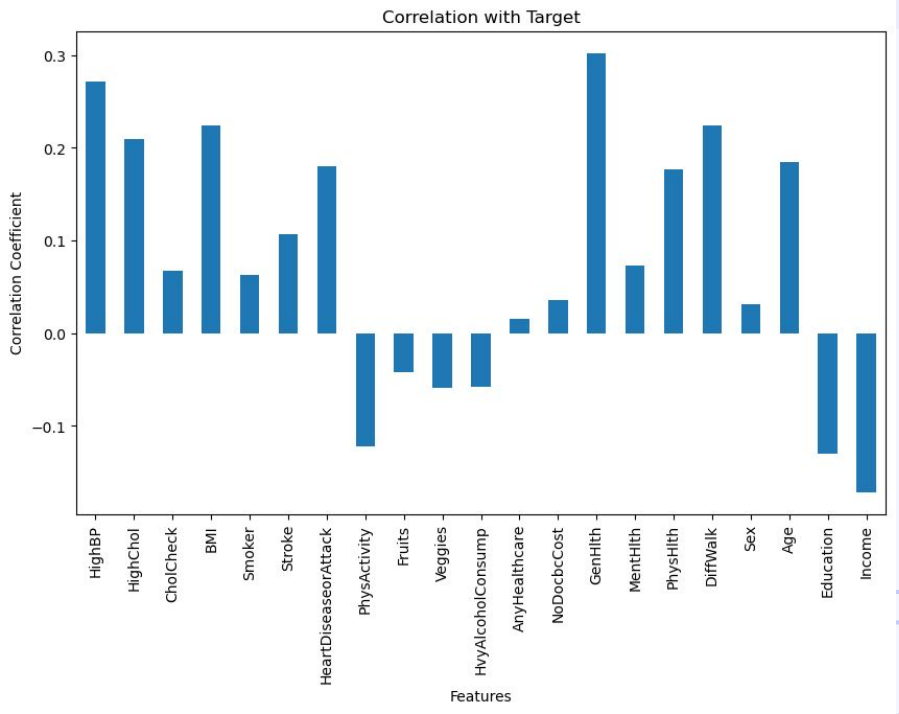
Key Findings & Insights

Major Findings

- Identified features such as BMI (Body Mass Index), high blood pressure, and Age as strong predictors of diabetes risk.

Visual Aids

- Presented charts that are used to highlight the relative contribution of each feature to the predictive accuracy of the model.



Real-World Application & Impact

The predictive model can proactively identify individuals at high risk of developing diabetes.

Application Scenario:

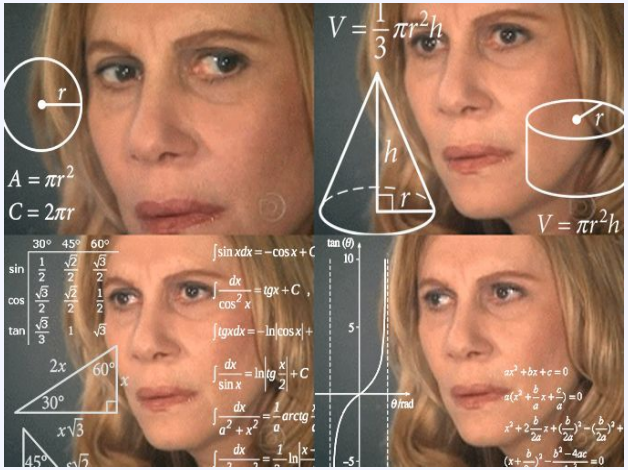
Healthcare providers can use the model to prioritize preventive actions such as lifestyle modifications and early screening tests.

Impact:

By intervening early, there is potential to mitigate the onset and progression of diabetes, leading to improved public health outcomes.

Challenges & Learnings

- **Technical Challenge**: Encountered significant runtime issues when attempting Gradient and GridSearch for evaluation metrics. The process extended beyond 200 minutes without completion, leading to the decision to abort and retain the initial classifier models that yielded the best results.
- **Challenge/Learning**: Selecting the optimal model amidst various algorithm choices & importance of feature engineering in enhancing model predictive power.
- **Performance**: Navigating our project's subject matter presented a learning curve in deciphering the optimal data formatting for achieving our objectives. Nevertheless, through collaborative effort and teamwork, we successfully navigated these challenges and powered through to attain the desired results.



Future Work & Improvements

Future Research:

- Incorporate more recent datasets to capture evolving health trends.
- Explore advanced feature engineering techniques such as deep learning for better feature extraction.

Improvements:

- Enhance model robustness through ensemble techniques.

THANK YOU!

Q&A

Classifiers: CDC Diabetes Health Indicators

Katherine Granados
Franklin Ledesma
Rosemary Medina-Casanova
Eros Reyes