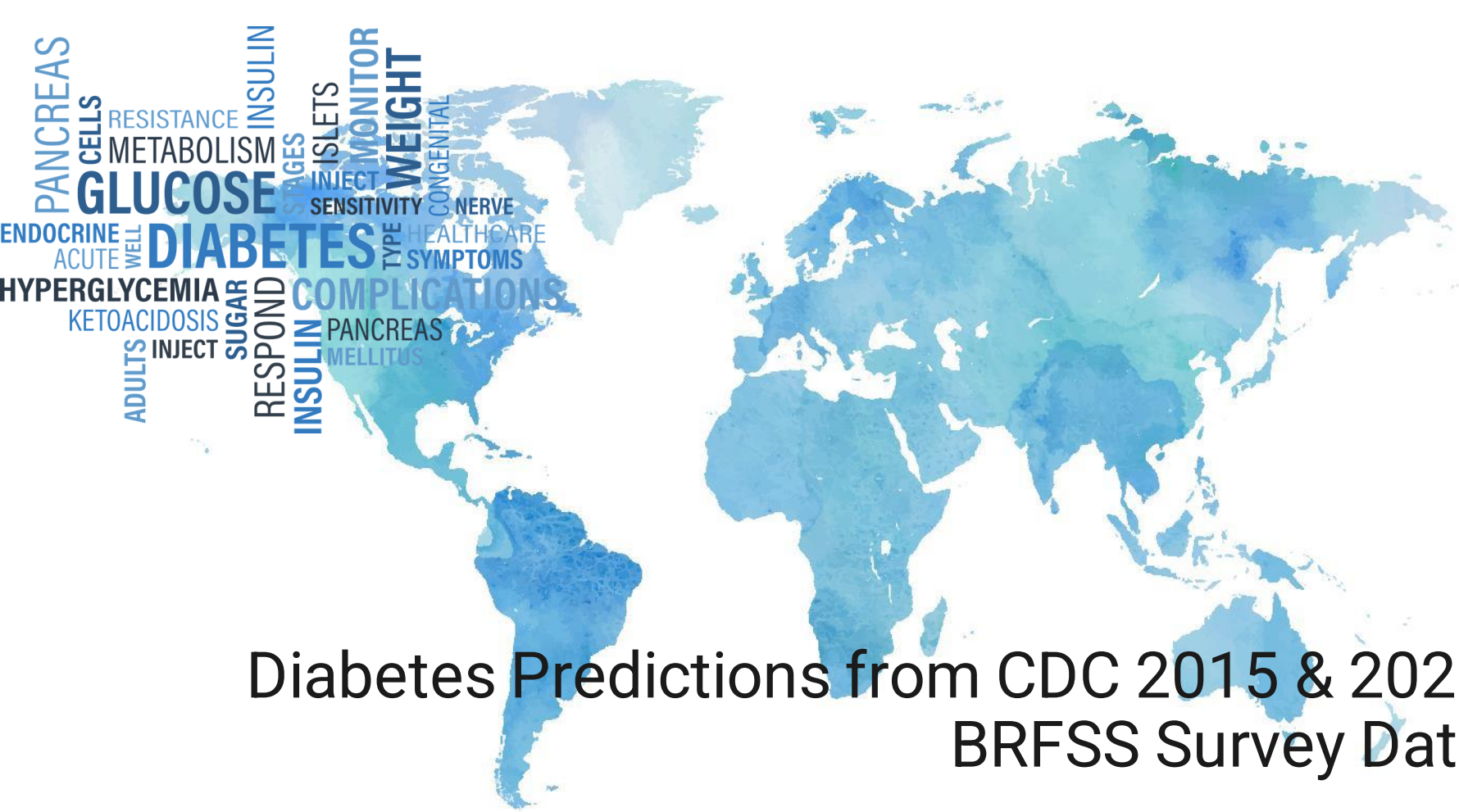


Diabetes Predictions from CDC 2015 & 2021 BRFSS Survey Data

DU Project 2 – Team 1 (Jeff Flachman, Elia Porter, Ava Lee)

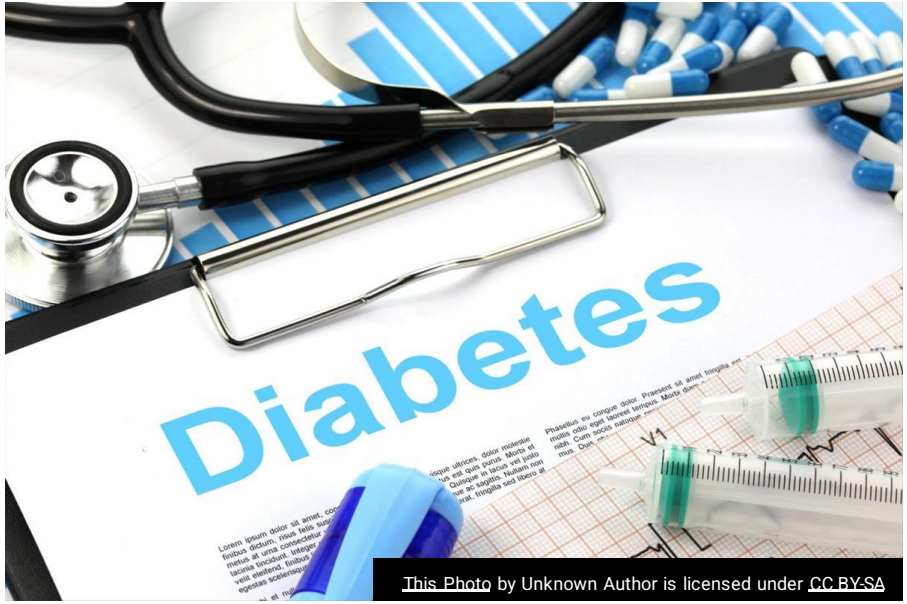
EXECUTIVE SUMMARY – PROJECT GOALS

Objective

- Analyze factors contributing to diabetes prevalence in the US
- Determine predictive value of these factors

Industry Relation

- Relevant to healthcare industry, focusing on predictive analytics for chronic disease management
- Addresses public health by identifying risk factors and potential interventions



Risk Factors:

- Age: Increased risk for individuals over 40
- Family history: Higher risk with diabetic relatives
- Ethnicity: Higher prevalence in certain races
- Inactivity: Lower physical activity increases risk
- Weight: Overweight and obesity are major risk factors
- Blood pressure: High BP can lead to insulin resistance
- Cholesterol: High levels increase diabetes risk
- Smoking: Smokers have a 30-40% higher risk of developing diabetes

Indicators/Symptoms:

- Frequent urination, excessive thirst, increased hunger, unintentional weight loss, fatigue, blurred vision, slow-healing wounds, itchy skin, infections, unusual sensations

DATA SOURCE, SELECTION, & COLLECTION PROCESS

2015 Dataset:

- UCI/Kaggle Dataset
  - Focused on 21 relevant features

2021 Dataset:

- Raw data pulled directly from CDC
  - Expanded to 36 features after cleaning and processing

Source:

- 2015 BRFSS dataset from UC Irvine Machine Learning Repository
- 2021 BRFSS dataset from CDC

Reason for Choice:

- Comprehensive annual survey data with relevant health and behavioral information
- BRFSS provides a large sample size, enhancing the reliability of predictive models

DATA CLEANUP & EXPLORATION PROCESS

Exploration Techniques:

- Automated processing of CDC codebooks to extract relevant features
- Evaluated features for relevance to diabetes risk
- Feature to Target Correlation
  - No single feature had strong correlation to target
- Feature to Feature correlation
  - Found and removed duplicate features

Feature Selection:

- Key features selected based on relevance to diabetes analysis

Dropped unknown/refused responses

Scaled certain values (e.g., weight from kg to standard units)

Transformed numeric responses (e.g., converting exercise days)

HOW DID WE ACHIEVE OUR PROJECT GOALS?

Steps Taken:

- **Used existing 2015 dataset and also evaluated, and cleaned 2021 survey data**
- **Defined two target variables:**
  - 0/1/2 (no diabetes, pre-diabetes, diabetes)
  - Binary 0/1 (no diabetes, diabetes)
- **Applied multiple classification** models to assess predictive power of each
- **Ranking method for Results**
- **Optimized best models/dataset combinations**

Developed Configuration Controlled Pipelines:

- **Codebook Pipeline:**
  - **Read codebook from CDC**
  - **Pulled all 300+ feature descriptions and provided a report for evaluation**
  - **Report used to select features applicable to diabetes risk factors**
- **Data Preparation Pipeline:**
  - **Read data for CDC or Downloaded zip**
  - **Applied additional feature transformation (imputation): feature scaling, and dropped rows with poor data quality**
- **Model Execution Pipeline:**
  - **Applied various imbalance methods: binary data, scaling and sampling methods**
  - **Trained models**
  - **Collected metrics**
  - **Generated performance report**
- **Optimization Pipeline:**
  - **Applied optimization methods with specified datasets and models**

MODEL TRAINING & EVALUATION

Models Used

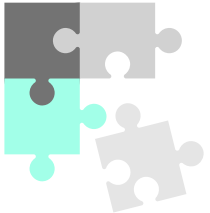
- KNeighborsClassifier
- DecisionTreeClassifier
- RandomForestClassifier
- ExtraTreesClassifier
- GradientBoostingClassifier
- AdaBoostClassifier
- LogisticRegression

Overfitting

- Observed a lot of overfitting with base dataset
- Largely due to imbalanced data
- Reduced using binary target feature, scaling techniques, and resampling methods

Evaluation Metrics:

- Score & Balanced Accuracy
- ROC AUC Score
- Mean Squared Error
- Accuracy
- Precision
- Recall
- F1 score
- Specificity
- False Positive Rate



HANDLING UNBALANCED DATA

Imbalanced Data:

- **Original target values:** 84% No diabetes, 2% Pre-diabetes, 14% Diabetes
- **Binary target values:** 86% No diabetes, 14% Diabetes

Sampling Methods:

- **RandomOverSampler**
- **RandomUnderSampler**
- **ClusterCentroids**
- **SMOTE**
- **SMOTEENN**
- **Note: Applied to Binary target with StandardScaler**

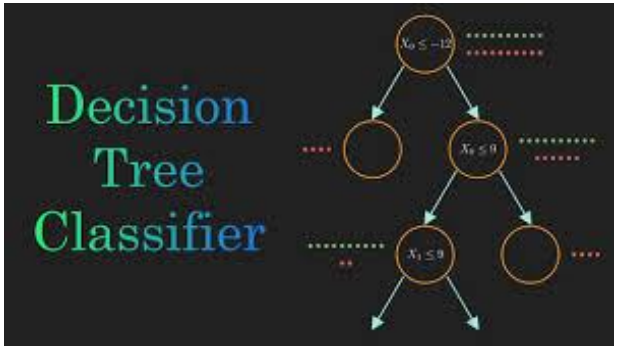
Outcome:

- **Binary target with StandardScaler and SMOTEENN or RandomUnderSampler** provided the smallest dataset and was used in the optimization phase for efficient training

- **Simplified Target Variable:**
  - **0/1/2 (no diabetes, pre-diabetes, diabetes)**
  - **Binary 0/1 (no diabetes, diabetes)**

Decision Tree Classifier + Randomized Search CV

- Sampled a fixed number of parameter settings from specified ranges for efficiency
- The optimization helped but not a substantial amount on this dataset




```
Original Decision Tree Classifier score: 0.6640871634638734
Balanced Accuracy Score: 0.6610727433209354
ROC AUC Score: 0.6610727433209355
```

```
Best Decision Tree Classifier score: 0.6808593307322444
Balanced Accuracy Score: 0.7252479681392824
ROC AUC Score: 0.7900964322506635
```

ADDITIONAL RESEARCH TOPICS WITH MORE TIME

Future Work:

- Expand dataset to select other years that surveyed features relevant to diabetes for broader analysis
- Implement real-time prediction models for use by healthcare providers and public health officials
- Try the pipeline we created to solve alternate problems

Potential Improvements:

- Incorporate more advanced machine learning techniques like deep learning
- Integrate data from other sources to enrich predictive models

RESULTS & CONCLUSIONS

Conclusions from 63 Model/Dataset Runs for each year. (126 total dataset/model combinations)

- We achieved good accuracy. But because of imbalance struggled with Precision
- Optimization helped some but did not make large gains for most models.

Top Models

- GradientBoostingClassifier
- AdaBoostClassifier
- LogisticRegression

Top Datasets

- Binary dataset with StandardScalar
- Binary, Standard Scalar & SMOTEEN sampling.

Project Goal: Achieved

- **Successfully identified key factors** contributing to diabetes prevalence.
- **Developed predictive models** with significant accuracy and reliability.
- **Strong Predictive performance through:**
  - Application of pipelines
  - Optimized datasets
  - Advanced classification models
  - Model performance ranking
  - Model Optimization


Questions