Predicting Type 2 Diabetes in California

By: Kelly Im, Brian Wang, Alvin Zou

Questions

1. How can we predict the risk of type 2 diabetes in adults in California?

2. Are there any events, including lifestyle practices, education level, family history, or health behaviors that are associated with the risk of type 2 diabetes?

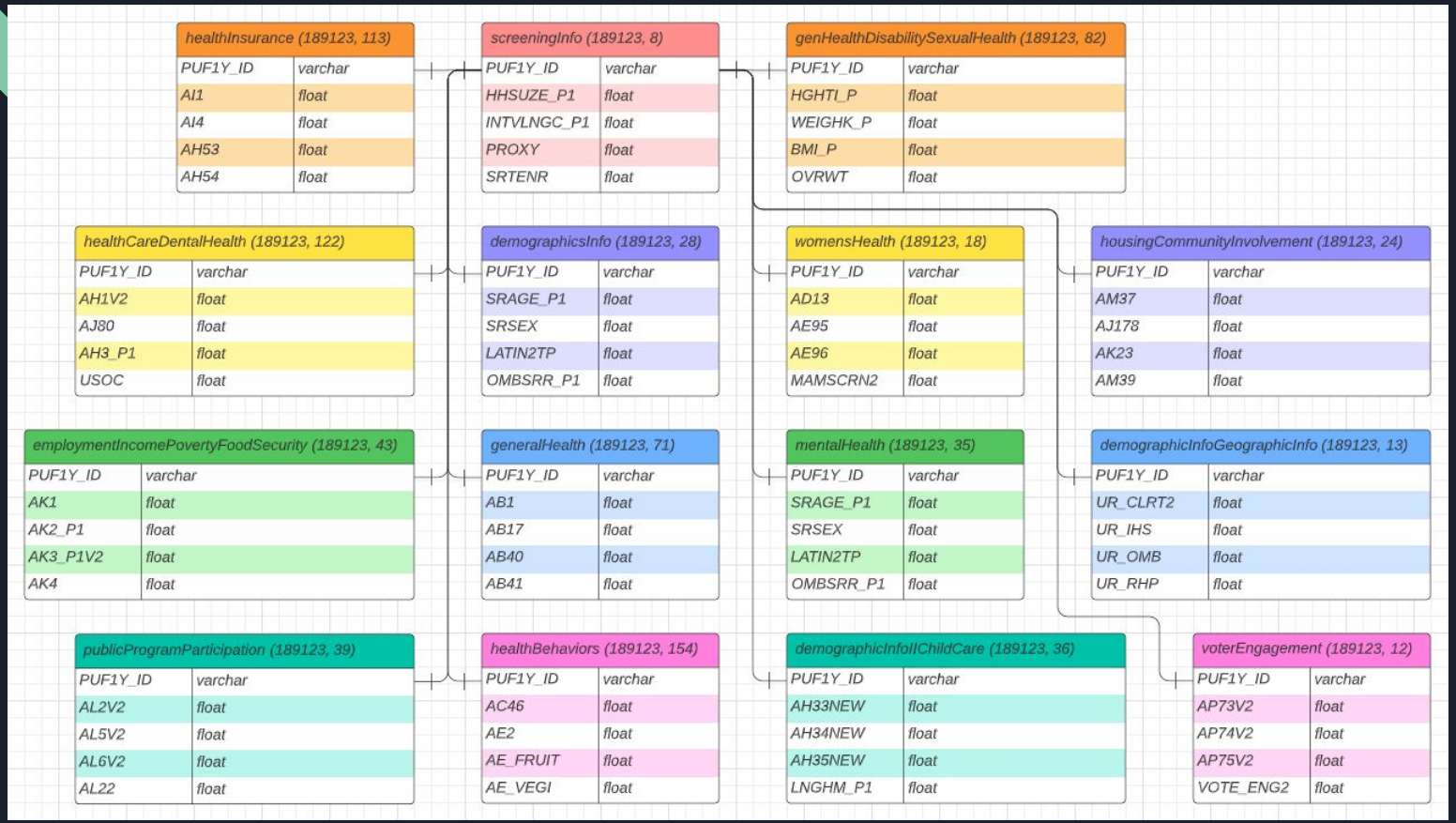
What is Type 2 Diabetes?

- “Diabetes is a chronic (long-lasting) health condition that affects how your body turns food into energy” (Centers for Disease Control and Prevention).
- Three main types of diabetes - Type 1, Type 2 and Gestational diabetes
- Significance of Type 2 diabetes:
  - 90 - 95% of people who have diabetes, have Type 2 diabetes.
  - Nearly half of adults in California have prediabetes or undiagnosed diabetes. While another 9% have been diagnosed with diabetes.
  - “Because diabetes is more common among older adults, the study’s finding that 33 percent of young adults aged 18-39 have prediabetes is of particular concern” (UCLA Health Policy Research).

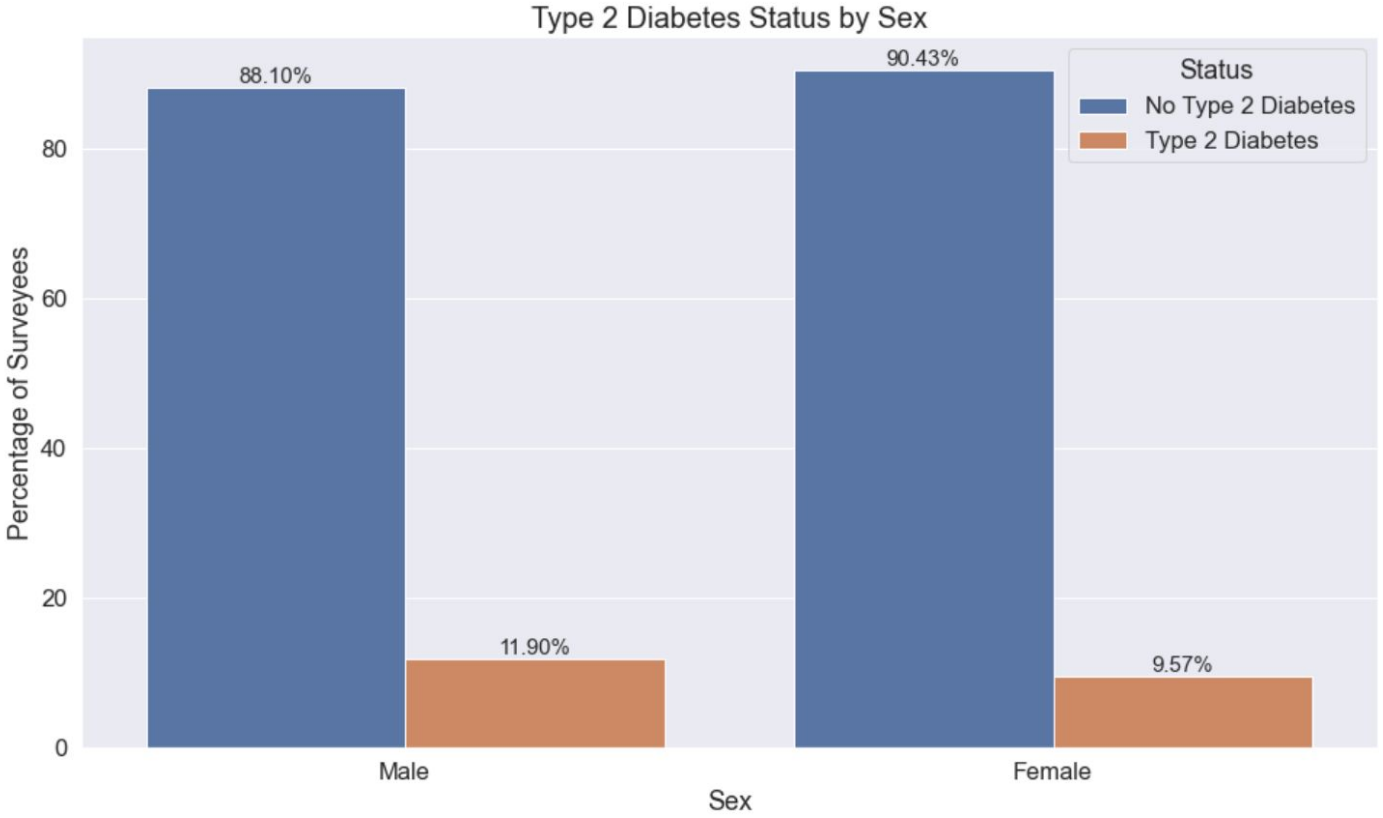
Dataset

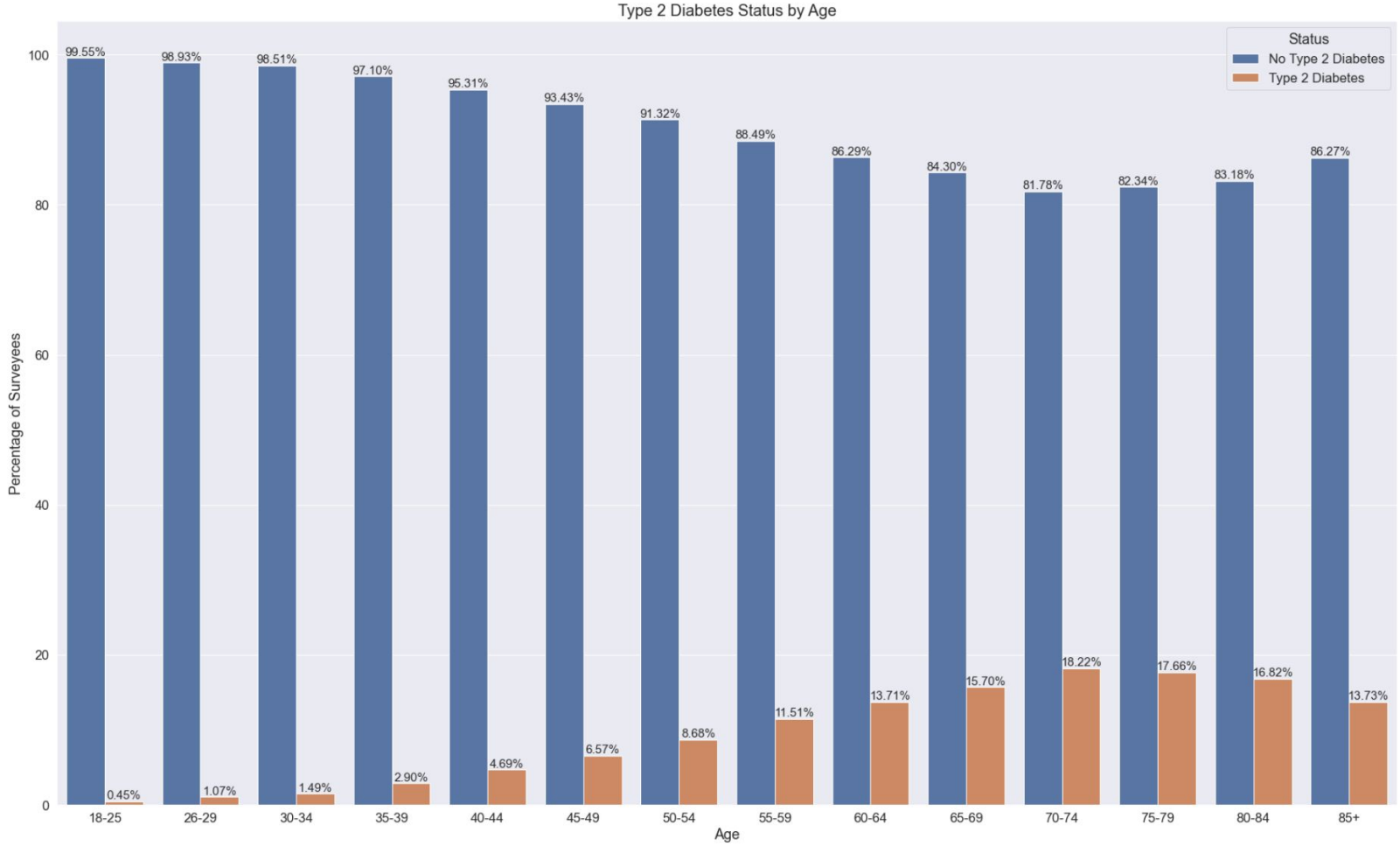
- California Health Interview Survey (CHIS)
  - Public Use Data
  - Individual Level Data
  - 9 Years
    - 2012 - 2020
- 189,123 surveyees and 785 covariates
- Encoded variables
  - Data dictionary (2020)
- Custom variables
  - T2D (Type 2 Diabetes - 1 / 0)
  - Year (2012 - 2020)

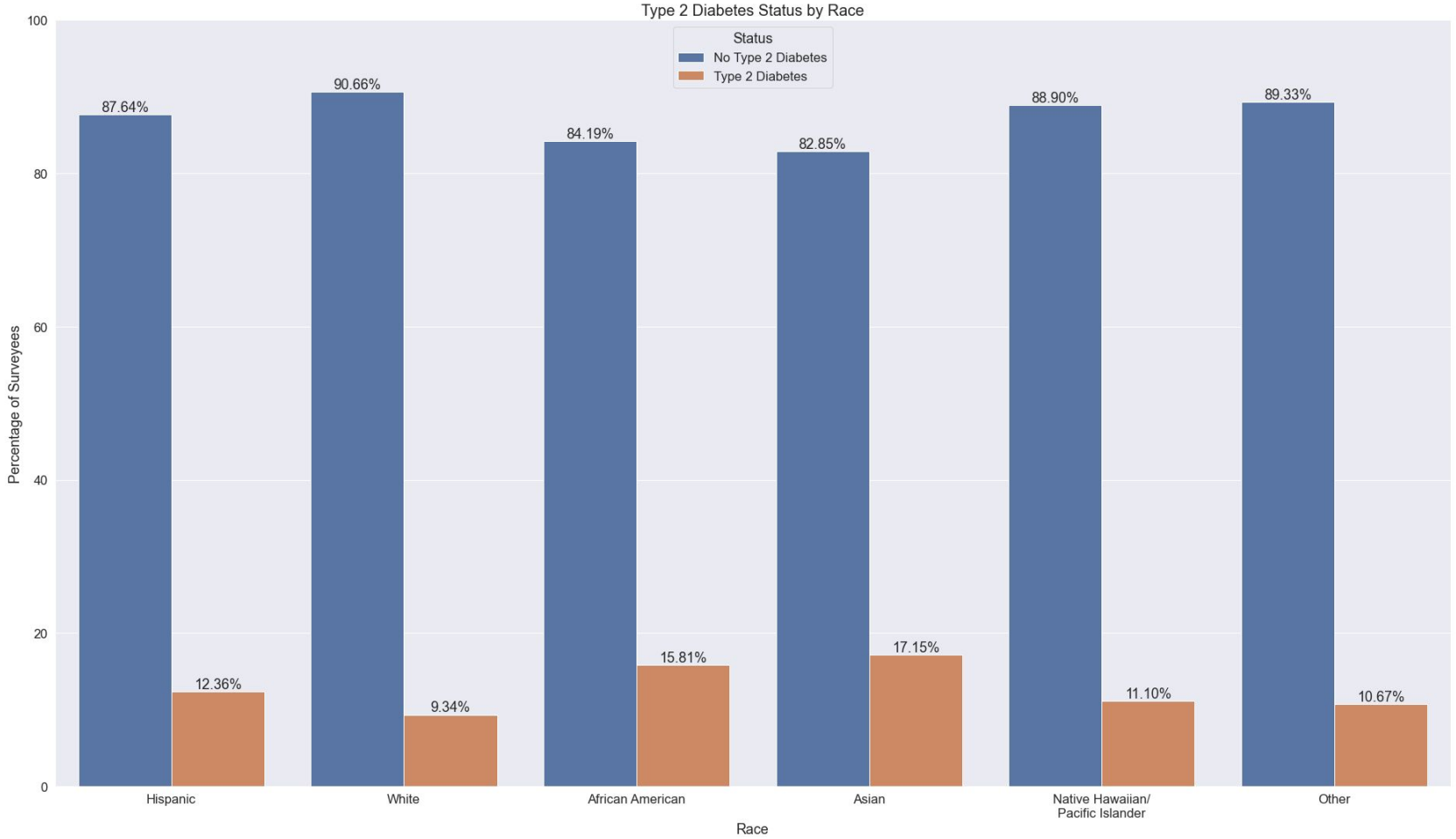
Data Management



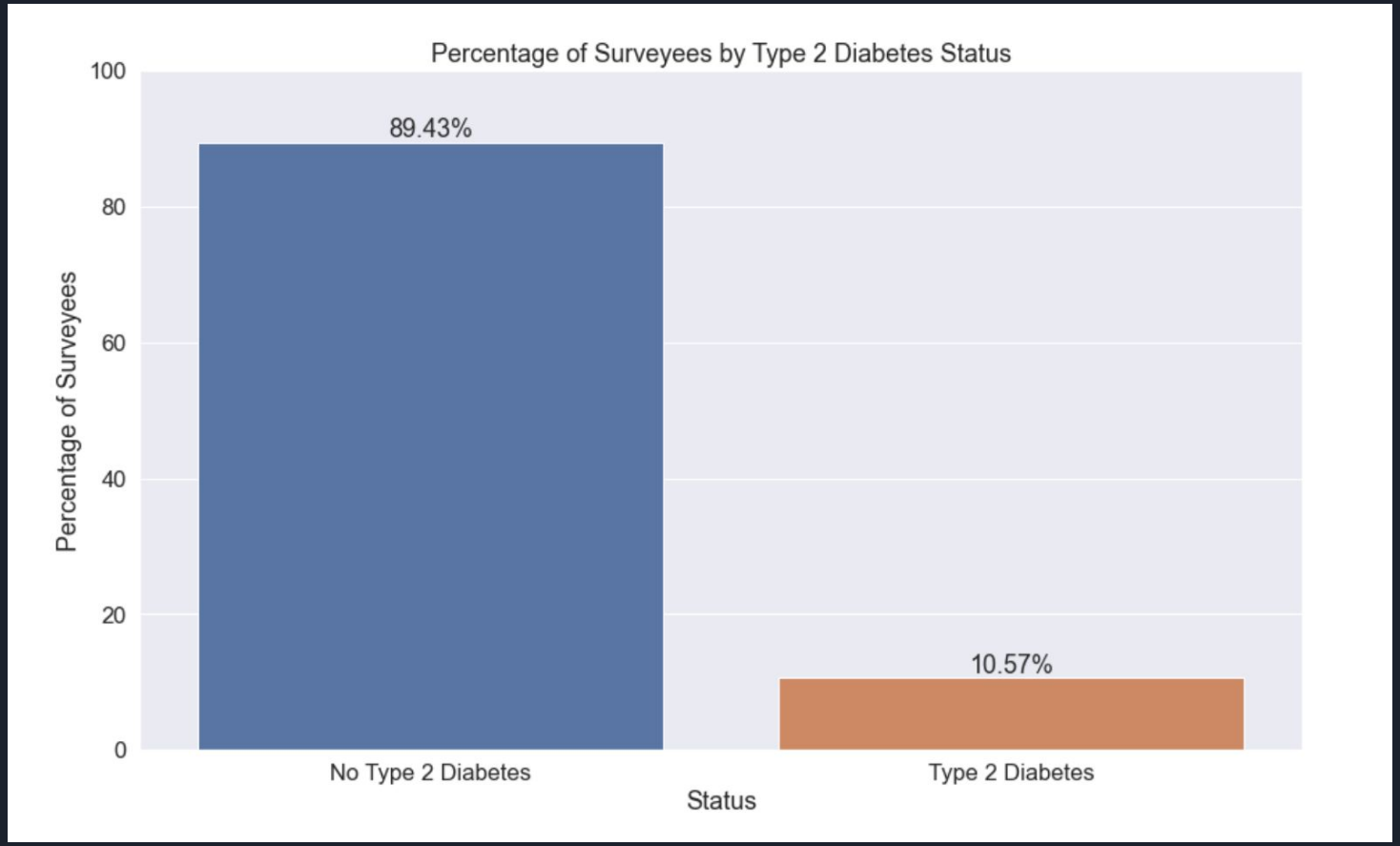
Exploratory Data Analysis




Type 2 Diabetes Status by Age
Status
No Type 2 Diabetes
Type 2 Diabetes
Percentage of Surveyees
Age
0
20
40
60
80
100
18-25
26-29
30-34
35-39
40-44
45-49
50-54
55-59
60-64
65-69
70-74
75-79
80-84
85+
99.55%
0.45%
98.93%
1.07%
98.51%
1.49%
97.10%
2.90%
95.31%
4.69%
93.43%
6.57%
91.32%
8.68%
88.49%
11.51%
86.29%
13.71%
84.30%
15.70%
81.78%
18.22%
82.34%
17.66%
83.18%
16.82%
86.27%
13.73%


Type 2 Diabetes Status by Race
Status
No Type 2 Diabetes
Type 2 Diabetes
Percentage of Surveyees
100
80
60
40
20
0
87.64%
12.36%
90.66%
9.34%
84.19%
15.81%
82.85%
17.15%
88.90%
11.10%
89.33%
10.67%
Hispanic
White
African American
Asian
Native Hawaiian/
Pacific Islander
Other
Race

Class Imbalance



Ways to Deal With Class Imbalance

- Cost-Sensitive Learning
  - Takes the costs of prediction errors (and potentially other costs) into account when training a machine learning model.
  - Assigns different costs to the types of misclassification errors
- Synthetic Minority Oversampling Technique (SMOTE)
  - Oversampling method that selects a random sample in the minority class, finds its *k* nearest neighbors, randomly selects a neighbor, and then creates a synthetic example between the two examples.

Features Not Included

Variable	Label	Reason
PUF1Y_ID	Public Use File ID	Participant's ID
AB111	Admitted to Hospital Overnight or Longer for Diabetes Past 12 Months	Universe = Adults who have diabetes
AB23_P1	Age First Told Have Diabetes	Universe = Adults who have diabetes
AB114_P1	Confidence to Control and Manage Diabetes	Universe = Adults who have diabetes
AJ82	Delayed Prescription for Diabetes	Universe = Adults who have diabetes who cited cost/lack of insurance as reason for delay in getting prescribed medicine
AB22V2	Doctor Ever Told Have Diabetes	May imply the participant already has diabetes
DIABETES	Doctor Ever Told Have Diabetes (Non-Gestational)	May imply the participant already has diabetes
AB81	Doctor Told Had Diabetes Only During Pregnancy	May imply the participant already has diabetes at the time of pregnancy
AB113	Have Written Copy of Diabetes Care Plan	Universe = Adults who have diabetes and medical providers developed diabetes care plan

Features Not Included Cont.

Variable	Label	Reason
AB112	Medical Providers Develop Diabetes Care Plan	Universe = Adults who have diabetes
AB51_P1	Type 1 or Type 2 Diabetes	Universe = Adults who have been told they have Type I or Type II diabetes
AB110_P	Visited ER for Diabetes Because Unable to See Own DR	Universe = Adults who have diabetes and visited ER for diabetes in past 12 months
AB109	Visited ER for Diabetes in Past 12 Months	Universe = Adults who have diabetes
AB24	Currently Taking Insulin	Universe = Adults who have diabetes
DIAMED	Taking Insulin or Pills	Universe = Adults who have diabetes
AB25	Currently Taking Diabetic Pills to Lower Blood Sugar	Universe = Adults who have diabetes
AB63	Last Eye Exam Dilated Pupils	Universe = Adults who have diabetes
AB27_P1	Number of Times Doctor Checked for Hemoglobin A1C Last Year	Universe = Adults who have diabetes

Features Not Included Cont.

Variable	Label	Reason
AB28_P1	Number of Times Doctor Checked Feet for Sores Last Year	Universe = Adults who have diabetes
AB26_P1	Number of Times Respondent/R's Family/Friend Check Glucose	Universe = Adults who have diabetes
DIABCK_P1	Number of Times Checking for Glucose/Sugar Per Month	Universe = Adults who have diabetes or sugar diabetes
AB27_P	Number of Times Doctor Checked Hemoglobin A1C Last Year	Universe = Adults who have diabetes
AJ80	Has Someone at Doctor's Office/Clinic Who Helps Coordinate Medica	Universe = Adults who have usual source of care and has personal doctor and have asthma or diabetes or heart disease
AH102_P	Number of Nights in Hospital Past 12 Months	Universe = Adults who were hospitalized for asthma, diabetes or heart disease during the past 12 months
AH102_P1	Number of Nights in Hospital Past 12 Months	Universe = Adults who were hospitalized for asthma, diabetes or heart disease during the past 12 months

Features Not Included Cont.

Variable	Label	Reason
WEIGHK_P	Weight - KG	Similar feature to another feature in the dataset
WGHTK_P	Weight - KG	Similar feature to another feature in the dataset
HEIGHM_P	Height - Meters	Similar feature to another feature in the dataset
HGHTM_P	Height - Meters	Similar feature to another feature in the dataset
WHOBMI	Body Mass Index - WHO Definition	Similar feature to another feature in the dataset

Feature Selection

- Used a Random Forest Classifier to get the important features for feature selection
  - Number of features: 785 → 245
- Some of the important features
  - AB99 - Doctor Ever Told Surveyee To Have Pre-Diabetes
  - BMI_P - Body Mass Index
  - WGHT_P - Weight (pounds)

Models

- K-Nearest Neighbors
  - Takes the *k* nearest neighbors of a data point and assigns the data point the same label based on majority rule and clusters the data points into *k* groups.
- Stochastic Gradient Descent
  - Iterative algorithm that starts from a random point on a function and randomly picks one point at each iteration to travel down its slope until it reaches the lowest point of the function.
- Decision Trees
  - Builds up a set of decision rules in the form of a tree which helps predict an outcome.
- Random Forest
  - Consists of many decision trees and takes the prediction from each tree and selects the final prediction based on majority rule.
- XGBoost
  - Algorithm that combines decision trees and gradient boosting.
  - Combines the predictions of numerous simple decision trees into one.

Performance Metric Definitions

- Accuracy
  - The proportion of respondents who have or don't have Type 2 diabetes and are correctly classified among the total number of cases examined.
- Precision
  - The proportion of people who have Type 2 diabetes and are correctly classified among all respondents who are predicted to have Type 2 diabetes.
- Recall
  - The proportion of respondents who have Type 2 diabetes and are correctly classified among those who actually have Type 2 diabetes.
- F1 Score
  - The harmonic mean of precision and recall. The value is between 0 and 1 - the closer the score is to 1, the more accurate the model.
- ROC AUC Score
  - The ROC curve plots the True Positive Rate against the False Positive Rate. The closer the curve is to the top left corner, the better the model is.
  - AUC stands for Area Under the Curve. The value is between 0 and 1 - the closer the score is to 1, the more accurate the model.
- Confusion Matrix
  - A table used to describe the performance of a classification model on a set of test data for which the true values are known.
- Stratified K-Folds Cross-Validation
  - Splits the training data into *k* folds with each fold containing a representative ratio of each class.

Confusion Matrix

Predicted

	Doesn't Have Type 2 Diabetes	Has Type 2 Diabetes
Doesn't Have Type 2 Diabetes	**True Negative (TN)**	**False Positive (FP)**
Has Type 2 Diabetes	**False Negative (FN)**	**True Positive (TP)**

Actual

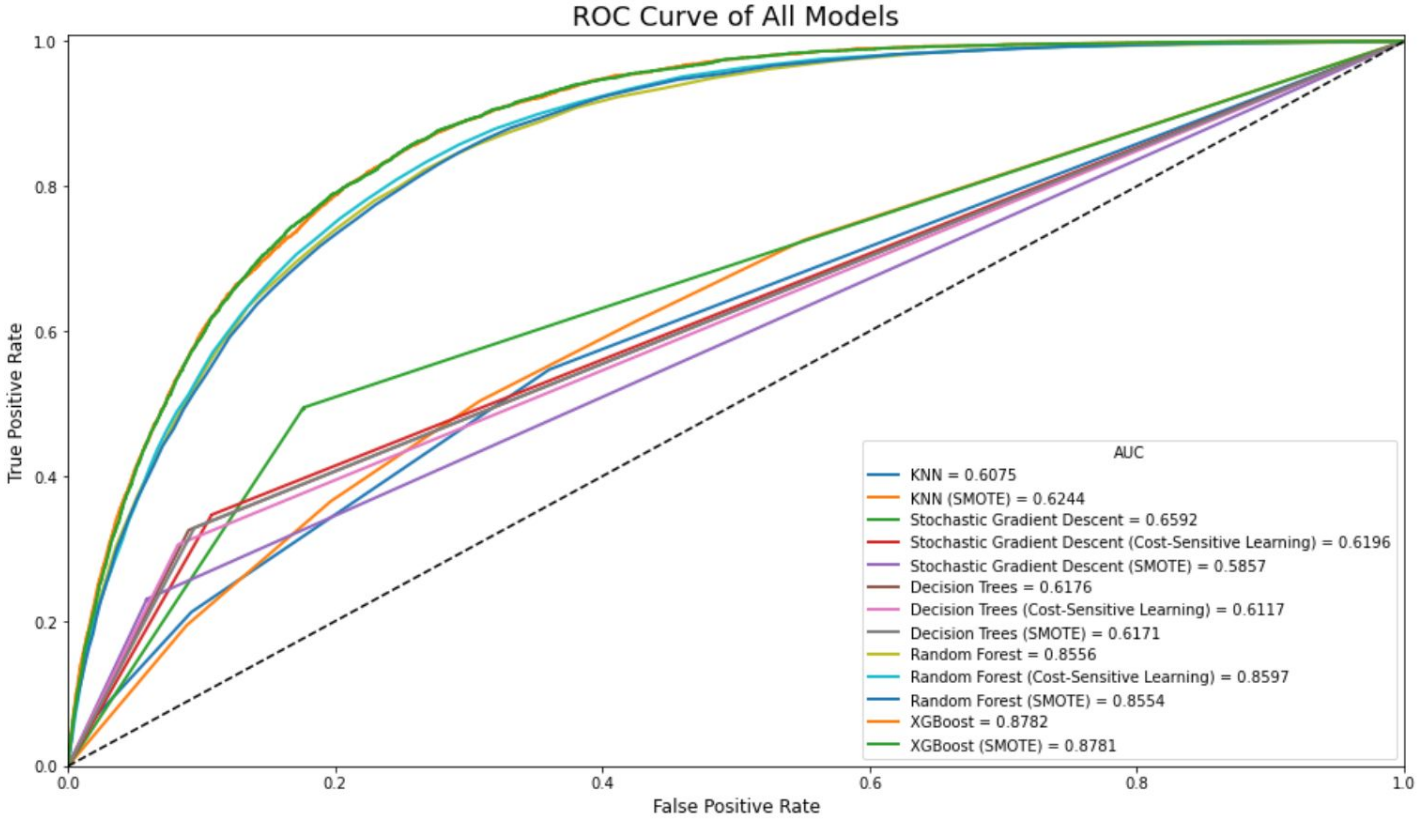
Performance Metric Equations

- Accuracy
  - (TP + TN) / (TP + FP + FN + TN)
- Precision
  - TP / (TP + FP)
- Recall
  - TP / (TP + FN)
- F1 Score
  - 2 * ((Precision * Recall) / (Precision + Recall))

Final Results Without Cross-Validation

Model	F1 Score	ROC AUC Score
KNN	0.1075	0.6075
KNN (SMOTE)	0.2437	0.6244
Stochastic Gradient Descent	0.2370	0.6592
Stochastic Gradient Descent (Cost-Sensitive Learning)	0.1083	0.6196
Stochastic Gradient Descent (SMOTE)	0.2624	0.5857
Decision Trees	0.3095	0.6176
Decision Trees (Cost-Sensitive Learning)	0.3031	0.6117
Decision Trees (SMOTE)	0.3074	0.6171
Random Forest	0.0952	0.8556
Random Forest (Cost-Sensitive Learning)	0.0913	0.8597
Random Forest (SMOTE)	0.1481	0.8554
XGBoost	0.3522	0.8782
XGBoost (SMOTE)	0.3605	0.8781

ROC-AUC Curve Of All Models



Final Results With Cross-Validation

Model	F1 Score	ROC AUC Score
KNN	0.8841	0.6041
KNN (SMOTE)	0.6704	0.6221
Stochastic Gradient Descent	0.7178	0.6872
Stochastic Gradient Descent (Cost-Sensitive Learning)	0.7665	0.7139
Stochastic Gradient Descent (SMOTE)	0.8181	0.7283
Decision Trees	0.8468	0.6140
Decision Trees (Cost-Sensitive Learning)	0.8510	0.6076
Decision Trees (SMOTE)	0.8460	0.6169
Random Forest	0.8961	0.8533
Random Forest (Cost-Sensitive Learning)	0.8957	0.8582
Random Forest (SMOTE)	0.8960	0.8524
XGBoost	0.8982	0.8744
XGBoost (SMOTE)	0.8981	0.8749

Conclusion

- How can we predict the risk of type 2 diabetes in adults in California?
  - XGBoost Model
    - Highest F1 Score (0.8982)
    - High ROC AUC score (0.8744)
- Are there any events, lifestyle practices, education level, family history, or health behaviors associated with the risk of type 2 diabetes?
  - Features including age, race, sex, weight, BMI, high blood pressure, heart disease, prediabetes, and diet, such as the number of times the surveyee drank soda per week, are associated with Type 2 diabetes.
- How can we continue this project?
  - Extend our findings to datasets that cover more regions
  - National Health Interview Survey (NHIS) dataset for nationwide coverage

References

- *What is Diabetes?* (2022, March 2). Centers for Disease Control and Prevention. Retrieved May 25, 2022, from https://www.cdc.gov/diabetes/basics/diabetes.html
- *Type 2 Diabetes*. (2022, March 2). Centers for Disease Control and Prevention. Retrieved May 25, 2022, from https://www.cdc.gov/diabetes/basics/type2.html
- *Majority of California Adults Have Prediabetes or Diabetes - U Magazine - UCLA Health - Los Angeles, CA*. (2016). UCLA Health. Retrieved May 25, 2022, from https://www.uclahealth.org/u-magazine/majority-of-california-adults-have-prediabetes-or-diabetes
- Goad, B. K. (2018, November 2). *What's Race Got to Do With Diabetes?* AARP. Retrieved May 25, 2022, from https://www.aarp.org/health/healthy-living/info-2018/role-of-race-in-diabetes.html#:%7E:text=What%20you%20may%20not%20know,American%20Diabetes%20Association%20(ADA)

Thank you!

Any Questions?