Predicting risk of Type 2 diabetes with AI in California

Brian Wang[1], Kelly Im[1], Alvin Zou[1], Shawna Tuli[2], Vishrut Chokshi[2], Mei-Chi Chen[2], Jeffrey Lu[2], Ella Liang[2], Manish Dasaur[2], Marty Hodgett[2], Jeffrey Vu[2], Sachin Dalal[2], Michelle Smith[2], Praneet Raj[2], Bryan S. Price[2], Erik Kronstadt[2], and Anthony Agbasi[2]

[2] *University of California, Irvine*
[2] *Accenture California*

September 2024

---

1 Abstract

1.1 Background

Type 2 diabetes is a rising problem among the California population, with nearly half of all adults in California being pre-diabetic or undiagnosed. The purpose of this study is to determine what factors may contribute to the presence of Type 2 diabetes, and whether the risk of Type 2 diabetes can be predicted using these factors. To achieve this, we analyzed public survey data from the California Health Interview Survey (CHIS) - a statewide health survey conducted yearly by the UCLA Center for Health Policy Research. Participants were aged 18 or older, and important demographics and health topics, such as marital status, education, and exercise tendencies, were recorded. The data analyzed spans over the course of 9 years, from 2012 to 2020 and the combined dataset contains about 190,000 surveyed and 785 covariates (Figure 1).

1.2 Methods

The chosen variables (Table 1) were fitted using KNN, stochastic gradient descent, decision trees, random forests, and XGBoost. To account for the class imbalance of the dependent variable, we implemented cost-sensitive learning and SMOTE to each model. The accuracy of each model was evaluated using its F1 and AUC scores (Table 2).

1.3 Results

The results indicate the XGBoost model had the highest scores. This model performed with an F1 score of 0.897 and an AUC score of 0.874 (Table 3). Given the results, the model demonstrated a strong capability of accurately predicting the risk of Type 2 diabetes in California adults and the fitted variables may be significantly tied to affecting the risk.

1.4 Conclusion

Therefore, it can be valuable for an individual to monitor these factors. In conclusion, the most significant factors for predicting the risk of Type 2 diabetes are BMI, high blood pressure, and if they have prediabetes.

2 Introduction

Diabetes is a long-term (chronic) health condition that is diagnosed among over 3 million people in California (10.5% of the adult population), with an additional 884,000 people who are undiagnosed. Nearly half of adults in California either have prediabetes or are unaware of their illness, which in turn greatly increases their health risk.

90-95% of people who have diabetes have Type 2 diabetes. In 2017, Type 2 diabetes was reported in nearly 2.6 million California adults, with 15.6% of adults estimated to have prediabetes. According to the California Department of Public Health (CDPH), there is an estimated 84 million people who have prediabetes, a condition marked higher than normal blood sugar levels, which raises the risk of developing Type 2 diabetes. Since 2013, the prevalence of prediabetes and diabetes among California adults has increased significantly. These rates are higher among racial and ethnic minorities, as well as older adults. American Indian/Alaska Native, Hispanic, Latino, African American, and Asian/ Pacific Islanders are about twice as likely to have Type 2 diabetes than white adults. Additionally, there are geographic areas in California that have a higher prevalence of diabetes.

Type 2 diabetes, in contrast to Type 1, can be avoided. To prevent the development of the disease, it is important to be aware of the risk factors for Type 2 diabetes which include older age, race/ethnicity, personal history of prediabetes or gestational diabetes, a family history of Type 2 diabetes, obesity, and inactivity.

To determine the factors that may contribute to the presence of Type 2 diabetes, and if we can use these factors to predict the risk of Type 2 diabetes, we analyzed public survey data from the California Health Survey (CHIS) and from a statewide health survey conducted yearly by the UCLA Center for Health Policy Research. These two surveys captured participants from different demographics and health topics including marital status, education, and exercise. Through these combined datasets and analytics, we deployed different models to predict the risks of Type 2 diabetes.

3 Methods

The California Health Interview Survey (CHIS) dataset includes both public use data and individual level data spanning 9 years from 2012 to 2020. The dataset details 189,123 surveyed and 785 covariates with encoded variables and custom variables T2D (Type 2 Diabetes – 1/0) and Year (2012-2020).

To deal with class imbalance, cost-sensitive learning, and synthetic minority, oversampling technique (SMOTE) were applied. Cost-sensitive learning takes the costs of prediction errors (and potentially other costs) into account when training a machine learning model and assigns different costs to the types of misclassification errors. The oversampling method 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 which were excluded are Public Use File ID, Admitted to Hospital Overnight or Longer for Diabetes Past 12 Months, Age First Told Have Diabetes, Confidence to Control and Manage Diabetes, Delayed Prescription for Diabetes, Doctor Ever Told Have Diabetes (Non-Gestational), Doctor Told Had Diabetes Only During Pregnancy, Have Written Copy of Diabetes Care Plan, Medical Providers Develop Diabetes Care Plan, Type 1 or Type 2 Diabetes, Visited ER for Diabetes in Past 12 Months, Currently Taking Insulin, Taking Insulin or Pills, Currently Taking Diabetic Pills to Lower Blood Sugar, Last Eye Exam Dilated Pupils, Number of Times Doctor Checked for Hemoglobin A1C Last Year, Number of Times Doctor Checked Feet for Sores Last Year, Number of Times Respondent/R's Family/Friend Check Glucose, Number of Times Checking for Glucose/Sugar Per Month, Number of Times Doctor Checked Hemoglobin A1C Last Year, Has Someone at Doctor's Office/Clinic Who Helps Coordinate Medica, Number of Nights in Hospital Past 12 Months, Weight – KG, Height – Meters, Body Mass Index – WHO Definition (Table 1).

Random Forest Classifier gets the important features for feature selection from 785 to 245 features; some of the important features are AB99 – Doctor Ever Told Surveyee To Have Pre-Diabetes, BMI_P–$BodyMassIndex$, $WGHT_P$–$Weight(pounds)(Table1)$.

The F1 score is the harmonic mean of precision, the proportion of people who have Type 2 diabetes and are correctly classified among those who actually have Type 2 diabetes, and recall, the proportion of respondents who have Type 2 diabetes and are correctly classified among those who actually have Type 2 diabetes. The ROC curve plots the True Positive Rate against 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, Area Under the Curve, is the value between 0 and 1 - the closer the score is to 1, the more accurate the model (Figure 2).

Accuracy is calculated as (True Positive + True Negative) / (True Positive + False Positive + False Negative + True Negative), Precision as True Positive / (True Positive + False Positive), Recall as True Positive / (True Positive + False Negative), and F1 Score as 2 * (Precision * Recall) / (Precision + Recall).

The chosen variables were fitted using KNN, stochastic gradient descent, decision trees, random forests, and XGBoost. To account for the class imbalance of the dependent variable, we implemented cost-sensitive learning and SMOTE to each model. The accuracy of each model was evaluated using its F1 and AUC scores (Table 3).

4 Results

Features including age, race, sex, weight, BMI, high blood pressure, heart disease, prediabetes, and diet, such as the number of times the surveyed drank soda per week, are associated with Type 2 diabetes. This project can be continued by extending our findings to datasets that cover more regions; the National Health Interview Survey (NHIS) dataset for nationwide coverage.

The results indicate the XGBoost model had the highest scores. This model performed with an F1 score of 0.897 and an AUC score of 0.874 (Table 3). Given the results, the model demonstrated a strong capability of accurately predicting the risk of Type 2 diabetes in California adults and the fitted variables may be significantly tied to affecting the risk. With an F1 score of 0.897 and an AUC score of 0.874, the XGBoost model has shown the highest potential in accurately predicting the risk of Type 2 diabetes in California (Table 3).

Based on the findings, features including age, race, sex, weight, SMI, high blood pressure, heart disease, prediabetes, and diet, are the top risk factors associated with Type 2 diabetes. We deliver a thorough analysis by combining several different datasets that are both publicly and privately available and encompass the many factors of a person's life as well as predicting several different health outcomes.

5 Conclusions

Among the different models that we employed, KNN, Stochastic Gradient Descent, Decision Trees, Random Forest, and XGBoost, XGBoost was the most accurate in predicting the risk of Type 2 Diabetes in California. It had an F1 Score of 0.8982 and a ROC AUC Score of 0.8744. To test for this, we used models that combine several other machine learning models to best predict an outcome. We are able to mimic this model against other diabetes data. Through our research, we learned that the top 10 risk factors for Type 2 diabetes is: age, race, sex, weight, BMI, high blood pressure, heart disease, prediabetes, and diet. XGBoost uses an algorithm that combines decision trees and gradient boosting as well as the predictions of numerous simple decision trees into one.

6 Figures

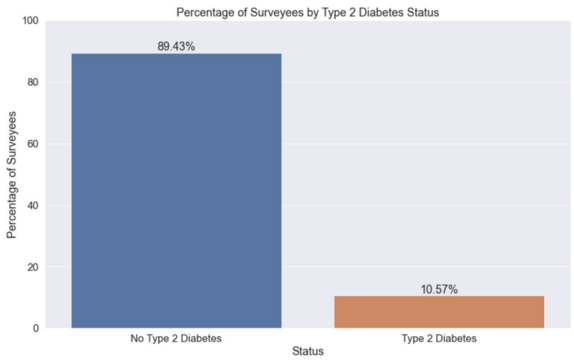

Figure 1: Percentages of people who responded if they do or do not have diabetes to the survey taken regarding people within Southern California.

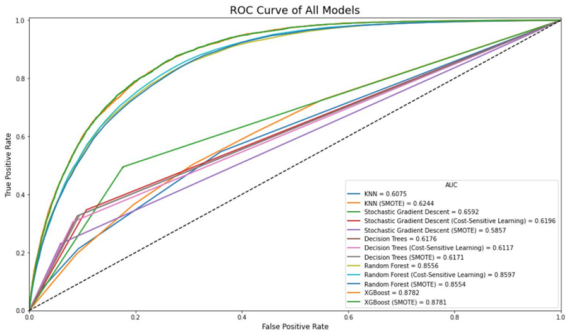

Figure 2: Comparison of different models and their accuracy, based on their false positive versus positive rates.

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

Table 1: The thirteen various models used, including their F1 and ROC AUC score. These results were taken without cross-validation.

Model 1 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

Table 2: The thirteen various models used, including their F1 and ROC AUC score. These results were taken with cross-validation.

7 References

1. Dinh A, Miertschin S, Young A, Mohanty SD. A data-driven approach to predicting diabetes and cardiovascular disease with Machine Learning - BMC Medical Informatics and Decision making [Internet]. BioMed Central. BioMed Central; 2019 [cited 2022Aug19]. Available from: https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-019-0918-5

2. Chen P-HC, Liu Y, Peng L. How to develop machine learning models for Healthcare [Internet]. Nature News. Nature Publishing Group; 2019 [cited 2022Aug19]. Available from: https://www.nature.com/articles/s41563-019-0345-0

3. Montazeri M, Montazeri M, Montazeri M, Beigzadeh A. Machine learning models in breast cancer survival prediction [Internet]. Technology and Health Care. IOS Press; 2016 [cited 2022Aug19]. Available from: https://content.iospress.com/articles/technology-and-health-care/thc1071

4. Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction [Internet]. Computational and Structural Biotechnology Journal. Elsevier; 2014 [cited 2022Aug19]. Available from: https://www.sciencedirect.com/science/article/pii/S2001037014000464

5. Haffner SM. Epidemiology of type 2 diabetes: Risk factors [Internet]. American Diabetes Association. American Diabetes Association; 1998 [cited 2022Aug19]. Available from: https://diabetesjournals.org/care/article/21/Supplement$_3$/$C3$/18526/$Epidemiology-of-Type-2-Diabetes-Risk-Factors$

6. Wilmot E, Idris I. Early onset type 2 diabetes: Risk factors, clinical impact and ... [Internet]. Sage Journals. Sage Journals; 2014 [cited 2022Aug19]. Available from:
https://journals.sagepub.com/doi/10.1177/2040622314548679

7. Panesar A. Machine learning and AI for Healthcare [Internet]. SpringerLink. Apress; 2020 [cited 2022Aug19]. Available from:
https://link.springer.com/book/10.1007/978-1-4842-6537-6

8. Shaheen MY. Cite this version: Shaheen, M. Y. (2021). Applications of Artificial Intelligence (AI) in healthcare: A review.1Applications of Artificial Intelligence (AI) in healthcare: A review [Internet]. ScienceOpen.com. ScienceOpen; 2021 [cited 2022Aug19]. Available from:
https://www.scienceopen.com/hosted-document?doi=10.14293/S2199-1006.1.SOR-.PPVRY8K.v1

9. Diamant AL, Babey SH, Brown ER, Neetu C. Diabetes in California- Findings from 2001 California Health Interview Survey [Internet]. UCLA Center for Health Policy Research; 2003 [cited 2022Aug19]. Available from:
http://ask.chis.ucla.edu/publications/Documents/PDF/
Diabetes%20in%20California%20Findings%20from

10. Taylor Cw, Downie C, Mercado V. Burden of Diabetes in California- June 2019 [Internet]. California Department of Public Health; 2019 [cited 2022Aug19]. Available from:
https://www.cdph.ca.gov/Programs/CCDPHP/DCDIC/CDCB/CDPH/
%20Document/%20Library/2019/%20Diabetes/%20Burden/%20Report/
%20(SCOTT$_9$$JUNE$2020).*pdf*