Analyzing mortality and its determinants among individuals with type 1 and type 2 diabetes

STATS 601

Tim White

Due April 27th, 2023

Abstract

Diabetes mellitus has consistently ranked among the ten leading causes of death in the United States for the past several decades. Diabetes-related mortality is prevalent across a wide range of sociodemographic subpopulations, and many previous studies have attempted to identify the risk factors and comorbidities that contribute most heavily to this phenomenon. However, surprisingly little attention has been paid to the distinction between the type 1 and type 2 variants of the disease in the context of mortality. In this report, we fill this gap in the literature by using unsupervised and supervised statistical learning methods to analyze the mortality risk profiles of individuals with type 1 and type 2 diabetes. We apply a dimension reduction technique to multiple-cause-of-death mortality data from the Centers for Disease Control and Prevention to explore the latent sociodemographic and health profiles of individuals whose deaths were attributed to diabetes in 2021, and we train several classification models to differentiate between type 1 and type 2 diabetes as a cause of death based on these characteristics. Our results suggest that sophisticated classification methods are capable of achieving moderate accuracy in distinguishing deaths due to type 1 diabetes from those due to type 2 diabetes, with tree-based and optimization-based classifiers such as random forest, AdaBoost, and kernel SVM providing a better holistic performance than model-based classifiers such as naive Bayes, quadratic discriminant analysis, and penalized logistic regression. We find that age is the most useful predictor for this classification task, followed by other sociodemographic predictors such as education, marital status, race, place of death, and sex. These findings provide important insights that could potentially improve the ability of practitioners to assess mortality risk in patients with type 1 and type 2 diabetes.

Contents

1 Introduction

1.1 Background

Diabetes mellitus — a chronic endocrine disorder that hinders the body's ability to produce and/or use insulin — has ranked among the ten leading causes of mortality in the United States since the middle of the 20th century [1, 2]. It was the eighth leading cause in 2021, accounting for three percent of the country's deaths [3]. Diabetes-related mortality is prevalent across many sociodemographic groups [4, 5], which makes it challenging to identify the underlying factors that contribute most heavily to the deaths of diabetic individuals. Overcoming this challenge is crucial, as the ability of practitioners to accurately assess mortality risk in people with diabetes could greatly improve the care they provide to these patients.

One important consideration in any study involving diabetes is the distinction between the two main types of the disease. As explained by Ozougwu et al. [6], type 1 diabetes is characterized by a lack of sufficient insulin production, which is caused by an autoimmune attack on pancreatic cells. The onset of type 1 diabetes typically occurs before the age of 20, and hence this variant of the disease has long been referred to as juvenile diabetes. In contrast, type 2 diabetes is usually diagnosed after the age of 30, and thus it is occasionally referred to as adult-onset diabetes. Type 2 diabetes is characterized by reduced insulin sensitivity — the pancreas can still produce insulin, but the body is resistant to it. Type 2 diabetes is much more common than type 1, as it accounts for approximately 90% of all cases of diabetes in the United States [6]. It is also much more common than gestational diabetes (i.e., diabetes during pregnancy), which is sometimes regarded as a third major type of diabetes but which is not considered in this report since it affects such a narrow demographic subpopulation.

Previous studies of diabetes-related mortality have concentrated on the disease's interaction with comorbidities like cardiovascular disease [7] and COVID-19 [8]. While these analyses provide important insights about the associations between diabetes and other causes of death, many of them fail to address the impact of the potentially stark sociodemographic and health-related differences between individuals with type 1 diabetes and those with type 2 diabetes. Some studies investigate just one of the two types [9, 10], while others treat the patient populations for the two types as one homogeneous group [4, 5]. In this report, we fill in the gaps of these prior studies by focusing on the distinction between type 1 and type 2 diabetes in the context of mortality. We use multiple-cause-of-death mortality data from the Centers for Disease Control and Prevention (CDC) [11] to predict whether individuals whose deaths were attributed to diabetes in 2021 suffered from the type 1 variant or the type 2 variant of the disease. We train several different classification models to make these predictions based on the decedents' demographic characteristics and underlying health conditions, and in doing so we aim to compare the mortality risk profiles of individuals across the two types of diabetes.

1.2 Guiding questions

The analysis in this report is centered around the following three questions:

1. **How accurately can sophisticated classification methods distinguish between type 1 and**

type 2 diabetes as a cause of death based on individuals' sociodemographic characteristics and underlying health conditions?

2. **Are there substantial differences in predictive accuracy for this task between model-based, tree-based, and optimization-based classifiers? Which methods achieve the best holistic classification performance?**

3. **Which sociodemographic and health characteristics are the most useful predictors for distinguishing deaths due to type 1 diabetes from those due to type 2 diabetes?**

Our approach to these answering these questions follows two distinct but related tracks. The first track involves unsupervised learning — we examine the social, demographic, and health profiles of those with type 1 and type 2 diabetes and use a dimension reduction technique to construct latent representations of these characteristics. The second track takes a supervised approach — we train six different classification methods to distinguish between deaths due to type 1 and type 2 diabetes, and we evaluate the performance of the resulting models using several different metrics. We implement this classification workflow twice — once using the original predictors from the CDC data set and once using the latent features mentioned above. This allows us to assess whether the latent predictors yield a similar (or perhaps higher) level of classification accuracy compared to the original features.

1.3 Outline

The remainder of this report is organized as follows. In Subsection 2.1, we describe how we assembled the CDC mortality data into a suitable data frame for dimension reduction and classification, with a particular focus on our construction of relevant predictors for the binary response variable `type1diabetes` and our use of synthetic minority oversampling (SMOTE) and random majority undersampling to handle class imbalance in this response variable. In Subsection 2.2, we apply logistic principal component analysis to the data and explore the latent space of our categorical predictors by interpreting the loadings and principal component scores. In Subsection 2.3, we introduce six different methods for classifying `type1diabetes`. We train these methods once using the original features and once using the principal components, and in Section 3 we report the classification performance of both sets of models on unseen test observations in terms of their accuracy, sensitivity, specificity, area under the receiver operating characteristic curve (AUC), and Brier score. Finally, we conclude in Section 4 by summarizing our main contributions, discussing the limitations of our analysis, and proposing potential remedies for these limitations.

2 Methods

2.1 Data preprocessing

The CDC multiple-cause mortality data set contains nearly 3.5 million death records for 2021 [11]. Of these 3.5 million deaths, just 46,206 have a primary International Statistical Classification of Diseases (ICD-10) code of E10 (type 1 diabetes mellitus) or E11 (type 2 diabetes mellitus) [12]. Another 57,218 deaths are attributed to diabetes mellitus but do not distinguish between type 1 and type 2, and hence these records

are unusable for the purposes of this report. Many other records list gestational diabetes as the primary cause of death or cite diabetes as an underlying but non-primary cause, but we do not consider these cases in our analysis. We claim that focusing on deaths for which type 1 or type 2 diabetes is listed as the primary cause is a reasonable approach to assessing mortality risk across the two types, and the number of records that satisfy this criterion is more than sufficient for a robust analysis.

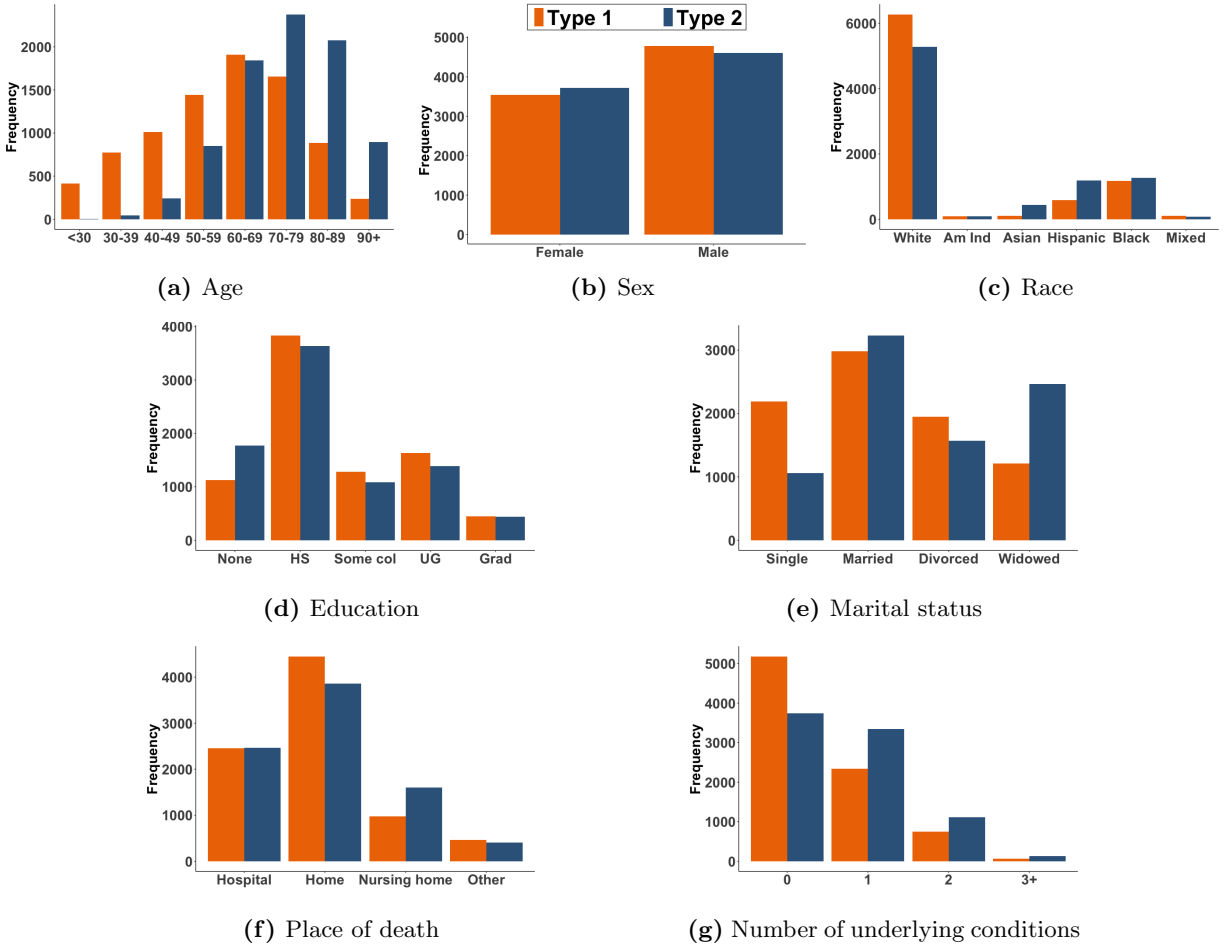
We construct a binary variable called `type1diabetes` based on the ICD-10 codes mentioned above — it takes a value of one for individuals whose primary cause of death is type 1 diabetes and a value of zero for those whose death is attributed to type 2 diabetes. This variable plays a critical role in both tracks of our analysis — for dimension reduction we examine the distributions of our latent predictors across the two classes of `type1diabetes`, and for classification we use it as our response variable.

We also construct ten categorical variables that serve as our observed features for dimension reduction and our predictors for classification: `age`, `sex`, `race`, `education`, `marital_status`, `place_of_death`, `cholesterol`, `covid`, `hypertension`, and `obesity`. The levels of the first six of these variables are listed in Figure 1. The latter four features are binary variables that indicate whether high cholesterol (ICD-10 code E78), COVID-19 (U071), hypertension (I10-I15), and obesity (E66) were listed among an individual's first six record-axis conditions (i.e., underlying health conditions). These four diseases are known to be comorbidities of diabetes [8, 13, 14], so it is reasonable to infer that they carry useful information for predicting `type1diabetes`. Note that the CDC reports up to 20 record-axis conditions for each deceased individual, but we focus only on the second through the seventh of these because the remaining record-axis conditions are missing for more than 95% of cases in the data set.

After filtering out records containing missing values for at least one of the above variables, we obtain a data set with 45,129 observations of `type1diabetes` and our ten predictors. However, as mentioned in Subsection 1.1, we encounter a problem of class imbalance: only 4,165 of these deaths are attributed to type 1 diabetes, while the remaining 40,964 are attributed to type 2. It is well-documented that classification models tend to perform poorly when the class distribution of the response variable is heavily skewed — learning algorithms tend to prioritize the majority class and ignore the minority class in this setting [15]. In our context, for instance, we could achieve roughly 90% classification accuracy by constructing a naive model that classifies all of the records as type 2 and none of the records as type 1. More sophisticated classification methods — including those considered in this report — tend to mimic this behavior in the presence of class imbalance, and hence they produce models that are not particularly useful.

To address the issue of the imbalanced classes in `type1diabetes`, we employ Chawla et al.'s synthetic minority oversampling technique (SMOTE) [16]. This approach generates synthetic data from the minority class by randomly combining the k nearest neighbors of each minority observation. We use SMOTE with $k = 5$ to double the size of our minority class from 4,165 to 8,330 — for each death attributed to type 1 diabetes, we generate one additional synthetic observation. In addition, we randomly undersample the majority class without replacement to achieve perfect balance between the two classes of `type1diabetes` — i.e., we reduce the number of deaths attributed to type 2 diabetes from 40,964 to 8,330. Thus, we obtain a balanced, semi-synthetic data set of 16,660 observations after SMOTE and random majority

Figure 1: Distribution of predictors by class after SMOTE and random majority undersampling



(a) Age **(b)** Sex **(c)** Race

(d) Education **(e)** Marital status

(f) Place of death **(g)** Number of underlying conditions

undersampling. This is the data set that we will use for dimension reduction and classification. We find that SMOTE works as advertised, as we verified that the class-specific distributions of the predictors in the SMOTED data set provide a close approximation to those in the un-SMOTED data set.

In Figure 1, we visualize the distributions of the predictors in the SMOTED data set. Note that in panel (g), we convert the indicator variables `cholesterol`, `covid`, `hypertension`, and `obesity` into a single numeric variable by computing their sum.

2.2 Dimension reduction

In Figure 1, we observe several prominent differences in the distributions of the predictors between individuals who died from type 1 diabetes and those who died from type 2 diabetes. Those who died from type 1 diabetes appear to younger and more educated on average, with fewer of the four comorbidities that contribute to the graph in panel (g). Individuals in the type 2 class appear to be more likely to be widowed and die in a nursing home, while those in the type 1 class are more likely to be single and die in their own home. These differences suggest that the original features in our data set carry at least some predictive power for the task of classifying `type1diabetes`.

However, it is possible that we could describe a similar amount of sociodemographic and health-related

Table 1: Loadings of the first three principal components

	Loading 1		**Loading 2**		**Loading 3**
Place of death: Other	0.415	Marital status: Married	0.497	Place of death: Nursing home	0.443
Place of death: Nursing home	0.387	Sex: Male	0.486	Sex: Male	0.152
Hypertension: Yes	0.232	Hypertension: Yes	0.188	Age: Thirties	0.097
Education: Graduate	0.208	Education: Graduate	0.174	Marital status: Married	0.082
Marital status: Widowed	0.189	Race: Asian	0.114	Education: High school	0.074
Place of death: Home	-0.508	Marital status: Widowed	-0.395	Hypertension: Yes	-0.750
Education: High school	-0.335	Marital status: Divorced	-0.360	High cholesterol: Yes	-0.324
Sex: Male	-0.318	Education: High school	-0.304	Place of death: Home	-0.241
Age: Forties	-0.155	Place of death: Home	-0.072	Marital status: Widowed	-0.115
Age: Thirties	-0.105	Age: Seventies	-0.058	Age: Seventies	-0.064

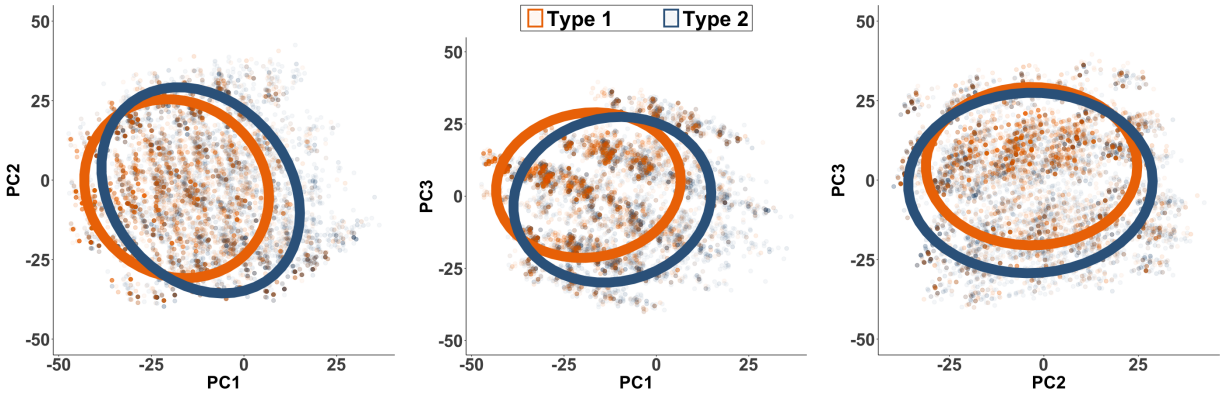
variation in the predictor space with a smaller number of latent representations of the original features. This is the concept that motivates data reduction techniques such as principal component analysis (PCA), which aims to identify a low-dimensional encoding of the original predictors that captures a considerable proportion of the variation of the observed data. We cannot directly apply the traditional version of PCA to our data since all of our predictors are categorical, as PCA assumes that the observed variables follow a continuous distribution. Fortunately, there exist extensions of PCA to binary data — instead of finding the low-dimensional latent representation that maximizes the deviance of some continuous distribution (e.g., Gaussian), we aim to maximize the Bernoulli deviance.

After encoding our ten categorical predictors as 27 dummy variables, we implement PCA on these binary features using the `logisticPCA` package in `R` [17]. We select 12 principal components, which is the smallest number of components that explains at least 95% of the deviance of the observed data. Table 1 reports the five largest and five smallest loadings of each of the first three principal components — these are the weights that describe the contributions of the original predictors to the latent components. Figure 2 plots the distribution of scores for each pair of the first three principal components, with the scores labeled by the two classes of `type1diabetes`.

For the first principal component, positive loadings seem to capture characteristics that are more associated with type 2 diabetes, while negative loadings seem to correspond to type 1 diabetes. In particular, the first principal component appears to separate deaths that occurred in the decedent's home from those that occurred in nursing homes or elsewhere, which is one of the differences between the two classes that we observed in Figure 1. This interpretation is also borne out in panel (a) of Figure 2, as we observe that the orange ellipse corresponding to type 1 diabetes is shifted slightly in the negative direction of the first principal component compared to the navy type 2 ellipse.

The second principal component captures variation in marital status and education — positive loadings correspond to married individuals with a graduate education, while negative loadings correspond to widowed and divorced individuals with a high school education. The third component seems to be based primarily on hypertension and cholesterol, although it also captures discrepancies in place of death, sex, marital status, and age. It appears that the second and third principal components draw less of a distinction between the two classes of `type1diabetes` — in Figure 2, the orange and navy ellipses overlap almost entirely in the second and third principal component directions.

Figure 2: Scores for the first three principal components, labeled by type of diabetes
**Orange and navy ellipses are 80% confidence bivariate Gaussian ellipses)*



(a) PC1 vs. PC2 **(b)** PC1 vs. PC3 **(c)** PC2 vs. PC3

Note that Table 1 and Figure 2 involve only the first three principal components, which is far fewer than the twelve components that we will use to train the classifiers described below. That being said, the first three principal components appear to establish very little separation between deaths due to type 1 and type 2 diabetes, which casts doubt on whether using these latent representations as predictors will lead to comparable accuracy in classifying `type1diabetes` as the original features themselves.

2.3 Classification

Recall that our balanced data set contains 16,660 observations of ten categorical predictors and the binary response variable `type1diabetes`. We construct a second version of this data set in which we replace the ten original features (which can be encoded as 27 dummy variables) with the 12 principal components from Subsection 2.2. We divide each version of the data set into a training set of 13,328 observations (80%) and a test set of 3,332 observations (20%), using the same random split for both versions.

We consider six different methods for classifying `type1diabetes`. Three of these classifiers are model-based (naive Bayes, quadratic discriminant analysis, and penalized logistic regression), two are tree-based (random forest and AdaBoost), and one is optimization-based (kernel SVM). None of these three families of classifiers is obviously better suited for our classification task than the others — if such an advantage exists, we hope to uncover it in our evaluation of these models. Below, we briefly summarize our approach to fitting these six methods on each version of the training set.

2.3.1 Naive Bayes

The first of the model-based classifiers is naive Bayes [18], a generative approach that requires no parameter tuning and imposes two simplifying assumptions on the predictors: (i) they are conditionally independent given the response variable, and (ii) their marginal distributions are categorical. Naive Bayes is well-suited for the original features from the CDC data set, which are all categorical. However, it can be extended in a fairly straightforward manner to accommodate continuous predictors — we simply assume a multivariate Gaussian marginal distribution for the predictors instead of a categorical distribution. We use the classical

version of naive Bayes to train a model using the original features, and we use the Gaussian extension to fit a model on the principal components.

2.3.2 Quadratic discriminant analysis

Quadratic discriminant analysis (QDA) [19] is another generative approach to model-based classification that is very similar to the Gaussian extension of naive Bayes. In fact, QDA can be viewed as a generalization of Gaussian naive Bayes — the only difference between the two methods is that QDA drops the conditional independence assumption on the predictors. Specifically, it models the conditional distribution of the predictors given the response variable as multivariate Gaussian with some mean and (potentially non-diagonal) covariance matrix, both of which are allowed to vary across the classes of the response variable.

There are two key remarks to be made about QDA in the context of our classification task. First, since QDA imposes a Gaussian assumption on the predictors, we cannot apply it to the version of our data set that contains the original categorical features. Second, since QDA and Gaussian naive Bayes differ only by a conditional independence assumption on the predictors, we can assess the plausibility of this assumption for the principal components by comparing the classification performances of the two methods. We circle back to this remark in Subsection 3.2.

2.3.3 Penalized logistic regression with elastic net regularization

The third model-based classification method we consider is penalized logistic regression with elastic net regularization [20]. For simplicity, we refer to this classifier as "elastic net" in Table 2 and Figure 4. Unlike naive Bayes and QDA, penalized logistic regression is a discriminative approach that directly models the conditional distribution of the response variable given the predictors without explicitly specifying the marginal distribution of the predictors. Penalized logistic regression also differs from the previous two model-based classifiers in that it requires parameter tuning. We use 10-fold cross-validation on each version of the training set to tune the shrinkage parameter λ and the LASSO-ridge mixing parameter α. We obtain optimal values $(\widehat{\lambda} = 0.0005, \widehat{\alpha} = 0.9)$ and $(\widetilde{\lambda} = 0.0856, \widetilde{\alpha} = 0.1)$, respectively, for the version with the original features and the version with the principal components.

2.3.4 Random forest

We now consider two tree-based models that offer a flexible alternative to model-based classification. The first of these tree-based classifiers is random forest [21], an extension of bagging that grows (and ultimately aggregates) a large number of decision trees on bootstrapped versions of the training data while considering a limited number of predictors m at each split. Limiting the number of predictors considered at each split reduce the correlation between the trees and tends to yield classifiers with lower variance, but it also introduces a tuning parameter. We use 10-fold cross-validation to tune m on each version of the training set. We obtain optimal values of $\widehat{m} = 3$ for the version with the original features and $\widetilde{m} = 9$ for the version with the principal components.

2.3.5 AdaBoost

The second tree-based method we consider is AdaBoost [22], a boosting algorithm that fits a sequence of so-called weak classifiers on iteratively reweighted versions of the data and outputs a weighted combination of the predictions of these weak classifiers. Boosting algorithms like AdaBoost are known to be well-suited for classifying difficult training examples that lie close to the decision boundary [23]. As such, we expect AdaBoost to perform relatively well on the principal component version of our data set, as we observed a substantial amount of overlap between the classes of `type1diabetes` in Figure 2. We run 200 iterations of AdaBoost on each version of the training set, growing each tree to a maximum depth of 30 nodes and pruning it with a cost-complexity parameter of 1e-6.

2.3.6 Kernel SVM

The final method we consider is a support vector machine (SVM) with a radial basis kernel [24]. Kernel SVM solves a constrained optimization problem — it projects the observed data onto a reproducing kernel Hilbert space and attempts to separate the classes of the response variable in this higher-dimensional space by maximizing the margin around the decision boundary. Kernel SVM is predominantly used in settings with continuous predictors, so it is arguably more suitable for the principal component version of our data set than the version with the original categorical features. However, we find that the method achieves a relatively strong classification performance using the categorical features if we encode these predictors as dummy variables. As such, we apply kernel SVM to both versions of the training set and report both sets of results in Section 3. We use 10-fold cross-validation on each version of the training set to tune the cost parameter, which controls the amount of constraint violation that is allowed when maximizing the margin. We obtain optimal cost values of 100 and 10, respectively, for the version with the original features and the version with the principal components.

3 Results

After fitting the classifiers from Subsection 2.3 on both versions of the training data, we evaluate their classification performances by using them to predict `type1diabetes` on the corresponding version of the test set. To ensure a comprehensive assessment of these methods, we consider five different performance metrics. Classification accuracy is just the proportion of test cases for which the cause of death (type 1 or type 2 diabetes) is correctly identified, assuming the usual decision threshold of 0.5. Sensitivity is the proportion of correctly identified type 1 deaths, while specificity is the proportion of correctly identified type 2 deaths. Area under the ROC curve (AUC) holistically evaluates classification accuracy across a range of decision thresholds between zero and one — a ROC curve plots the false positive rate (1 - specificity) against the true positive rate (sensitivity) at each threshold, and AUC is just the area under this curve. Finally, Brier scores provide a more direct evaluation of the probabilistic predictions of our classification models — a lower Brier score indicates higher accuracy, and vice versa [25]. This metric is computed as $\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{p}_i)^2$, where N is the number of observations in the test set, $y_i \in \{0, 1\}$ is the actual value of `type1diabetes` for the ith test observation, and $\hat{p}_i$ is the predicted probability of a type 1 death for the

Table 2: Performance of classifiers trained on original features and principal components
**Underline identifies the best-performing method(s) for each metric*

	Original features					Principal components				
	Accuracy	**Sens.**	**Spec.**	**AUC**	**Brier**	**Accuracy**	**Sens.**	**Spec.**	**AUC**	**Brier**
Naive Bayes	0.672	0.640	0.704	0.745	0.206	0.654	0.674	0.635	0.714	0.217
QDA	—	—	—	—	—	0.667	0.667	0.667	0.725	0.219
Elastic net	0.688	0.646	0.729	0.750	0.201	0.630	0.731	0.531	0.704	0.225
Random forest	0.709	0.674	0.743	0.784	0.202	0.697	0.686	0.707	0.761	0.221
AdaBoost	0.695	0.705	0.686	0.769	0.201	0.694	0.702	0.687	0.766	0.200
Kernel SVM	0.703	0.641	0.762	0.766	0.198	0.700	0.694	0.707	0.758	0.199

ith test observation. Table 2 reports these five performance metrics for both sets of classification models — those trained on the original features and those trained on the principal components.

3.1 Classification using original features

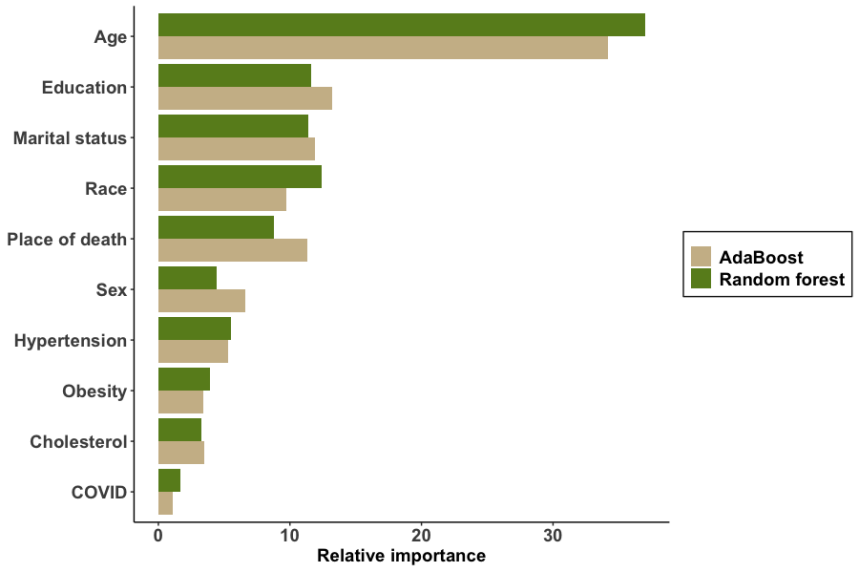
We first examine the performance of the classifiers trained on the original categorical predictors. The five methods perform similarly overall, although **random forest**, **AdaBoost**, and **kernel SVM** hold a slight advantage over naive Bayes and elastic net across all five metrics. Random forest achieves the highest accuracy and AUC, kernel SVM achieves the lowest Brier score and highest specificity, and AdaBoost achieves the highest sensitivity. The weaker performance of the model-based classifiers relative to the tree-based and optimization-based methods suggests that the assumptions imposed by naive Bayes and logistic regression may not hold for the CDC mortality data. In particular, the naive Bayes assumption of conditionally independent predictors is potentially suspect — there may be a nontrivial amount of association in each class between at least some of the predictors.

Recall the third guiding question from Subsection 1.2 — it is of interest to know which of the features in our data set carry the most predictive information for distinguishing deaths due to type 1 diabetes from those due to type 2 diabetes. To address this objective, we extract variable importance scores from the tree-based classifiers trained on the original features. Figure 3 reports the relative importance of the ten categorical predictors according to random forest and AdaBoost. These relative importance scores reflect the average change in the Gini index when a particular variable is added or removed from a tree. We standardize these two sets of variable importance scores so that they fall between 0 and 100, and we find that the two sets of rankings generally agree with one another. In particular, the tree-based methods suggest that **age** is by far the most important predictor for classifying `type1diabetes`, followed in some order by **education**, **marital status**, **race**, and **place of death**. The four comorbidities are found to carry the least amount of predictive information about mortality among individuals with diabetes.

3.2 Classification using principal components

We now turn our attention to the classifiers trained on the principal components. We observe essentially the same trends described in Subsection 3.1, as **random forest**, **AdaBoost**, and **kernel SVM** generally

Figure 3: Relative importance of original predictors according to random forest and AdaBoost



perform better across the five metrics than the model-based classifiers. Interestingly, the elastic net model achieves the highest sensitivity of the six methods, but its specificity is almost as low as that of a random classifier. This result is a bit puzzling, particularly because we found that it holds for a wide range of values of the elastic net shrinkage and mixing parameters. Elsewhere among the model-based classifiers, we find that QDA performs slightly better than Gaussian naive Bayes across almost all of the metrics. Recall from Subsubsection 2.3.2 that these two methods differ only by the naive Bayes assumption of conditional independence among the predictors. Thus, our results indicate that this assumption may not hold for the principal component version of the data set.

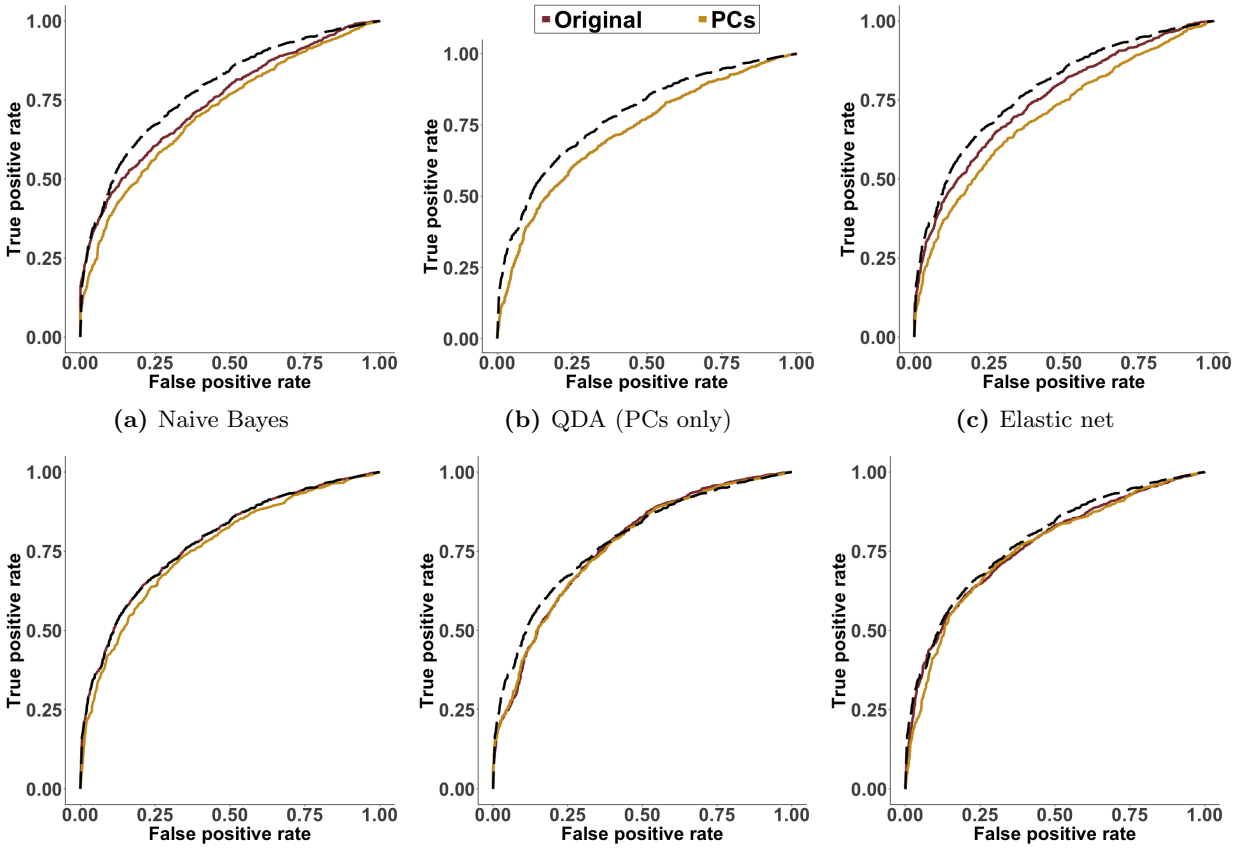
Arguably the most notable takeaway regarding the classifiers trained on the principal components is that they do not perform as strongly as the models that use the original features. Our findings suggest that using the principal components as predictors yields slightly worse classification performance across all five metrics compared to the models trained on the original features. The exception to this is AdaBoost, which performs almost exactly the same with the principal components as it does with the original features; this is consistent with our hypothesis from Subsubsection 2.3.5. We visualize the differences in performance between the two sets of models in Figure 4 by plotting the ROC curves for both versions of each classifier. The slight advantage of random forest, AdaBoost, and kernel SVM over the model-based classifiers is visually apparent in this figure, as the ROC curves in panels (d)-(f) adhere more closely to the best ROC curve (achieved by random forest using the original predictors) than those in panels (a)-(c).

4 Discussion

4.1 Summary of main contributions

In this report, we have applied a variety of unsupervised and supervised statistical learning methods to CDC multiple-cause mortality data from 2021 with the objective of assessing the determinants of mortality

Figure 4: ROC curves for the six classifiers using original features and principal components
**Black dashed line is the best ROC curve (achieved by random forest on original predictors)*



(a) Naive Bayes **(b)** QDA (PCs only) **(c)** Elastic net

(d) Random forest **(e)** AdaBoost **(f)** Kernel SVM

among individuals with type 1 and type 2 diabetes. We used dimension reduction and classification techniques to address the three guiding questions laid out in Subsection 1.2 — i.e., we examined (1) the accuracy with which classification methods can distinguish between type 1 and type 2 diabetes as a cause of death, (2) the similarities and differences in predictive performance between model-based, tree-based, and optimization-based classifiers, and (3) the sociodemographic and health characteristics that play the most important roles in separating the two classes of the response variable `type1diabetes`. By focusing on the often overlooked distinction between type 1 and type 2 diabetes, we have expanded upon previous studies and contributed new insights about the mortality risk profiles of diabetic individuals.

Our results suggest that distinguishing between deaths attributed to type 1 and type 2 diabetes is a difficult task, as even the best-performing methods in Section 3 misclassify roughly three out of ten cases. The challenges encountered by these classifiers are not too surprising, as we observed a nontrivial amount of class overlap in the predictor space for both the original features (Figure 1) and the principal components (Figure 2). However, all of the classifiers considered in this report would outperform a random classifier (i.e., one that achieves 50% accuracy) by a considerable margin. As such, our findings indicate that sophisticated classification methods can distinguish between type 1 and type 2 diabetic mortality with moderate accuracy based on individuals' sociodemographic and health profiles.

The six classification methods considered in this report achieve similar performance overall, although

random forest, AdaBoost, and kernel SVM outperform naive Bayes, QDA, and penalized logistic regression by a slight margin in terms of their accuracy, sensitivity, specificity, AUC, and Brier score. This advantage holds for both sets of models constructed in Section 2 — those trained on the original categorical features and those that use the principal components. These findings suggest that tree-based and optimization-based classifiers are better equipped to predict `type1diabetes` than model-based classifiers. We hypothesize that the shortcomings of the model-based classifiers may arise because the distributional assumptions that they impose on the predictors and the response do not hold for the CDC mortality data.

Finally, regarding the third guiding question from Subsection 1.2, we find that age seems to be the most useful predictor for the task of differentiating deaths due to type 1 and type 2 diabetes. Age dominates the other predictors in the tree-based classifiers trained on the original features (Figure 3), and it is also an important contributor to the loadings of the first three principal components (Table 1). Sociodemographic characteristics like education, marital status, race, place of death, and sex are also important contributors to the tree-based classifiers and the first few principal components, while underlying health conditions like high cholesterol, COVID-19, hypertension, and obesity seem to wield little predictive power that is not already captured by the other features.

4.2 Limitations and potential remedies

While the analysis presented in this report addresses all of our stated objectives, it is not without its shortcomings. One such limitation is the large number of unlabeled diabetes-related deaths in the CDC mortality data set. As mentioned in Subsection 2.1, we found it necessary to omit the nearly 60,000 diabetes-related deaths in this data set that do not specify whether the individual died from the type 1 or type 2 variants of the disease. We could have doubled our sample size and enriched the predictor space by including these cases in our analysis, but unfortunately the lack of a specified type for these records renders them useless for supervised learning. One could utilize missing value imputation techniques to handle these unlabeled records [26], but this is a potentially risky strategy since the number of unlabeled cases in the data set is greater than the number of labeled cases. Alternatively, one could incorporate past years of CDC mortality data to bolster the number of labeled records [11]. This would be particularly helpful because it would provide more training examples for the minority class of type 1 deaths.

A second hurdle that we encountered in our analysis is the absence of more detailed medical information in the CDC data set. It is unrealistic to expect individual health records to be included in a public use data set, but this information would be invaluable for the dimension reduction and classification tasks in this report. In particular, measures of glycemic variability such as hemoglobin A1C (HbA1C) are known to be strong predictors of mortality among individuals with type 1 and type 2 diabetes [7, 27], and there tends to be greater variability in these indicators among those with type 1 diabetes than those with type 2 diabetes [28]. We infer that including these diabetes-specific indicators as features in our data set would greatly improve the predictive performance of our classification models. Future analyses that aim to predict type 1 and type 2 diabetic mortality would likely benefit from merging the CDC mortality data with other data sources such as electronic health records.

References

[1] Centers for Disease Control and Prevention. *Leading Causes of Death, 1900-1998*. 2015. URL: https://www.cdc.gov/nchs/nvss/mortality_historical_data.htm.

[2] Centers for Disease Control and Prevention. *Mortality tables*. 2017. URL: https://www.cdc.gov/nchs/nvss/mortality_tables.htm.

[3] Jiaquan Xu et al. *Mortality in the United States, 2021*. 2022. DOI: 10.15620/cdc:122516.

[4] Joanna Buscemi et al. "Diabetes mortality across the 30 biggest U.S. cities: Assessing overall trends and racial inequities". In: *Diabetes Research and Clinical Practice* 173.108652 (2021). DOI: 10.1016/j.diabres.2021.108652.

[5] Fatima Rodriguez et al. "Diabetes-attributable mortality in the United States from 2003 to 2016 using a multiple-cause-of-death approach". In: *Diabetes Research and Clinical Practice* 148 (2019). DOI: 10.1016/j.diabres.2019.01.015.

[6] J. C. Ozougwu et al. "The pathogenesis and pathophysiology of type 1 and type 2 diabetes mellitus". In: *Journal of Physiology and Pathophysiology* 4.4 (2013). DOI: 10.5897/jpap2013.0001.

[7] Sridharan Raghavan et al. "Diabetes mellitus-related all-cause and cardiovascular mortality in a national cohort of adults". In: *Journal of the American Heart Association* 8.4 (2019). DOI: 10.1161/JAHA.118.011295.

[8] Bruce Bode et al. "Glycemic characteristics and clinical outcomes of COVID-19 patients hospitalized in the United States". In: *Journal of Diabetes Science and Technology* 14.4 (2020). DOI: 10.1177/1932296820924469.

[9] Mauro Tancredi et al. "Excess mortality among persons with type 2 diabetes". In: *The New England Journal of Medicine* 373.18 (2015). DOI: 10.1056/NEJMoa1504347.

[10] Christopher C. Patterson et al. "Worldwide estimates of incidence, prevalence and mortality of type 1 diabetes in children and adolescents: Results from the International Diabetes Federation Diabetes Atlas, 9th edition". In: *Diabetes Research and Clinical Practice* 157.107842 (2019). DOI: 10.1016/j.diabres.2019.107842.

[11] Centers for Disease Control and Prevention. *Mortality Multiple Cause-of-Death Public Use Record*. 2023. URL: https://www.cdc.gov/nchs/data_access/vitalstatsonline.htm.

[12] Centers for Disease Control and Prevention. *ICD-10 tabular list, 2022*. 2022. URL: https://www.cdc.gov/nchs/nvss/manuals/2022/2e_volume1_2022.htm.

[13] Denis Daneman. "Type 1 diabetes". In: *The Lancet* 367.9513 (2006). DOI: 10.1016/S0140-6736(06)68341-4.

[14] Kristy Iglay et al. "Prevalence and co-prevalence of comorbidities among patients with type 2 diabetes mellitus". In: *Current Medical Research and Opinion* 32.7 (2016). DOI: 10.1185/03007995.2016.1168291.

[15] Ronaldo C. Prati, Eapa Gustavo, and Maria Carolina Batista. "Data mining with imbalanced class distributions: concepts and methods". In: 4th Indian International Conference on Artificial Intelligence, 2009. URL: http://sites.labic.icmc.usp.br/pub/mcmonard/PratiIICAI09.pdf.

[16] N. V. Chawla et al. "SMOTE: Synthetic minority over-sampling technique". In: *The Journal of Artificial Intelligence Research* 16 (2002). DOI: 10.1613/jair.953.

[17] Andrew Landgraf. *logisticPCA: Binary Dimensionality Reduction*. 2016. URL: https://cran.rstudio.com/web/packages/logisticPCA/index.html.

[18] Michal Majka. *naivebayes: High Performance Implementation of the Naive Bayes Algorithm*. 2020. URL: https://cran.rstudio.com/web/packages/naivebayes/index.html.

[19] Brian Ripley et al. *MASS: Support Functions and Datasets for Venables and Ripley's MASS*. 2023. URL: https://cran.rstudio.com/web/packages/MASS/index.html.

[20] Jerome Friedman et al. *glmnet: Lasso and Elastic-Net Regularized Generalized Linear Models*. 2023. URL: https://cran.rstudio.com/web/packages/glmnet/index.html.

[21] Andy Liaw. *randomForest: Breiman and Cutler's Random Forests for Classification and Regression*. 2022. URL: https://cran.rstudio.com/web/packages/randomForest/index.html.

[22] Esteban Alfaro, Matias Gamez, and Noelia Garcia. *adabag: Applies Multiclass AdaBoost.M1, SAMME and Bagging*. 2018. URL: https://cran.rstudio.com/web/packages/adabag/index.html.

[23] Peter Bartlett et al. "Boosting the margin: a new explanation for the effectiveness of voting methods". In: *Annals of Statistics* 26.5 (1998). DOI: 10.1214/aos/1024691352.

[24] David Meyer et al. *e1071: Misc Functions of the Department of Statistics, Probability Theory Group*. 2023. URL: https://cran.rstudio.com/web/packages/e1071/index.html.

[25] Kaspar Rufibach. "Use of Brier score to assess binary predictions". In: *Journal of Clinical Epidemiology* 63.8 (2010). DOI: 10.1016/j.jclinepi.2009.11.009.

[26] Wei-Chao Lin and Chih-Fong Tsai. "Missing value imputation: a review and analysis of the literature (2006–2017)". In: *Artificial Intelligence Review* 53.2 (2020). DOI: 10.1007/s10462-019-09709-4.

[27] Stuart S. Wightman, Christopher A. R. Sainsbury, and Gregory C. Jones. "Visit-to-visit HbA1c variability and systolic blood pressure (SBP) variability are significantly and additively associated with mortality in individuals with type 1 diabetes: An observational study". In: *Diabetes, Obesity & Metabolism* 20.4 (2018). DOI: 10.1111/dom.13193.

[28] Masafumi Koga et al. "Comparison of annual variability in HbA1c and glycated albumin in patients with type 1 vs. type 2 diabetes mellitus". In: *Journal of Diabetes and its Complications* 27.3 (2013). DOI: 10.1016/j.jdiacomp.2012.12.001.