Predicting Diabetes Diagnosis Based on Demographics, Socio-Economic Status, and Dietary Factors

Nicole Nagata

Seattle University DATA 5322 02

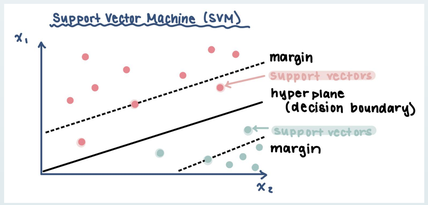
Abstract

The following study will use data from the National Health Interview Survey and various support vector machine (SVM) models to predict whether an adult has been diagnosed with diabetes (1). According to Medical News Today, the most common age to get a diabetes diagnosis is between 45-65, so this study will mainly focus on this demographic (2). This study will explore various factors that could affect someone with diabetes like demographics, socio-economic status, body compositions, health, diet, and more.

Technical Background

This study will utilize various support vector machine models with applying different kernels to analyze health survey data and predict a diabetes diagnosis. **Support Vector Machine (SVM) Models** are classification machine learning models, which aim to find the optimal hyperplane that separates classes with the maximum margin. When data can not be separated linearly, SVMs use kernel functions to project data to higher dimensions, enabling separation by a linear hyperplane in that space.

1. **Linear SVM Model**: utilizes a linear kernel (dot product), equivalent to a support vector classifier. The *cost parameter* controls the trade-off between maximizing the margin and minimizing classification error.
2. **Polynomial SVM Model**: applies a polynomial kernel to allow non-linear decision boundaries. The *degree parameter* determines the polynomial's complexity, where higher degree allows for more curvature.
3. **Radial SVM Model**: uses the radial kernel to map data to an infinite-dimensional space, enabling highly flexible boundaries. The *gamma parameter* controls how far the influence of a training example spans, lower values imply broad influences and vice versa for high values.



Methodology

The data for this study comes from the National Health Interview Survey and is accessed through the IPUMS Health Survey (1). The data is imported and filtered for adults who are between the ages of 45-65; this subset is targeting the most common age that adults get diagnosed with diabetes. To clean the data, the target variable (diabetes diagnosis) is recoded into 1 for 'Yes' diabetes and 0 for 'No' diabetes and changed into a factor, the appropriate data type for an SVM classification model.

All initial SVM models are implemented on subset data based on specified predictor variables like socio-economic, body and health, and dietary consumption. All variables are cleaned by removing codes (ie. 99, 990) and filtered to a realistic range (ie. Height between 50 – 84 inches). All subsets are under-sampled to address the class imbalance, randomly selecting 1,000 data points for 'No' diabetes and 1,000 points for 'Yes' diabetes. Initial models are run with a 70/30 train test split and analyzed using training error, test error, confusion matrix, and ROC curves. Then tuned using cross validation and ranges for cost, degree, and gamma depending on kernel used.

References

[1] Lynn A. Blewett, Julia A. Rivera Drew, Miriam L. King, Kari C.W. Williams, Daniel Backman, Annie Chen, and Stephanie Richards. IPUMS Health Surveys: National Health Interview Survey, Version 7.4 [dataset]. Minneapolis, MN: IPUMS, 2024. https://doi.org/10.18128/D070.V7.4. Links to an external site.http://www.nhis.ipums.orgLinks to an external site..

[2] Huizen, J., & Buggers, A. (2025, March 4). Type 2 diabetes: Average age of onset, risk factors, prevention. Medical News Today. https://www.medicalnewstoday.com/articles/317375

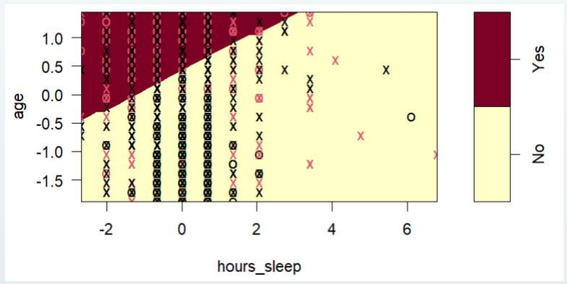
Linear SVM Model: Diabetes ~ Socio-Economic

The first model is a linear SVM model that predicts whether an adult (ages 45-65) has diabetes given certain demographic and socio-economic status like age, sex, martial status, education level, poverty, hours worked, and hours slept. The first model has low test errors, but the confusion matrix reveals that the data is heavily imbalanced, and the model is only predicting the majority class (no diabetes). To improve the linear SVM model, the final iteration reduces and balances the training data to 2,000 data points with even target variable classes, as well as runs a cross validation testing different cost tuning parameters from 1 to 100.

The optimal model utilizes 1,168 significant points in the training data, with 584 support vectors for no diabetes and 584 support vectors for a diabetes diagnosis. This model has similar training and test errors of about 40%, which tells us the model generalizes the data well. However, the confusion matrix shows that the model typically misclassifies non-diabetes diagnoses. A linear SVM model may not be robust enough to capture underlying socio-economic relationships, but graphs can provide some general insights towards factors that may influence diabetes.

true	pred No	pred Yes
No	182	86
Yes	171	161

Fig 1. Linear SVM Plot: Age vs Hours Slept



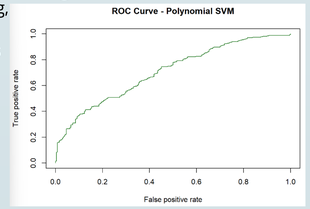
This plot graphs age and hours of sleep, with the SVM model separating the diabetes classes. Older individuals who sleep less are more likely to be classified has having diabetes. In this case anyone who is about 50-65 and sleeping less than 7 hours has a higher chance of a diabetes diagnosis.

Results

Performance Metrics	Linear SVM	Polynomial SVM	Radial SVM
Training Error	0.381	0.342	0.369
Test Error	0.405	0.364	0.468

Based on the performance metrics, the polynomial SVM model, in combination with the body and health predictor variables, had the best predictive power of a diabetes diagnosis, with the lowest test error overall at about 36%. The polynomial SVM model likely performs the best on this dataset, as it balances flexible boundaries with complex relationships among the predictor and target variables. Although this test error is not amazing, it does a good enough classification job with confidence to analyze factors that could impact diabetes in adults. The ROC curve shows that the poly-nomial model can be improved, but is starting to take shape and hug the corner which proves this is a fair model and classifier of diabetes.

Fig 4. ROC: Polynomial SVM



The Linear SVM model is the next best model, as it may not capture the nuances of complex relationships among the variables but does predict a diabetes diagnosis relatively well. This model is simple and can save computational power, while also providing easily explainable graphs.

The poor performance of the Radial SVM model likely implies that the model is too flexible. The largest gap, almost ten percent, between the training and test error tells us the model is overfitting and does not generalize well.

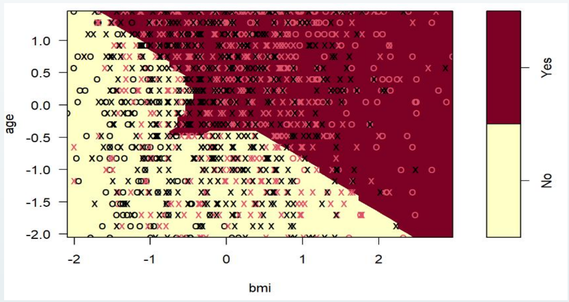
Polynomial SVM Model: Diabetes ~ Body + Health

A polynomial SVM model will be used to predict whether an adult (ages 45-65) has diabetes given certain body compositions and health habits like age, weight, sex, BMI, alcohol intake, and smoke intake. A polynomial model will be used as these variables might have a non-linear or nuanced relationship with a blood sugar diagnosis like diabetes. When testing the model, increasing the degrees improve the models test errors which confirms the complex relationship with diabetes.

The final model will under sample to 2,000 data points, as well as cross validate degrees of 2, 3, and 4, and costs between 0.01 and 100. The optimal model has a degree of 3 and a cost of 1. 1,000 points are significant in predicting a diabetes diagnosis, with 498 no diabetes support vectors and 502 being a diabetes diagnosis support vector. The test error is about 36%, which is a better prediction than the linear SVM model, implying more nuanced relationships than initially explored. The confusion matrix shows a more balanced prediction, as this polynomial model does a better job at predicting a diabetes diagnosis based on body and health factors.

true	pred No	pred Yes
No	164	96
Yes	95	169

Fig 2. Polynomial SVM Plot: BMI vs Age



This plot graphs an adults age (between 45-65) and their BMI (10-70). Based on this plot, adults who are on the older side of the 45-65 age range and have a higher BMI are more likely to be diagnosed with diabetes. This may imply that keeping track of a healthy weight and height ratio could help predict a diabetes diagnosis or start preventative measures earlier.

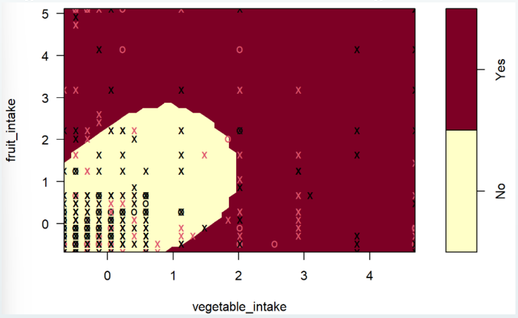
Radial SVM Model: Diabetes ~ Dietary Habits

Lastly, a radial SVM model will be used to predict whether an adult between 45-65 has diabetes given certain dietary habits like fruit, vegetables, juice, salad, soda, pizza, and fried potato intake. This model will attempt to explain the complex relationships between diet and diabetes, as the radial kernel allows for highly flexible boundaries.

The final radial SVM model will include 2,000 under-sampled training data points, utilizing cross validation and testing optimal costs between 0.01 to 50 and gamma values between 0.001 and 1. The optimal radial model has a cost of 5 and a gamma value of 0.1. 1,287 points are significant with 640 no diabetes support vector and 647 diabetes diagnosis support vectors. The test error is significantly higher (+10%) than the training error, which tells us this model does not generalize well.

In case a radial kernel is too complex, a linear SVM model was tested to predict diabetes using dietary habits. The results are also a poor test error (47%), just slightly better than guessing. This likely implies that dietary habits does not necessarily have a strong impact in predicting diabetes.

Fig 3. Radial SVM Plot: Fruit Intake vs Vegetable Intake



Since the Radial SVM model was the best model that predicted diabetes using dietary habits, we can use a plot to visualize what we know but take this information with a grain if salt. This plot graphs fruit intake versus vegetable intake. We can see that adults who eat an excess of both (above average) is more likely to be diagnosed with diabetes. We might relate this to figure 2, where we found adults to might weight more (or have a higher BMI) are also more likely to be diagnosed with diabetes.

Discussions

The linear SVM model considers various demographic and socio-economic factors that may impact a diabetes diagnosis for adults ages 45-65. This model is simple and may loose some of the nuances between the relationships of the predictor and target variables, but nevertheless does a fair job at predicting a diabetes diagnosis. Figure 1 implies that adults who are around fifty to sixty years old and gets less than 7 hours of sleep have a higher chance of a diabetes diagnosis. Similar graphs (see full code for additional graphs) show that this model utilizes age as a strong indicator in predicting a diabetes diagnosis, and socio-economic factors like marital status or hours worked have less of an impact. This simple linear SVM model highlights the importance of preventative measures and health awareness as adults near the fifty- to sixty-year-old range, as this increases changes of a diabetes diagnosis. Small steps like improving sleep can likely help be preventative.

The polynomial SVM model considers body composition and health habits to predict a diabetes diagnosis in adults ages 45-65. The plot compares age again BMI (body mass index, which is calculated using one's height and weight). The relationship between these two variables are not linear, but may be separated by a more complex class division like a polynomial. Generally based on this plot, adults who are on the older side of the 45-65 age range and have a higher BMI are more likely to be diagnosed with diabetes. This can imply that as adults near their fifty's, being aware of their height and weight ratio can provide insights and inform healthier eating and living habits throughout later adult life. This polynomial model implies that a healthier and more balanced lifestyle can be one step towards preventing a diabetes diagnosis later on in life.

The radial SVM model explores various dietary habits and how they impact diabetes in later adult life. This model generally does not perform well, only predicting diabetes slightly better than guessing, likely due to the overly complex model and overfitting on the training data. However, the same dietary predictor variables were used to train a linear SVM model for simplicity, and found even worse results. This may imply that dietary habits do not necessarily inform a diabetes diagnosis.

However, since the radial model was best at capturing the nuances of dietary habits and diabetes we might be able to gain some insights. Figure 3 compares fruit and vegetable intake, with the radial model showing that any excess consumption (above an average daily intake) results in a likely diagnosis. While taking this information with a grain of salt, this does align with Figure 2 that implies that adults should watch their BMI. Eating a healthy and reasonable amount could be preventative measures to avoid diabetes.

Conclusions

For broader implications, it is important to spread awareness for diabetes, as 14.7% of adults in the United States have diabetes (2). This study shows that the most direct lifestyle change an adult can make nearing their fifty's is implementing a healthy and balanced lifestyle. It may be worth while to implement healthy habits like monitoring amount of sleep with a goal of reaching 8 hours per night, tracking healthy weight and height ratios, and promoting a balanced diet to prevent diabetes diagnosis. This is a common diagnosis for many adults, but signs can be noticed earlier, and prevention can start before and major progression occurs.