Title: Analyzing the Early Stage of Diabetes Risk

Author: Mei Ying Tan

Abstract: The number of the population with diabetes rose from 108 million in 1980 to 422 million in 2014 and the majority of the people are not aware of the warning signs of prediabetes. In this paper, the data was collected based on the questionnaires of a list of common symptoms and was partitioned randomly into training and testing sets and the model prediction performance was evaluated. 6 supervised classification methods which were discriminant analysis, logistic regression, classification trees, random forest, support vector machine, and neural networks were used in the analysis. Random forest yielded the best final model since it resulted in the smallest test set misclassification error rate with the largest AUC of the ROC curve. Based on our findings, most of the models suggested that Polyuria (excessive urination) and Polydipsia (excessive thirst) are the two key early symptoms in predicting patients' early stage of diabetes risk.

1.0 - INTRODUCTION

Diabetes Mellitus or commonly known as diabetes is a chronic health condition in which it affects how our body uses sugar or glucose. According to the Centers for Disease Control and Prevention (CDC), diabetes is the seventh leading cause of death among men and women in the United States. Based on a National Diabetes Statistics Report released by the CDC in 2018, there are approximately 34.2 million people (about 10.5% of the US population) of all ages have diabetes. Furthermore, 7.3 million adults aged 18 years old or older who met laboratory criteria for diabetes are not aware of or did not report having diabetes. In general, there are three types of diabetes: Type I, Type II, and Gestational diabetes, of which about 90-95% of the diabetic patients are Type II. Unlike Type I and Gestational diabetes, Type II diabetes is largely due to the result of excess body weight (obesity) and physical inactivity. Given the fact that diabetes is one of the leading causes of death among men and women in the United States and most of the people are not aware that they have diabetes until they have suffered from the long-term complications caused by the disease, some questions about the early signs or symptoms of diabetes have naturally risen. Our group was motivated to predict the early stage diabetes risk based on the signs and symptoms shown by the patients. The data set was found online from the Kaggle website: https://www.kaggle.com/ishandutta/early-stage-diabetes-risk-prediction-dataset
The data is collected from direct questionnaires from the patients of Sylhet Diabetes hospital in Sylhet, Bangladesh. The questionnaires are composed of a list of common diabetes symptoms and demographic questions and patients are being told to answer the questions honestly. The data set has a sample size of 520 and consists of a total number of 16 predictors that are potentially contributing to whether a given patient will be at positive risk for diabetes or not.

1.1 - METHODS

Upon fitting the models, we first did some data modification. We divided the full data set of 520 observations into two subsets where the first subset called the training set was used to fit the model while the second data set called the test set was used to validate or test the model built or in other words, to evaluate the model performance. In this analysis, we allocated 70% of our data into the training data set (362 observations) randomly with seed number 202112 while the other 30% was allocated into the test set (158 observations). In our analysis, our group chose to apply several classification methods such as discriminant analysis, logistic regression, classification tree, random forests, support vector machine (SVM), and neural networks to our training set to predict the likelihood that the input data will fall into one of the classes, in this case, the two classes would be either patient will be at positive risk for diabetes or negative risk for diabetes.

All of the classification methods listed above belong to the supervised learning methods in which the algorithm learns from the training set by iteratively making predictions on the data using labeled groups and features. In other words, supervised learning methods assume a given structure within the data.

For discriminant analysis method, linear discriminant analysis (LDA) is used when a linear boundary is required between classifiers and two classes are maximally separated with the supervision of y. The assumption required for LDA is the common covariance matrices across all of the response classes. As for the quadratic discriminant analysis (QDA), it is used when a non-linear boundary is needed between classifiers under the normality assumption that the distribution of observation in each of the response classes is normal. Besides that, logistic regression is also one of the supervised learning methods that predicts a binary outcome and the predictors are analyzed to determine the binary outcome with the results falling into one of two categories. Model selection process particularly stepwise selection procedure is applied to select a reduced number of significant predictors for selecting the best fit for the data.

Classification trees or decision trees are also one of the popular classification algorithms. It is non-parametric (distribution free) and it separates the data points by identifying the categories that best differentiate the data points based on the categorical outcome variable. Classification tree generates a multi-level tree diagram and uses the recursive binary splitting that is based on the greedy algorithm in the splitting process. Apart from that, instead of growing one single large tree that would result in overfitting problem, tree pruning would help to remove a subtree that is redundant or not useful in the splitting process. Unlike growing a single classification tree, in a random forest, many classification trees are grown. Since each tree in the forest uses a random subset of predictors to grow, thus it will induce randomness in the tree growing process as well as decorrelate the trees in the forest. As for the support vector machine (SVM) method, a hyperplane that distinctly classifies the classes in the feature space is introduced. Through the use of kernels, nonlinearities in support-vector classifiers can be well controlled. In neural networks, it is a subset of a machine learning method that comprises a node layer, containing an input layer, one or more hidden layers, and an output layer. Neural networks are modeled loosely after the human brain and help in extracting the features and allowing us to deal with classification problems.

Finally, our group is inspired to evaluate and compare the model performances among the six methods: Discriminant analysis, Logistic regression, Classification tree, Random forests, Support vector machine (SVM), and Neural networks by choosing the best final model that produces the smallest test set misclassification error rate and the highest area under the curve (AUC) of the Receiver Operating Characteristic (ROC) curve.

1.2 - EXPLORATORY DATA ANALYSIS

The diabetes data set had a sample size of 520 and no missing values were found. The response variable, class, consisted of a binary outcome where “Negative” represents patients will be at negative risk for diabetes, while “Positive” represents patients will be at positive risk for diabetes. Furthermore, the data set used was a balanced data since the response variable for the event of interest (patients will be at positive risk for diabetes) constituted about 62% of the total sample size (320 observations) while the remaining 38% was made up of the patients that will be at negative risk for diabetes. A

total of 16 predictors were used in the analysis and the detailed information of the respective variables was shown in Table 1.

Variable	Type	Levels	Description
Age	Numerical	Min: 16; Max: 90	Age of the patients
Gender	Categorical	Female or Male	Gender of the subjects
Polyuria	Categorical	Yes or No	Is the patient having excessive/frequent urination?
Polydipsia	Categorical	Yes or No	Is the patient experiencing excessive/increased thirst?
Sudden weight loss	Categorical	Yes or No	Is the patient having unexplained sudden weight loss or unplanned weight loss?
Weakness	Categorical	Yes or No	Is the patient experiencing fatigue, weak, tired feeling?
Polyphagia	Categorical	Yes or No	Is the patient experiencing extreme hunger?
Genital thrush	Categorical	Yes or No	Is the patient having genital thrush?
Visual blurring	Categorical	Yes or No	Is the patient having blurred vision ?
Itching	Categorical	Yes or No	Is the patient having itchy skin?
Irritability	Categorical	Yes or No	Is the patient feeling irritable and having mood swings?
Delayed healing	Categorical	Yes or No	Is the patient experiencing delayed wound healing or slow-healing sores?
Partial paresis	Categorical	Yes or No	Is the patient having weakening muscles?
Muscle stiffness	Categorical	Yes or No	Is the patient experiencing muscle stiffness?
Alopecia	Categorical	Yes or No	Is the patient experiencing loss of hair?
Obesity	Categorical	Yes or No	Is the patient obese or overweight?

Table 1: *This table includes all the variables used in this study. It includes the variable name, variable type (numerical or categorical), variable levels, and variable description.*

Out of the 16 predictors, 15 predictors were categorical while only predictor Age was continuous. The minimum age of patients was 16 years old while the maximum age was 90 years old. On average, the age of the patients that participated was around 48.03 years old.

To get an overview of the 15 categorical predictors: Sex, Polyuria, Polydipsia, Sudden weight loss, Weakness, Polyphagia, Genital thrush, Visual blurring, Itching, Irritability, Delayed healing, Parital paresis, Muscle stiffness, Alopecia, and Obesity, we plotted the bar plots of each of the categorical predictors by the diabetes risk class.

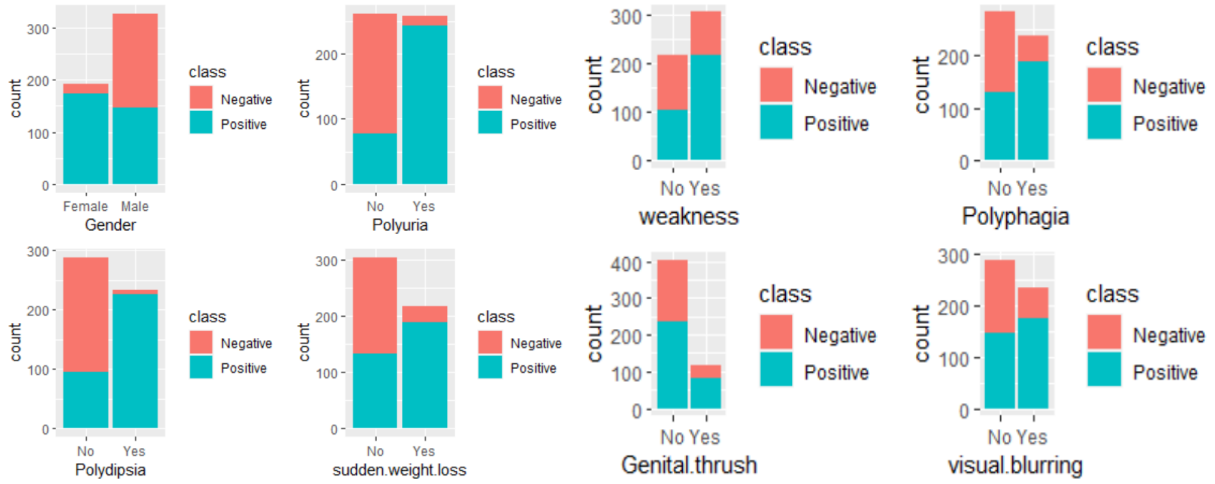

Figure 1: *Bar plots for the categorical predictors, Gender, Polyuria, Weakness, Polyphagia, Polydipsia, Sudden weight loss, Genital thrush, and Visual blurring by the diabetes risk class.*

From the bar plots in figure 1, we were able to observe that female patients possessed a significantly higher ratio of positive risk to negative risk for diabetes as compared to the male patients. Furthemore, patients that experienced polyuria, were more likely to be at positive risk for diabetes, as compared to those that did not experience polyuria. Patients that experienced polydipsia and sudden weight loss had a significantly higher ratio of positive risk to negative risk for diabetes, compared to those patients that did not experience Polydipsia and sudden weight loss. As for the four bar plots at the right, we were able to observe that patients with symptoms such as weakness, visual blurring had a relatively high ratio of positive risk to negative risk for diabetes while patients with polyphagia and genital thrush had a significantly higher ratio of positive risk to negative risk for diabetes as compared to those that did not experience the symptoms.

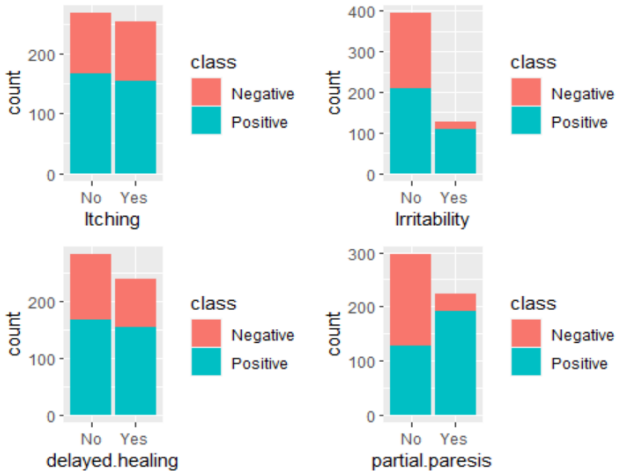

Figure 2: *Bar plots for the categorical predictors, Itching, Irritability, Delayed healing, Partial Paresis by the diabetes risk class.*

Symptom itching had about the same ratio of positive risk to negative risk for diabetes across patients that had itching and had no itching while patients with irritability and

partial paresis had a significantly higher ratio of positive risk to negative risk for diabetes. Besides that, patients with delayed healing also had a relatively high ratio of positive risk to negative risk for diabetes as well.

Figure 3: *Three bar plots for the categorical predictors, Muscle Stiffness, Alopecia, and Obesity by the diabetes risk class and a density plot of the diabetes risk class versus the Age.*

Referring to figure 3 above, we were able to observe that patients that experienced muscle stiffness had a relatively high ratio of positive risk to negative risk for diabetes while overweight patients also had a significantly higher ratio of positive risk to negative risk for diabetes. Patients that experienced no alopecia had a relatively high ratio of positive risk to negative risk for diabetes as compared to the patients that experienced alopecia. 50 years old patients were the majority age in the positive risk class of diabetes.

In addition to that, since most of the predictors were categorical, thus, our group was not able to examine the correlation matrix of the variables. Furthermore, our group also did some data manipulation by applying the one hot encoding to create a new binary feature for each possible category of the categorical predictors. This process was extremely crucial since some of the classification methods such as logistic regression, support vector machine, and neural networks can only be applied to continuous numerical data or in other words, the input predictors must be in numeric value or in the form of dummy variables.

Data standardization was also applied on the predictor Age after partitioning the original data into training set and testing set since Age was the only continuous predictor among the 16 predictors. Normalizing the data or feature scaling was important especially in neural networks since neural networks were particulary sensitive to scaling and normalizing the data through rescaling the data from the original range to mean 0 and standard deviation of 1 would ensure all of the predictors on the same scale.

Additional plots and discoveries were included in Appendix A.

1.3 - STATISTICAL ANALYSIS / MODELING

In order to use the discriminant analysis method, we started off by using the training set to fit into the Box's M-test by testing the homogeneity or equal covariance matrices assumption. Since the p-value of the Box's M- test was 2.2e-16 which was smaller than alpha 0.05, hence we would reject the null hypothesis and conclude that the homogeneity of covariance matrices assumption did not hold. Thus, Linear Discriminant Analysis (LDA) was not able to apply into our data set because our data did not satisfy the equal covariance matrices assumption. Our group also further evaluated the Quadratic Discriminant Analysis (QDA) method. We came to the conclusion that QDA also failed to apply in our case because in our data, 15 out of the 16 predictors were categorical and the normality assumption for QDA was not satisfied.

```
        Box's M-test for Homogeneity of Covariance Matrices

data:  data.train[, -1]
Chi-Sq (approx.) = 949.96, df = 136, p-value < 2.2e-16
```

Figure 4: *This figure shows the result of Box's M-test for testing the homogeneity of covariance matrices.*

Apart from that, we fitted the training set with the Conventional Logistic Regression by including all of the 16 predictors and we then ran the model selection process which in this case, we chose stepwise selection since it performed both backward elimination and forward selection. In the end, as shown in figure 5, 12 predictors were kept in the model with an Akaike Information Criterion (AIC) value of 144.8.

(Intercept)	Age	GenderMale	PolyuriaYes	PolydipsiaYes	weaknessYes	PolyphagiaYes
2.0492322	-0.0462558	-4.4367281	4.6601708	4.9622869	0.8704835	0.9994824
Genital.thrushYes	visual.blurringYes	ItchingYes	IrritabilityYes	delayed.healingYes	partial.paresisYes	
2.2880439	1.6780496	-2.8125865	2.7268178	-0.9383595	1.1596561	

Figure 5: *This figure shows the coefficients of the 12 covariates after fitted using the conventional logistic regression with stepwise selection.*

We also further assessed the Hosmer-Lemeshow goodness of fit of the model. Since the p-value was 0.9039, thus, it was not significant at alpha level 0.05 and we were able to conclude that there was no lack of fit issue in the model. Furthermore, the model prediction accuracy using the test set resulted in a misclassification error rate of 6.962025% which 0.5 was used as the classification threshold and the area under the curve (AUC) of the ROC curve was relatively high, approximately 0.9718776.

Beyond that, we also fitted the training set using the classification tree method. As shown in figure 6, at the top of the tree, variable Polydipsia (excessive/increased thirst) was first selected in the splitting process since Polydipsia was the best predictor that gave the maximum reduction of the classification error rate. Polyuria (excessive urination) was the next best predictor selected since it gave the next maximum reduction of classification error rate in the tree building process. The best split was made based on the recursive binary splitting with the greedy algorithm in which the best split was always made at the particular or current step. The full grown large tree consisted of 15 terminal nodes and variables used in the tree construction were

Polydipsia, Polyuria, Gender, Alopecia, Visual blurring, Age, Irritability, Partial Paresis, Itching, and Obesity. The test set misclassification error rate using the fitted model with classification tree was 7.594937% and the AUC of the ROC curve was roughly 0.9643507.

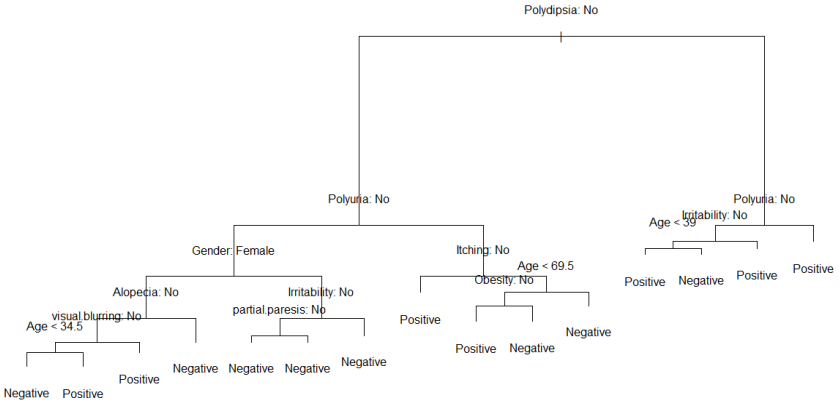

Figure 6: *This figure shows the full tree with 15 terminal nodes.*

Aside from fitting the training set using the classification tree, we also pruned the tree in order to avoid any overfitting issue. By using the 10-fold cross validation, the best subtree size that associated with the smallest cross validation deviance was 8. In the pruned tree, there were 8 terminal nodes and variables such as Polydipsia, Polyuria, Gender, Alopecia, Itching, Age, and Obesity were used in the tree building process. The fitted model with pruned tree resulted in a test set misclassification error rate of 6.962025% which was the same error rate as the conventional logistic regression with stepwise selection. The AUC of the ROC curve in this case was 0.9379653.

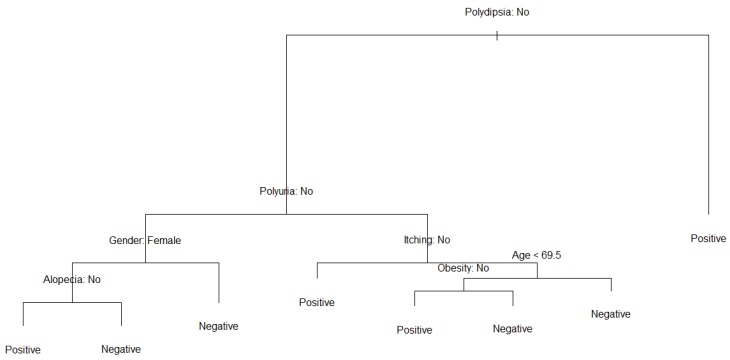

Figure 7: *This figure shows the pruned tree with 8 terminal nodes.*

Due to the fact that the performance of a single tree could be unstable, our group went ahead to fit the training set by growing many classification trees in a random forest. In random forests, since there were a total of 16 input predictors, we set the number of random variables used in each tree to be sqrt(16)=4 and number of trees used in the forest to be 500. In other words, each tree in the forest would use a random subset of 4 predictors during the splitting process to ensure less correlation among the 500 trees in the random forest. As shown in figure 8, Polydipsia and Polyuria were the two most important variables because removing these variables would reduce the model

accuracy by 47.93700% and 46.02677% respectively. Again, Polydipsia and Polyuria were the two best predictors in order to split the data since these two variables would cause a maximum reduction of Gini Index by 33.610388% and 34.998751% respectively. The resulting test set misclassification error rate was 1.898734% and the AUC of the ROC curve was significantly high, 0.9990074.

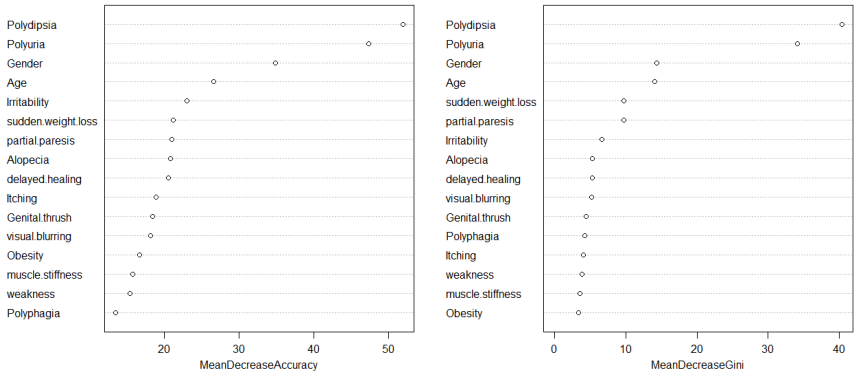

Figure 8: *This figure shows the important scores of the 16 predictors in which Polydipsia and Polyuria are ranked the two most important variables among all.*

Apart from that, our group also fitted the training set using the support vector machine (SVM). Different types of kernel functions such as linear, radial, and polynomial were tested and 10-fold cross validation was used to choose the best parameters for each of the kernels. For the linear kernel, the best cost parameter chosen from a sequence of alphas was 100 since it gave the smallest performance rate and the corresponding test set misclassification error rate using the fitted model was 6.962%. The test set prediction accuracy using the fitted model with radial kernel that consisted of gamma=0.5 and cost=100 resulted in a misclassification error rate of 3.797%. Furthermore, polynomial kernel with gamma=2, cost=0.001, and degree=3 was chosen to be the best among the three since it produced the smallest test set misclassification error rate of 1.8987% while the test set performance based of the AUC of the ROC curve was about 0.983871.

In neural networks, the standardized version of the training set was trained using a neural net model and 2 hidden layers were chosen because our data sample size was not relatively that huge. Different combinations of hidden nodes or hidden units were tested (3,2), (2,1), (5,3), (2,2) and (3,2) was selected among all because the test set yielded the smallest misclassification error rate of 1.89% with an AUC of 0.9899917.

1.4 DISCUSSION

In conclusion, for early stage diabetes risk prediction dataset, although the test set misclassification error rate of 1.898734% was the same among the model fitted with random forest, support vector machine using polynomial kernel, and neural networks with 2 hidden layers, the AUC for the test set using the model fitted with random forest was the highest in which the AUC was almost close to 1. Thus, our best final model selected was the model fitted with random forest since it had the best test set prediction

accuracy and outperformed the other classification methods in doing the job of classifying and distinguishing the diabetes class risk.

Methods	Test set misclassification rate	AUC for Test set
Discriminant Analysis	Not available	Not available
Logistic Regression with stepwise	6.962025%	0.9718776
Classification tree	7.594937%	0.9643507
Classification tree (Pruned)	6.962025%	0.9379653
Random Forest (500 trees and 4 random subset of predictors)	***1.898734%***	***0.9990074***
Support Vector Machine (Polynomial Kernel)	1.898734%	0.983871
Neural Networks (2 hidden layers)	1.898734%	0.9899917

Table 2: *Comparison of the test set misclassification rate and area under the curve for the test set among the 6 classification methods.*

In short, Polyuria (excessive urination) and Polydipsia (excessive thirst) were the two most important variables since the models fitted with logistic regression, classification tree as well as random forests suggested that Polyuria and Polydipsia were the two key early signs or symptoms in predicting patients' early stage of diabetes risk. Furthermore, in the logistic regression with stepwise model, the odds of patients having a positive risk of diabetes increased by a factor of 105.65412203 for patients that had polyuria symptoms as compared to the odds for patients that have no polyuria symptoms. The odds of having a positive risk of diabetes increased by a factor of 142.92026155 for patients that had polydipsia symptoms as compared to the odds for patients that have no sign of polydipsia. In the pruned classification tree, we were also able to observe that patients that had no signs of polydipsia and itching but with signs of polyuria were predicted as positive risk for early stage diabetes in the third split of the tree while patients that experienced signs of polydipsia were classified as positive risk for early stage diabetes in the first split.

Although random forest is a robust model that assumes no assumption, no feature scaling required, handles both categorical and continuous variables and overfitting issues, training a large number of trees may be complex and would require more computational power and resources. Therefore, there is no such existence of the best classification method that fits for all since some classification methods could perform well in some circumstances while not in others and each method has their respective advantages and disadvantages. Our group suggests exploring other classification methods that could possibly yield a better fit model such as K-Nearest Neighbors classification method, a type of lazy learning method that simply stores instances of the training data and the classification is computed from a simple majority vote of the k nearest neighbors of each point or Naive Bayes method which based on the Bayes's theorem with the assumption of independence between every pair of features.