Predictive Modeling of Contraceptive Use Among Indonesian Women in 1987

Authors:
Nidhi Chalgeri, Tyler Nunez, Louis Spencer

Table of Contents

Abstract

Using data from the 1987 National Indonesia Contraceptive Prevalence Survey, we aimed to build a predictive model that is able to classify contraceptive types based on features given in the dataset. However, the features provided in the data did not produce an accurate model for this classification problem. Instead, in order to simplify the problem, we tried to predict instead whether or not an Indonesian woman uses contraceptives, reducing the problem from a multi-class classification problem to a binary classification problem.

Introduction

The 1987 Indonesian survey (hereby referred to as NICPS) was the first of its kind to be conducted in Indonesia and was a part of a larger project intended to gather information on family planning and fertility behaviors of people in a variety of countries for policy makers and organizations to enhance and provide family planning programs. Using this data beyond its intended purpose poses ethical dilemmas considering that our data provides very personal information of individuals - however, individuals' information is protected because the data is anonymized. Obtaining information regarding family planning and contraceptive methods will always pose ethical issues because this information is private and individuals may not want to divulge such information - anonymity of subjects is effective for getting around this. Despite these issues, moving forward with our analysis we will assume that this data accurately represents fertility behaviors of individuals in Indonesia in 1984

We must also understand that there are serious limitations to the scope of our findings. Because this is the first year of such a survey, it is quite likely that a subset of Indonesian people were unable to be a participant in the study - thus we must call into question the population that this data represents. Furthermore, we cannot apply our findings to today's Indonesian population because availability to all contraceptives has increased in the past 40 years.

Description of Methods

Our goal is to answer the question of whether it is possible to predict if an Indonesian woman uses contraceptives based on the features in the dataset. To do this, we trained and validated a logistic model, a random forest model, a K-nearest neighbors model, and a neural network to find the best model for the classification task.

Various transformations were applied to the raw data to make it ready for classification. To start, the response labels were transformed from multi-class labels into binary labels, with an Indonesian woman being in class 1 if her original response label was ‘2’ or ‘3’, indicating that she is taking some kind of contraceptive, and class 0 if her original response label was ‘1’, indicating no use of contraceptives. In addition, since there was only 2 numeric features in the entire dataset, in an effort to try to find useful features for prediction, we added various transformations to the original 2 features and produced new ones, such as **e**^(num_child), **log**(wife_age), and **num_child**^4. To finish our preprocessing, we one-hot encoded all of the categorical variables in the dataset to be used in prediction.

To begin, we created a Logistic Regression model. We used cross-validation to find the most optimal regularization parameter out of a list of several possible parameter values. We also trained a Random Forest model in which n_estimators were set to 1500 to avoid overfitting. We then created a helper function that returned the training and validation accuracy of the given model, using the model and training data as inputs.

Next, we generated our K-nearest neighbors model by using the KNeighborsClassifier. We built a helper function that took in the number of neighbors as the input and returned the training and validation accuracy of a K-nearest neighbors model whose hyper parameter n_neighbors is equal to the input. We proceeded to loop over different values of n_neighbors and found the value of n_neighbors that maximized the validation accuracy for that particular model.

The Neural Network model was generated using a KerasClassifier. The neural network has two hidden layers consisting of 20 neurons each, with a dropout(0.5) layer added after the last hidden layer. The ReLU activation function was used in the hidden layers, while in the output layer, the logistic activation function was used.

Summary of Results

The logistic model returned a training accuracy of around 73.55% and a validation accuracy of around 71.2% The random forest model returned a training accuracy of around 98.6% and a validation accuracy of around 70.3%. The K-nearest neighbors model returned a training accuracy of around 72.2% and a validation accuracy of around 70.2%. The neural network model returned a training accuracy of 75% and a validation accuracy of around 70.7%

The logistic model had the best validation accuracy out of all the models (Figure 6). Using the test data, the logistic model at a final test accuracy of around 72.89%. The worst model was apparently the random forest model with a validation accuracy of around 69.74%.

We ran several visualizations of the data and created a heatmap, a confusion matrix and a principal component analysis. The heatmap (Figure 1) showed the correlation between each of the features in the dataset, allowing us to choose which features to select when implementing our model. In our confusion matrix (Figure 2), we found that we had a high number for True Positives and True Negative, 559 and 253 respectively, but saw that the number of our False Positives was quite high as well, 212. This PCA (Figure 3) revealed that should we have used the features that we selected, we would not have had an incredibly accurate model for predicting the correct contraceptive. Should our features accurately predict contraceptive type, we would see distinct blue, red, and green clusters in this principal component analysis. However, it is possible that we could see these distinct clusters in a three dimensional space. After generating it in 3 dimensional space (Figure 4), our principal component analysis did not display distinct clusters for the different contraceptive types. The visualizations supported our findings from modeling that our model is not able to accurately predict different contraceptive methods from these features.

Analysis

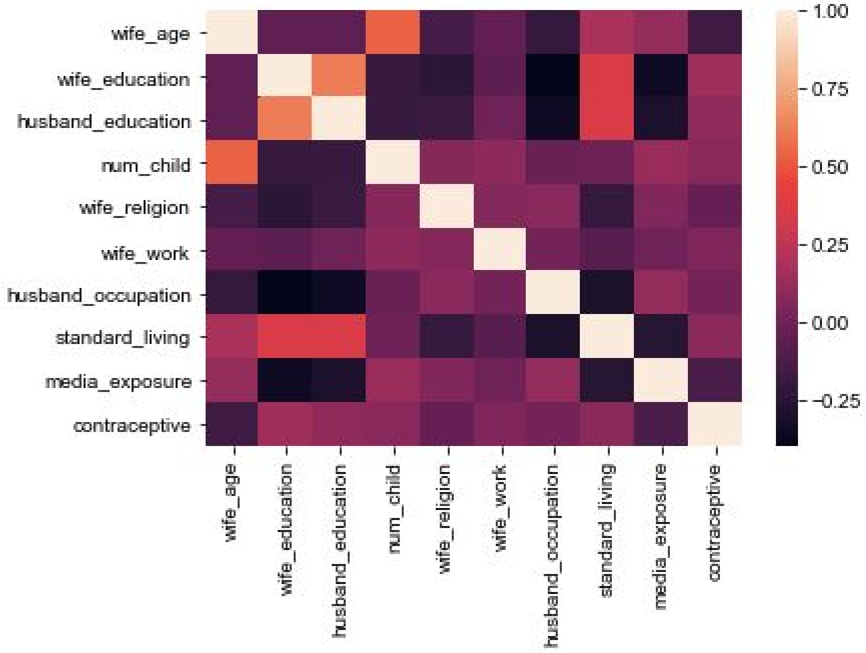
Given our results, we could see that the age of the woman and the number of children the woman had were the most important features in addressing our question. These two features had a fairly different distribution between non-contraceptive users and contraceptive users, prompting us to explore whether these features had a significant effect in determining if women did use contraceptives.

We found that, given our best test accuracy was around 73%, our model could fairly predict the use of contraceptives among Indonesian women. There were some features, such as the transformation of existing features, that we thought would be more successful but turned out to be ineffective. We were challenged when trying to prove the test accuracy, but were able to find it once we made the classes binary.

Some limitations of our analysis was that much our data was categorical. It would have most likely improved our model if we included numerical data, as well. Another limitation was the very specific demographic we were given when trying to determine the answer to our project question. It would have been to explore different types of women.

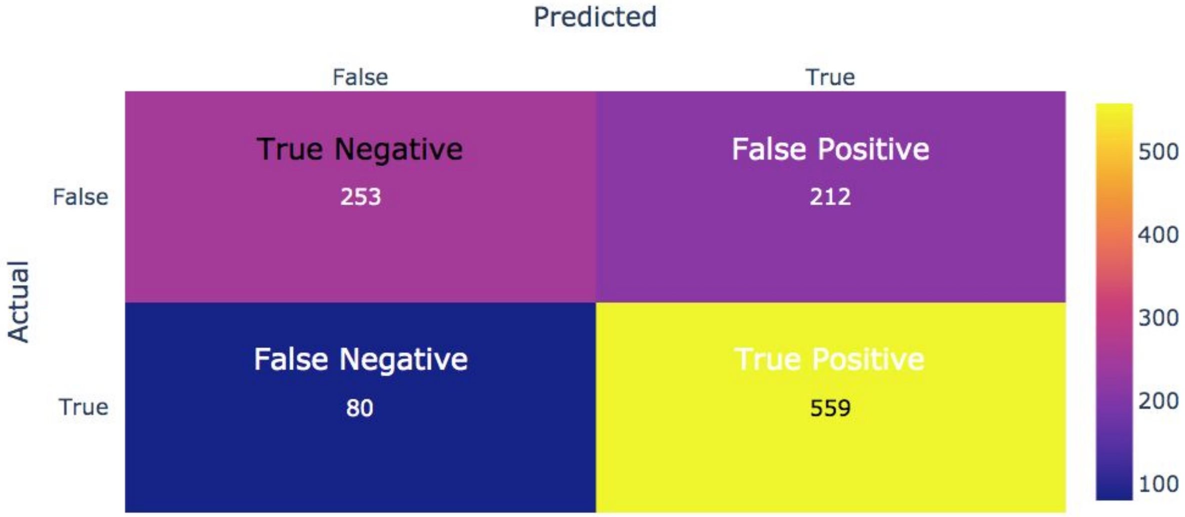
Figures & Graphs

Figure 1: HeatMap



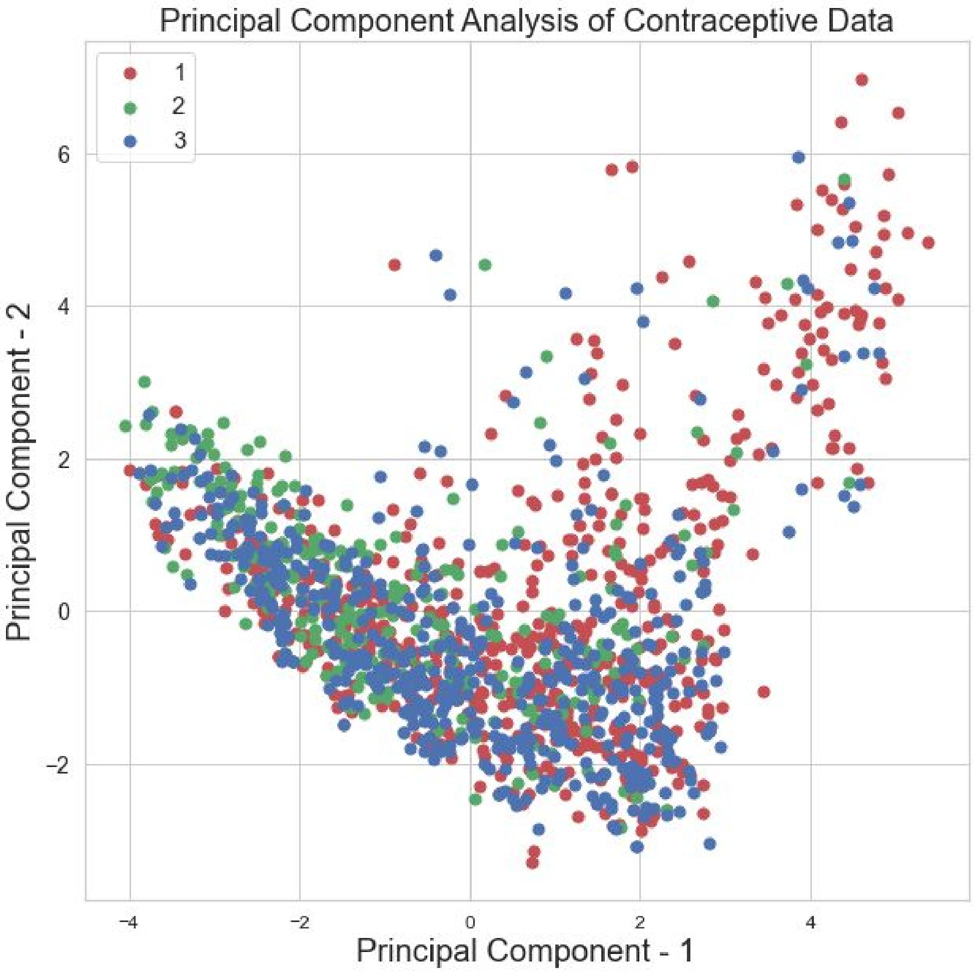
In the figure above, we plot the features against each other to determine the correlation between all of them.

Figure 2: Confusion Matrix for Logistic Model



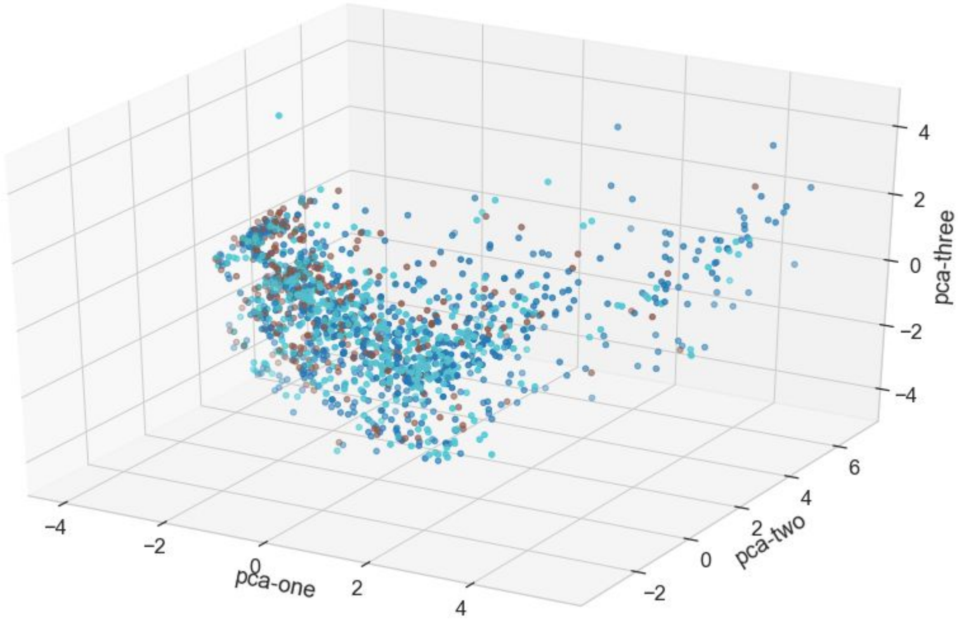
In the figure above, we used the training and test set data to determine the number of false positives, false negatives, true negatives, and true positives there are. This will better help to see if our model worked in accurately creating predictions.

Figure 3: Principal Component Analysis



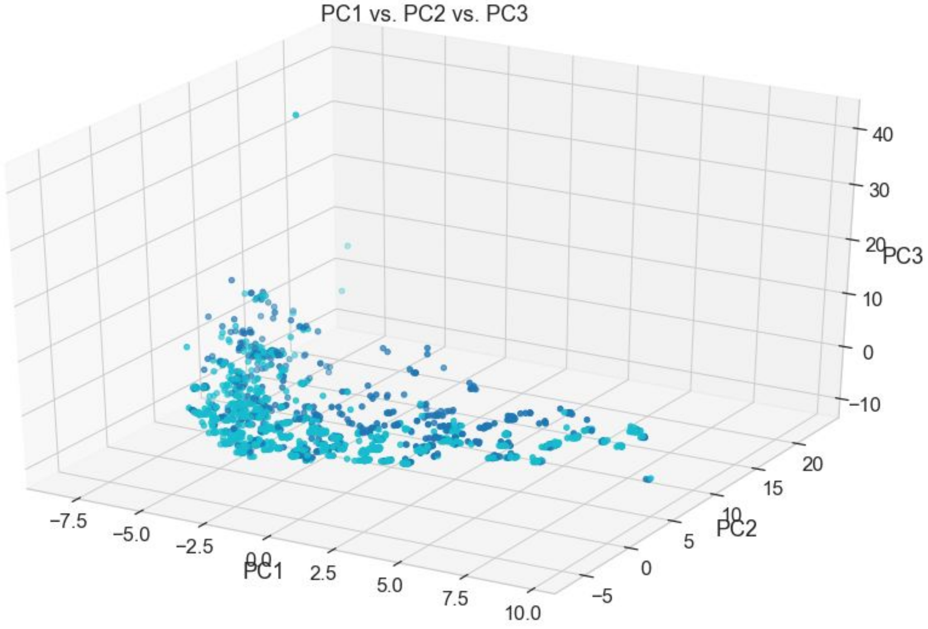
In a 2-Dimensional Principal Component Analysis to test how well our features worked for predicting long term, short term, and no contraceptive use

Figure 4: Principal Component Analysis (3D)



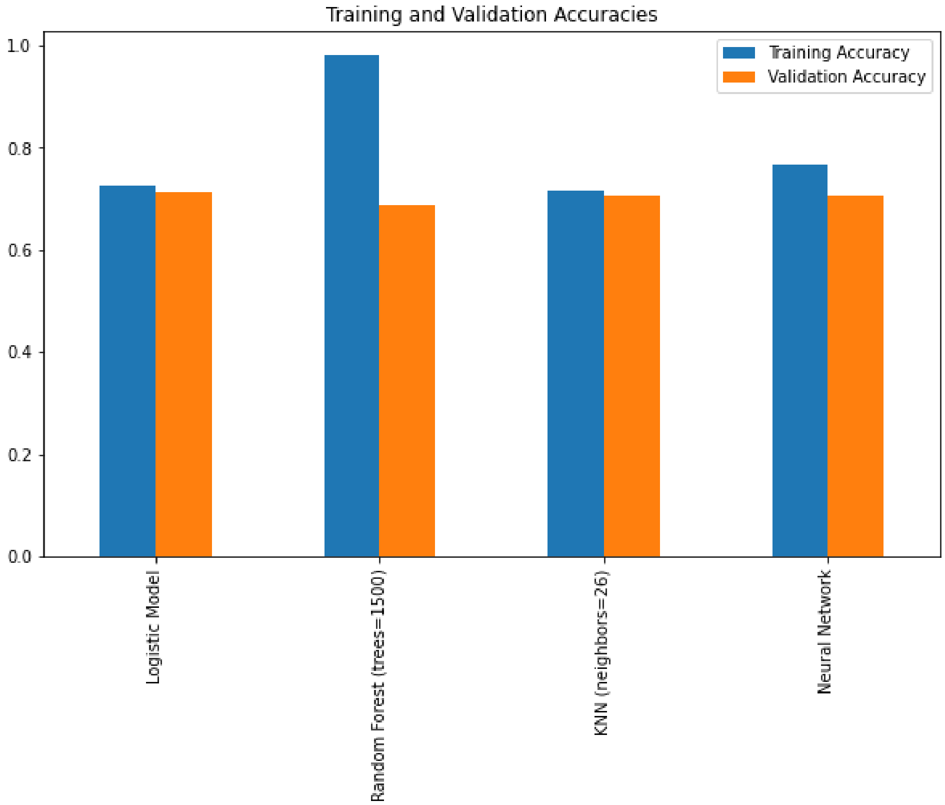
3-Dimensional Principal Component Analysis to test how well our features worked for predicting long term, short term, and no contraceptive use

Figure 5: Principal Component Analysis (3D)



3-Dimensional Principal Component Analysis for predicting binary contraceptive use.

Figure 6: Bar Plot Comparing the Training and Validation Accuracies of all models



Plotted above, the Logistic Model has the best validation accuracy.