Study of contraceptive method prediction model in Indonesia

Harry Chan

updated: 2022-01-22

Contents

0.1 Exercise 1: Audience Personas

Crystal is an undergraduate computer science student interested in machine learning. She has taken courses in machine learning. She is curious about any kinds of problem machine learning could be a viable solution to and how technology can benefit the general public.

0.2 Exercise 2: Write your report

1 Abstract

We built a classification model using the SVC classifier algorithm which can help predict the use/or no use contraceptives by women based on their demographic and socio-economic characteristics. The data used in our project was a subset of 1987 National Indonesia Contraceptive Prevalence Survey. The samples are married women who were either not pregnant or do not know if they were at the time of interview.

The overall accuracy of the model was 74% and the area under the curve (AUC) was of 78%. Given that the data-set was limited this seems to be a decent score. However the model still had a few false predictions for the non usage of contraceptive. These cases where false positives, that was predicting the usage of

contraceptive when in fact the person did not use contraceptives. These kind of predictions gave wrong insights of contraceptive usage, thus further work to improve model prediction was needed before we could put this model in the real world.

When evaluating the training set, the DT model with 6 of the best features selected by Random Forest Importance (RFI) produces the highest cross-validated accuracy score. Similarly, when evaluating the test set, the DT model performed the best on accuracy score of approximately 55%. Therefore, it can be concluded that the DT method was the most suitable model for this classification problem based on the given data-set.

2 Introduction

Family planning offers immediate health benefits, potential non-health benefits that encompass expanded education opportunities and empowerment for women as well as economic advancement. Consequently, the proportion of women of reproductive age who are in need for modern contraceptive methods has increased gradually from 73.6 percent in 2000 to 76.8 percent in 2020 (Economic and Affairs 2020). Reasons for the slow growth may arise from lack of access to services, poor quality of available services, gender barriers and personal bias against some methods. These limitations are not addressed in some regions due to their cultural/traditional beliefs, financial background, religious influence, etc.

In last two decades, decision tree-based and multivariate regression-based prediction models have received some attention in the study of family planning, but it generally exhibit modest predictive performance, especially for certain large and complicated datasets (T. Lim, Loh, and Cohen). Machine learning (ML) such as SVC can improve the performance by exploiting large data repositories to identify novel predictors and non-linear interactions between them.

In this study, we conducted a prediction model research based on a specific background in Indonesia. The objective is to predict whether the contraceptive method is used by Indonesian women based on their demographic and socio-economic factors. In 1991, a dataset were taken from the 1987 National Indonesia Contraceptive Prevalence Survey and a cohort of 1473 samples were selected.

3 Methods

3.1 Data Collection

This data-set is a subset of the 1987 National Indonesia Contraceptive Prevalence Survey (T.-S. Lim, Loh, and Shih 2000). It was sourced from the UCI Machine Learning Repository (Dua and Graff 2017). The samples are from single or married women who were either not pregnant or do not know if they were at the time of interview. It is made up of 9 attributes (2 numerical attributes and 7 categorical attributes), 1473 observations and three classes.

The Indonesian Central Bureau of Statistics conducted the survey from September to December 1987 to collect information on sociodemographic status (ethnicity, education level, financial status, occupation, marital status, religious belief and so on). Interviews were conducted with 11,884 ever-married women age 15-49, from a sample of households representing 93 percent of the population of Indonesia.

3.2 Ethics Approval

The study entailed secondary analyis of data containing no personal identifying information. The participants were informed about the objectives and methods of the study. They were informed that their participation was totally voluntary and that they could withdraw from the study at any time without citing any reason. Written and signed or thumb printed informed consent was obtained from those who agreed to participate, or from their guardians.

3.3 Data Pre-Processing

3.3.1 Summary of the dataset

Exploratory data analysis was performed on the training data-set and was found to have no missing values. Table 1 shows the description of the attributes of this dataset. The attribute `Contraceptive method used` was taken as a target variable, and the ermaining eight attributes were taken as feature variables.

Table 1: Summary of the dataset

Column name	Description	Type	Values
Wife age	Wife's age	Numberical	any positive values
Wife education	Wife's education	Categorical	1=low, 2, 3, 4=high
Husband education	Husband's education	Categorical	1=low, 2, 3, 4=high
Number of children even born	Number of children evern born	Numberical	any positive values
Wife religion	Wife's religion	Binary	0=Non-Islam, 1=Islam
Wife now working	Is wife working or not	Binary	0=Yes, 1=No
Husband occupation	Husband's occpation	Categorical	1,2 ,3 ,4
Standard-of-living index	Standard-of-living Index	Categorical	1=low, 2, 3, 4=high
Media Exposure	Media exposure	Binary	0=Good, 1=Not good
Contraceptive method used	Contraceptive method used	Categorical	1=No-use, 2=Long-term, 3=Short-term

3.3.2 Distribution of the target class

There are three classes in this dataset - with 1 ("No-use"), 2 ("Long-term use"), followed by 3 ("Short-term use"). In this study, we aimed to examine whether the women would use any contraceptive methods. Therefore, we have combined 2 and 3 as "use" case and have left 1 as it is ("no-use" case). After the transformation, the distribution of the classes: `0 = No-use`: 636 observations, `1 = use`: 837 observations.

3.3.3 Data Transformation

Based on the EDA (Exploratory Data Analysis) performed earlier and variable descriptions, there were no missing values in the dataset. However, the variables were of different data types. In order to perform operations on data, the consistency of data types should be guaranteed. Feature scaling (Standardization) was performed on the numeric data while integer-encoding were done for the ordinal categorical features. The following table shows different variables in the data-set and the respective transformation performed on each of them.

Table 2: Transformation of the variable

Data Type	Variable	Transformation	Technique used
Numerical	Wife's age, Number of children evern born	Scaling	Standarization
Ordinal	Wife's education, Husband's education	Encoding	Integer Encoding
Ordinal	Husband's occpation, Standard-of-living Index	Encoding	Integer Encoding
Binary	Wife's religion, Is wife working or not, Media exposure	None	Pass through

3.4 Data Analysis

3.4.1 Model Selection

We have explored the following four Machine Learning classifier algorithms to predict the target feature:

1. Decision Tree
2. kNN
3. Logistic Regression
4. RBF SVC

3.4.2 Train-Test Split

Data sampling was not required as the original data-set was not a significantly larger one. As shown below, the model has been trained and tuned on 1031 rows of training data and tested on 442 rows of test data. This was constituted from 70:30 ratio of training vs test observations in the dataset.

3.4.3 Model Evaluation

To measure the performance of the model we used the area under the receiver operating characteristics (ROC) curve (AUC or AUROC). 10-fold cross validation has been used as the model evaluation strategy. Accuracy results are presented as mean SD calculated over the tenfold validation sets.

3.4.4 Cross Validation Results

From the table below, it can be clearly inferred that the RBF SVC algorithm gave the best score on both training and validation set.

Table 3: Cross Validation Result

X	decision.tree	kNN	Logistic.Regression	RBF.SVM
fit_time	0.01	0.01	0.03	0.04
score_time	0.01	0.01	0.01	0.02
test_score	0.63	0.65	0.66	0.69
train_score	0.99	0.77	0.68	0.75

3.4.5 Hyperparameter Tuning

The randomized search for hyperparameter tuning of each classifiers has been performed via cross-validation approach. Given that the performance of RBF SVC was the best, we applied hyper-parameter tuning and experiment with the pre-defined hyperparmeters in a total of 200 iterations to find the optimal parameters. In each iteration, we compute the validation accuracy. A dictionary for the hyperparameters of the RBF SVC classifier is defined as below:

- value range for "gamma": Between -3 and 4 on a log scale
- value range for "C": Between -2 and 6 on a log scale
- values for `class_weight`: None, Balanced

Among all of the combination of the hyperparameters, we settled for the combination of hyperparameter `gamma = 0.01, C = 10, class_weight = None`. The results of the top 5 models are shown in the table below.

Table 4: Hyperparameter Selection

X	X1	X2	X3	X4	X5
mean_test_score	0.70	0.70	7e-01	0.70	0.69
param_svc__gamma	0.01	0.00	1e-02	0.01	0.10
param_svc__C	10.00	1000.00	1e+03	100.00	10.00
param_svc__class_weight	NA	NA	NA	NA	NA

X	X1	X2	X3	X4	X5
mean_fit_time	0.21	0.15	4e-01	0.17	0.19

4 Results

4.1 Confusion Matrix

The accuracy of the machine learning algorithm can be calculated from the confusion matrix. In the abstract term, the confusion matrix is given below. Here, FP = False Positive, FN = False, Positive, TN = True Negative, and TP = True Positive. Accuracy, Recall, Precision and F-measure were used to calculate the performance measurement of the classification.

	Predicted Not Use (0)	Predicted Use (1)
Actually Not Use (0)	TN	FP
Actually Use (1)	FN	TP

The confusion matrix of RBF SVC classifier for test set is shown in the figure 1.

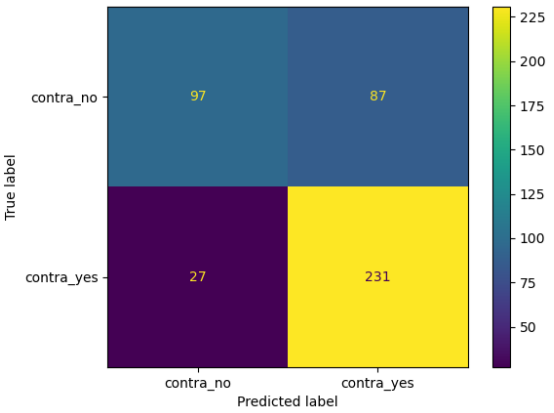

Figure 1: Confusion Matrix (Actual vs Predicted)

We have considered the use of contraceptive method as positive class. The findings of Confusion Matrix suggested that the model preformed well on the total number of `True positives` i.e 231 which were the ones that the model predicted correctly to be using contraceptive method.and `True Negatives` i.e 97 which were predicted correctly for not using contraceptive method.

However, it was found that there were some false positives and false negatives. `False positives` were indicated when the model affirmatively predicted the use of contraceptive method when in fact, the person did not use contraceptives i.e in our matrix 87 .and `False Negatives` indicated when the model incorrectly predicted the person was not using, when they did not use contraceptives.

4.2 Scoring Metric

The recall, precision and the f1-score were calculated while considering each class to be the positive class. Our findings showed that the recall value was approximately **0.90**, indicating a better true positive rate (TPR) for the `1` class and **0.53**, indicating the TPR of the `0` class. The accuracy and weight average accuracy obtained were 74% and 73% respectively.

Table 6: Scoring Metrics

X	precision	recall	f1.score	support
contra_no	0.78	0.53	0.63	184
contra_yes	0.73	0.90	0.80	258
accuracy	NA	NA	0.74	NA
macro avg	0.75	0.71	0.72	442
weighted avg	0.75	0.74	0.73	442

4.3 ROC Curve

As mentioned above, the ROC curves were plotted at a threshold of 0.5. In order to obtain an overall score for our model, the Area under the curve was observed which resulted in a decent score of 78% from the figure 2: AUC ROC Curve.

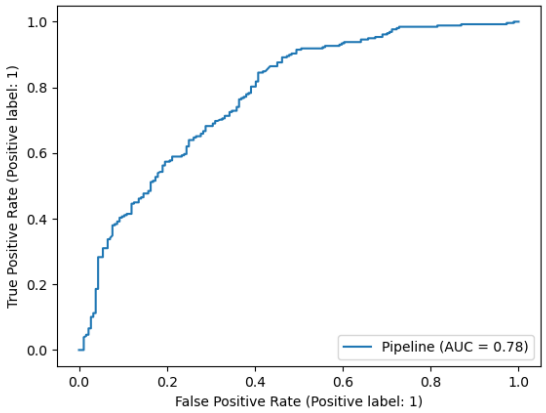

Figure 2: AUC ROC Curve

5 Conclusion

The process intended to predict the use of contraceptives in married women based on their socio-economic and demographic information. In the process, 4 different models were tried. In our study, the best model RMB SVC performed well given the size of the data-set with an accuracy of 74% , `recall` of 90%, `precision` of 73% , `f1_score` of 80% and `AUC` 78%. The precision indicates, that out of all predicted positives, how many are positive, i.e. out of 100 predicted positive samples, 73 were using the contraceptives. High Recall indicated, out of total positives how many were predicted positive A high recall showed that our model could identify most of them correctly. These results were in line with the validation scores outlined previously. The result of high `recall value` of 90% also indicated that `False Negatives` were very low and showed an appreciable f1_score of 0.8.

Since the size of the data set was significantly small (only 1473), the accuracy rate of the models developed could be considered low to represent the entire county (Indonesia). Also, this dataset could be a biased one, considering its small size of it. As a well-known limitation of the supervised machine learning is that it requires a large number of data to achieve a reasonably accurate model. Therefore, if a larger data set is available for this exercise, then a more accurate model could have been developed. In opposed to supervised machine learning, deep learning would have been a better approach for this kind of problem with limited data sets. There is room to improve the model in future by considering more parameters and more ensemble methods during the hyperparameter tuning process.

6 Acknowledgment

This data-set is a subset of the 1987 National Indonesia Contraceptive Prevalence Survey . It was sourced from the UCI Machine Learning Repository.

The Python programming languages (Van Rossum and Drake Jr 1995) and the following Python packages were used to perform the analysis: altair (VanderPlas et al. 2018), docopt (de Jonge 2018), matplotlib (Hunter 2007), numpy (Harris et al. 2020), pandas (McKinney et al. 2010), sckit-learn (Pedregosa et al. 2011). The code used to perform the analysis and create this report can be found here.

References

de Jonge, Edwin. 2018. *Docopt: Command-Line Interface Specification Language*. https://CRAN.R-project.org/package=docopt.

Dua, Dheeru, and Casey Graff. 2017. "UCI Machine Learning Repository." University of California, Irvine, School of Information; Computer Sciences. http://archive.ics.uci.edu/ml.

Economic, United Nations Department of, and Social Affairs. 2020. "World Family Planning 2020 Highlights." 2020. https://www.un.org/development/desa/pd/sites/www.un.org.development.desa.pd/files/files/documents/2020/Sep/unpd_2020_worldfamilyplanning_highlights.pdf.

Harris, Charles R., K. Jarrod Millman, Stéfan J van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, et al. 2020. "Array Programming with NumPy." *Nature* 585: 357–62. https://doi.org/10.1038/s41586-020-2649-2.

Hunter, John D. 2007. "Matplotlib: A 2d Graphics Environment." *Computing in Science & Engineering* 9 (3): 90–95.

Lim, Tjen-sien, Wei-yin Loh, and W. Cohen. "A Comparison of Prediction Accuracy, Complexity, and Training Time of Thirty-Three Old and New Classification Algorithms." In *Machine Learning*, 2000.

Lim, Tjen-Sien, Wei-Yin Loh, and Yu-Shan Shih. 2000. "A Comparison of Prediction Accuracy, Complexity, and Training Time of Thirty-Three Old and New Classification Algorithms." *Machine Learning* 40 (3): 203–28.

McKinney, Wes et al. 2010. "Data Structures for Statistical Computing in Python." In *Proceedings of the 9th Python in Science Conference*, 445:51–56. Austin, TX.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, et al. 2011. "Scikit-Learn: Machine Learning in Python." *Journal of Machine Learning Research* 12 (Oct): 2825–30.

Van Rossum, Guido, and Fred L Drake Jr. 1995. *Python Tutorial*. Centrum voor Wiskunde en Informatica Amsterdam, The Netherlands.

VanderPlas, Jacob, Brian Granger, Jeffrey Heer, Dominik Moritz, Kanit Wongsuphasawat, Arvind Satyanarayan, Eitan Lees, Ilia Timofeev, Ben Welsh, and Scott Sievert. 2018. "Altair: Interactive Statistical Visualizations for Python." *Journal of Open Source Software* 3 (32): 1057.