Family Planning: A Data Based Analysis

Cole Rose
University of California, Berkeley
colerose@berkeley.edu

Rayan Alatab
University of California, Berkeley
ralatab@berkeley.edu

Abstract

This paper aimed to predict contraceptive use in women based on various demographic and socioeconomic characteristics. Additionally, the efficacy of logistic regression, decision trees, and random forest models on predicting contraceptive use was evaluated. The problem was initially framed as a three class problem. The initial three classes of contraceptive use were No Use, Short Term Use, and Long Term Use. The problem was later reduced to a two class case by combining No Use and Short Term Use into a single class. Various techniques such as feature engineering, hyperparameter tuning, oversampling, undersampling, and SMOTE sampling were performed in an attempt to optimize the models. The most successful model was a fine tuned random forest model on the two class case. The model used 110 learners with a max depth of 6. The final model had a resulting training and test accuracy of 77.50 % and 73.76 % respectively.

1. Introduction

This paper presents a data based analysis of the *Contraceptive Method Choice Data Set* from the University of California Irvine's Machine Learning Repository [4]. The data is a subset of the 1987 National Indonesia Contraceptive Prevalence Survey. The data collected contains contraceptive choice, demographic, and socioeconomic data on married women that were either not pregnant or did not know that they were pregnant at the time. The goal of this paper is two fold. The first goal will be to attempt to create a model that can predict the type of contraceptive a family will use based on a variety of factors, such as religion, number of children, and level of education. The three classes of contraceptive that will be predicted are: No Use, Long Term Use, and Short Term Use. The second goal will be to explore the efficacy of different classification models on this data, as well as the ethical and social constraints that are imposed by predicting contraceptive choice. First and foremost, the data was cleaned to allow for an effective analysis. Second, exploratory data analysis **(EDA)** was conducted to determine factors such as class balance, predictive power of features, and covariance relationship between features. Various models such as multi-class logistic regression, decision trees, and random forest trees were trained. After fine tuning the hyperparameters of these models via cross validation with grid search, various feature transformations were applied to the continuous features of the model in an attempt to better separate the classes of each data. After each class of model was fully tuned, various metrics were computed on the models. After analyzing these metrics along with the graphs generated from EDA, the multiclass model was reduced to a binary class model, where the new classes were 1 (women using contraceptives long term) and 0 (women using no contraceptives or using contraceptives short term).

2. Motivation

Research has shown that contraceptive choice is often associated with a variety of socioeconomic and demographic characteristics [3]. This research investigates the factors that most contribute to contraceptive method choice. Some of the factors investigated involved standard of living, level of education, and age. Different models were tested to see if high prediction accuracy was achievable given that ethical and social limitations might prevent this.

Moreover, there may be some bias in the dataset that hinders the prediction of contraceptive choice. This dataset was generated from Indonesia, which has the highest Muslim population out of any country in the world [1]. As a result, selection bias may be present since studies have shown that religious women are less likely to use contraception. The fact that all the women in the survey were married may also attribute to selection bias. Thus, models developed from this data may not generalize well if the targeted population is all women. Furthermore, a response bias may be present as Muslim women may be less likely to say they are using contraceptives due to religious stigma that may be associated with doing so [5].

Many machine learning models exist for classification, with three of the most popular being logistic regression, decision trees, and random forest trees [2]. In addition to predicting contraceptive choice, this research aims to eval-

uate the efficacy of these three models. Metrics such as accuracy, precision, and recall will be used to determine how much influence the socioeconomic and demographic characteristics of the surveyed women influence the contraceptive choice. Determining the factors that best influence contraceptive choice is important to expanding awareness of available contraceptives to groups that may benefit from contraceptive use. These groups may be inhibited from choosing the contraceptive choice that is best suited to their needs due to their socioeconomic and demographic backgrounds.

3. Exploratory Data Analysis

3.1. Data Cleaning

First, the data was checked for null values. None were found in the dataset. Much of the given data contained categorical variables. Although these features were ordinal and already presented in numerical form, it was necessary to one-hot-encode the categorical variables since there was no clear metric to determine the difference between two scores for a given feature. For example, it is subjective how worse a score of '1' for standard of living is then a score of '2'. The categorical features are: wife education, wife religion, wife work, husband occupation, standard living, media exposure. All these features were one hot encoded. In addition to one-hot-encoding, the continuous features were standardized since they had different units of measurement.

3.2. Feature Correlation

A correlation heatmap was generated to compute the correlation coefficients of the features in the dataset. Figure 1 shows the results of this correlation map. Correlation measures the association of two variables. For a model to be stable, the variance should be reasonably low. If the variance of the weights is high, then the model will be highly sensitive to noise in the data. If two variables are highly correlated then the variance of the weights will be large.

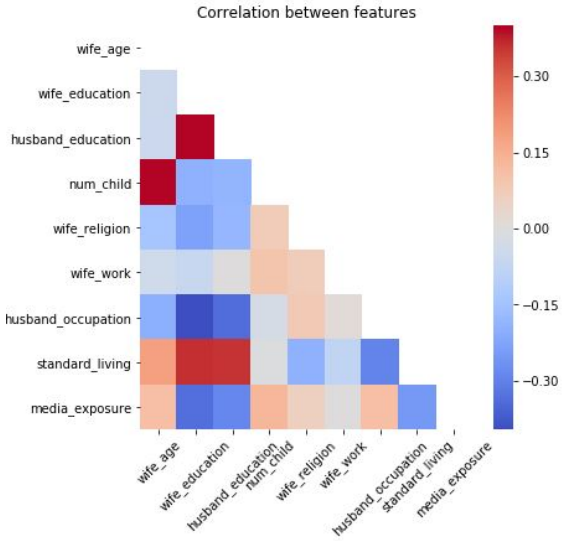

Figure 1. Correlation of Features

The resulting correlation map in Figure 1 shows that none of the features are highly correlated.

3.3. Principal Component Analysis

After cleaning the data, Singular Value Decomposition **(SVD)** was computed on a basic design matrix of the data that included the one-hot-encoded and standardized features. The designed matrix did not include the wife's religion as a feature. It was hypothesized that this would be irrelevant since the majority of Indonesia's population are Muslim. The singular values of the design matrix were used to analyze the proportion of variance that each principal component captured. These proportions are shown in Figure 2.

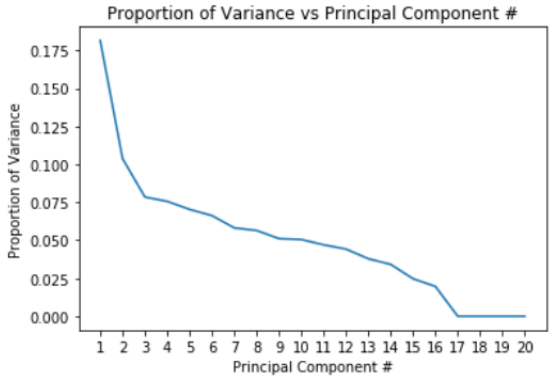

Figure 2. Proportion of variance contributed by each principal component.

These proportions were calculated to see if the data would be a good candidate for Principal Component Analysis **(PCA)**. PCA is a technique that can be used to reduce

irrelevant dimensions in data. Reducing irrelevant dimensions in data can prevent overfitting and speed up computation in learning algorithms. However, according to Figure 2, it appears that each principal component contributes a only small percentage of the variance in the data. PCA may have been useful if a small number of principal components represented a majority of the variance in the data. However, based on Figure 2, that is clearly not the case. This was further verified by projecting the design matrix onto the first two principal components. A graph showing the results of the projection is found in Figure 3.

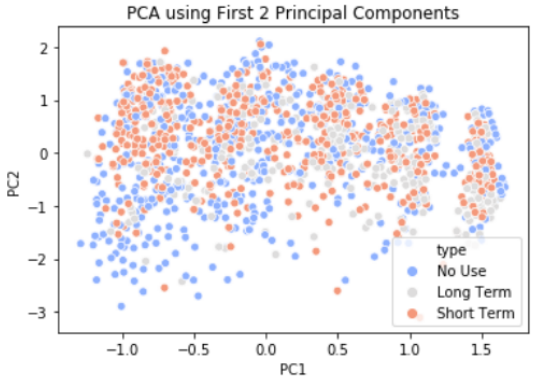

Figure 3. The figure shows the results of centering the design matrix and projecting it to the first two principal components.

It is difficult to discern any relationship between the data projected to the first two principal components. It appears that the variance in the data is widely spread among its features. As a result, it is unlikely that any overfitting in the models constructed will be attributed to the design matrix.

3.4. Data Visualization

First and foremost, the number of samples belonging to each class was calculated to see how the data was distributed. The data was observed to be unevenly proportioned, with most of the samples belong to the 'No Use' category. Figure 4 shows the results of this calculation.

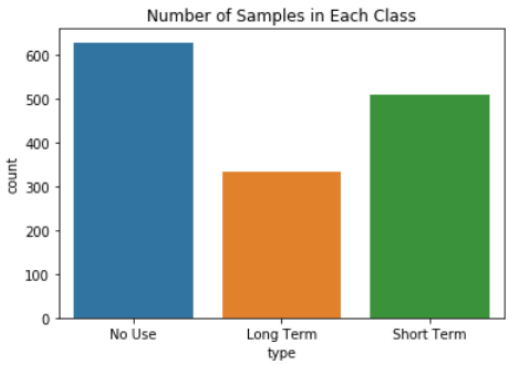

Figure 4. Number of samples in each of the three contraceptive classes.

Next, the distribution of the wife age was calculated for each contraceptive choice. Figure 5 shows the results of this calculation.

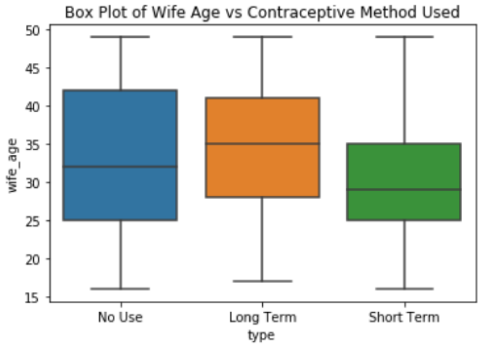

Figure 5. Distribution of wife age based on the type of contraceptive used.

It can be seen in Figure 5 that the median age is highest in women that used contraceptives Long Term, while the median age was lowest in women that used contraceptives Short Term.

The proportion of feature values in each of the three contraceptive classes were calculated for the data on religion, wife education, husband education, and standard of living. The results of the calculations can be found in Figures 6, 7, 8, and 9 respectively.

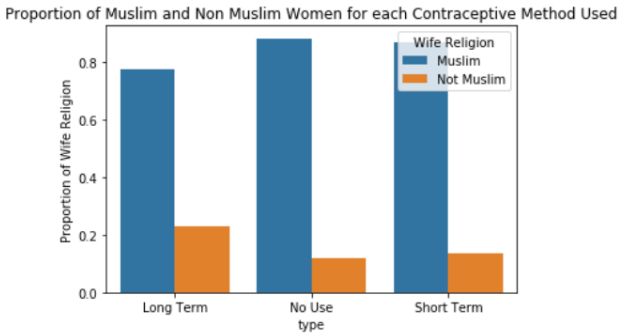

Figure 6. Proportion of Muslim and Non Muslim women for each contraceptive class.

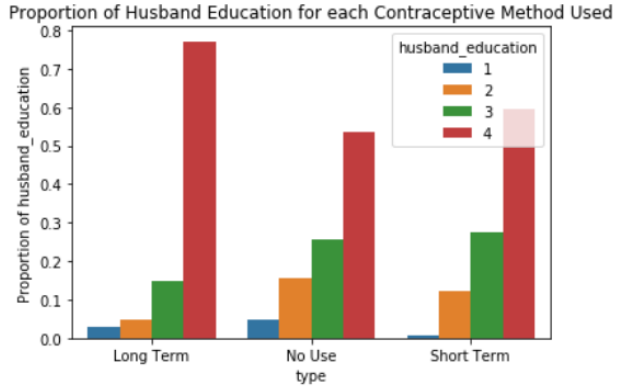

Figure 7. Proportion of the four possible levels of husband education for each contraceptive class.

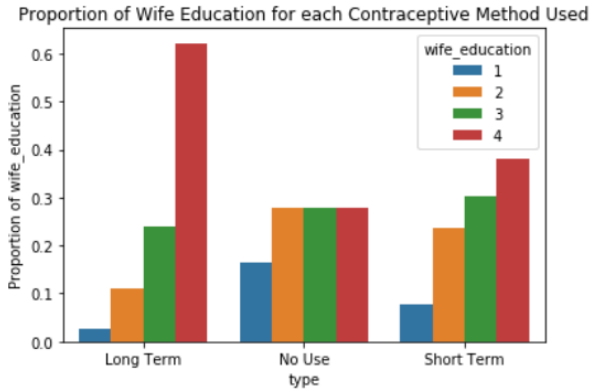

Figure 8. Proportion of the four possible levels of wife education for each contraceptive class.

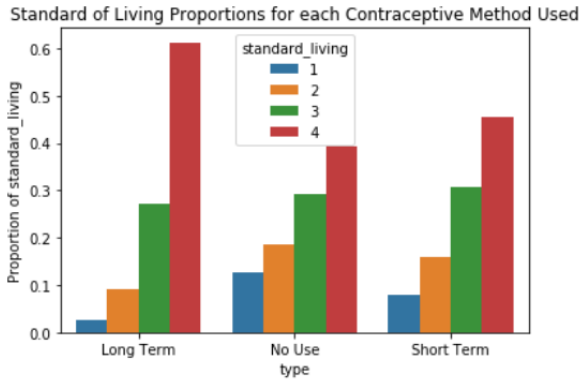

Figure 9. Proportion of the four possible values for standard of living for each contraceptive class.

Looking at Figure 6, there does not appear to be any significant difference in Muslim vs non Muslim women in reference to the type of contraceptive used. In other words, this feature won't give us any new useful information. This supports the hypothesis that wife religion has poor predictive power as a feature.

However, analyzing wife education, husband education, and standard of living in Figure 7, 8, and 9 all show very different distributions when analyzing the possible values for Long Term contraceptive use in comparison to Short Term or No Use. It is worth nothing that the Short Term and No Use distributions for these graphs all look fairly similar. However, in terms of the Long Term distributions, there is a disproportionate amount of women in the 4th category of feature values. This suggests that in terms of standard of living, wife education, and husband education, those that use contraceptives Long Term are much more likely to have high levels of education and standards of living in comparison to those in the No Use or Short Term class.

4. Models

4.1. Logistic Regression

After performing EDA and data cleaning, the refined data was split into 90% training data and 10% test data. Multiclass Logistic Regression was used as the first model attempt. Sklearn's LogisticRegressionCV was used to find the optimal inverse of regularization strength parameter.

	Training	Validation	Test
Accuracy	0.5154	0.5079	0.5405

Table 1. Training, Validation and Test Accuracy for Logistic Regression

As a first model attempt, logistic regression's accuracy was fairly low across training, validation and test data. Moreover, one can suspect that the model was underfitting due to its low training accuracy.

To further investigate this model, the average precision and recall were computed. The results are shown in Table 2. This was achieved by averaging the precision and recall for each class individually.

	Precision	Recall
Long Term	.56	.29
No Use	.63	.68
Short Term	.42	.52
Average	.53	.50

Table 2. Precision and Recall values evaluated on the test set of the final logistic regression model.

The average low precision indicates that the fraction of class predictions that actually belonged to the class they were assigned to was fairly low. This shows that few of the positive predictions for a given class actually captured that class. The No Use class had the highest precision, which was expected since it was the majority class. On the other hand, the low average recall indicates that a large amount of the positive values for a given class were never predicted.

Depending on the application behind this dataset, one can alter the probability threshold at which the classes are classified. To achieve higher precision, the probability threshold should be increased. On the other hand, if the user was interested in achieving higher recall, the probability threshold should be reduced.

4.2. Decision Tree

4.2.1 Attempt 1: Simple Naive Model

Sklearn's DecisionTreeClassifier was used to implement the decision tree model.

The classifiers hyperparameters were not explicitly assigned or tuned for the first attempt.

	Training	Validation
Accuracy	0.9472	0.4778

Table 3. Training, Validation and Test Accuracy for the naive decision tree.

Contrary to logistic regression, high training accuracy and low validation accuracy for the simple decision tree suggested that the model had overfit. This made sense, since deep decision trees tend to have high variance and low bias. This issue is often fixed by ensemble methods such as bagging or random forest. Both approaches rely on averaging the results of multiple decision trees, which tends to reduce variance. Before exploring these other options, fine tuning the depth of the tree and the minimum sample split was tested.

4.2.2 Attempt 2: Tuning Max Depth

The training and validation accuracy were computed for multiple classifiers with max depth ranging from 1 to 20. All other parameters remained untouched.

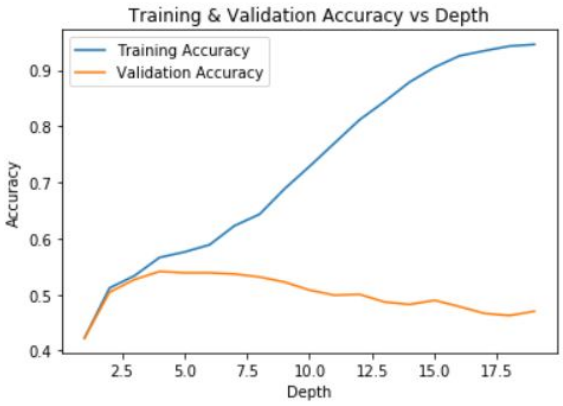

Figure 10. Training and Validation accuracy vs depth.

Table 10 shows the results achieved. It appears that the model was overfitting at high depths. On the other hand, validation accuracy was higher from depths in the 2 to 6 range. The highest validation accuracy was 0.5411 at depth 4.

4.2.3 Attempt 3: Tuning Min Samples Split

In this model, the maximum depth was fixed at 4 based on the previous results and the minimum number of samples required to split an internal node was tuned in an effort to reduce overfitting. Training and validation accuracy were computed and plotted in Figure 11.

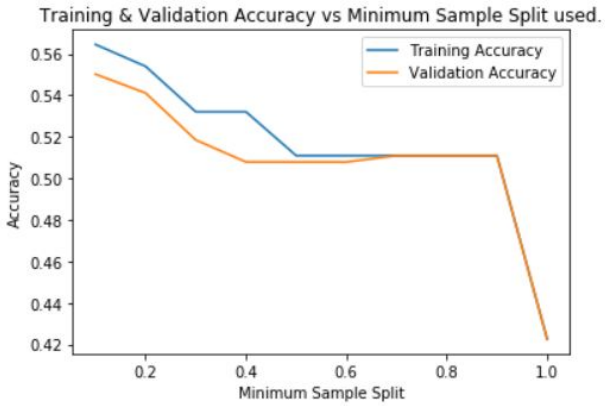

Figure 11. Training and Validation Accuracy vs Fraction of Minimum Samples Split.

The training accuracy dropped significantly which definitely reduced overfitting at the risk of underfitting. The

highest validation accuracy was 0.5426 with 0.1 fraction minimum sample split.

The highest validation accuracy was achieved so far as a result of tuning both parameters.

Next, to improve the prediction accuracy even more, random forests were explored.

4.3. Random Forests

4.3.1 Attempt 1: Simple Naive Model

Sklearn's RandomForestClassifier was used to build random forest classifier.

Random Forest is an ensemble of decision trees, usually trained with the "bagging" method. Bagging is a randomized method for creating many different learners from the same data set. Random Forest builds multiple decision trees and merges them together to get a more accurate and stable prediction. Random sampling on a subset of the features is done at every tree node. This guarantees that the same feature split is not used at the root of every tree.

The classifier's parameters and hyperparameters were not explicitly assigned or tuned for the first attempt.

	Training	Validation
Accuracy	0.9472	0.5162

Table 4. Training, Validation and Test Accuracy for Simple Random Forest

Compared to the naive model of decision trees, the training accuracy was lower and the validation accuracy was significantly higher. This is because random forests help reduce variance.

4.3.2 Attempt 2: Tuning Number of Trees

The training and validation accuracy were computed for multiple classifiers with number of trees ranging from 20 to 1000. All other parameters remained untouched. The results are shown in figure 12.

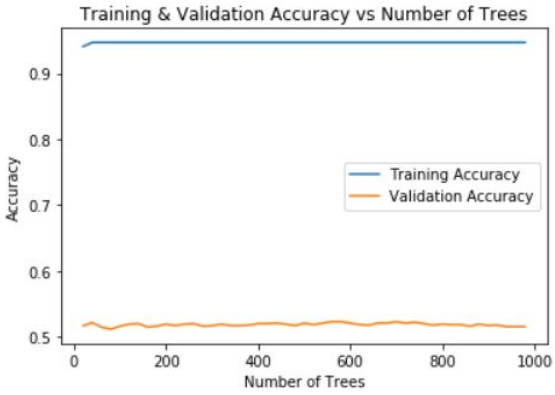

Figure 12. Training and Validation Accuracy vs number of trees.

The training and validation were roughly constant at 0.9471 and 0.5183 respectively. The highest validation accuracy was 0.5230 with 560 trees.

This value will be held constant in future model modifications.

4.3.3 Attempt 3: Tuning Max Depth

While keeping the number of trees constant at 560, the max depth was tuned from values ranging from 1 to 20. The results can be found in Figure 13.

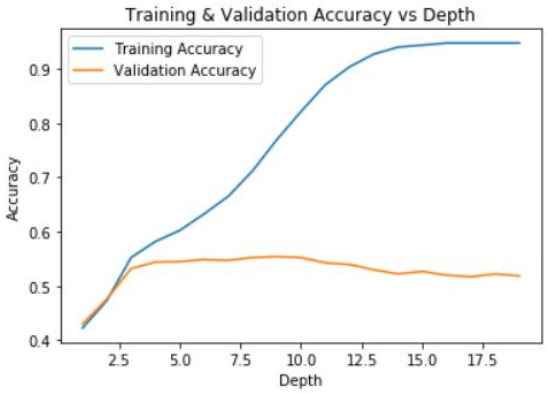

Figure 13. Training and Validation Accuracy vs max depth.

The results show that the training accuracy was increasing as maximum depth increased. The maximum validation accuracy was 0.5539 at depth 9.

As a result of tuning the maximum depth and number of trees, the highest validation accuracy, 0.5539, was achieved so far.

4.3.4 Attempt 4: Tuning Number of Trees and Max Depth Simultaneously

As a final attempt, Sklearn's GridSearchCV was used to find a more optimal combination for max depth and number of trees. Grid search is the process of performing hyper parameter tuning on a combination of hyperparameters in order to determine the optimal values for a given model. The grid search returned the highest validation accuracy with max depth and number of trees to be 6 and 110, respectively.

The validation score achieved by random forest with the above hyperparameters was 0.5524. There was no improvement in validation accuracy compared to the model in attempt 3.

All in all, it was determined that random forest with 560 trees and max depth 9 performed best with 0.5539 validation accuracy.

4.4. Feature Transformations

After tuning the max depth and number of trees on the random forest model, various transformations were performed on the continuous features of the data in order to lift it to a higher dimensional feature space. This was done in an attempt to add variance since the random forest model appeared to be underfitting. A total of 8 feature functions were tested on the new data which included: the original feature function without any modification, a feature function that applied a tanh function to the continous data, feature functions that applied polynomial features in conjunction with tanh, a feature function that lifted the data to a parabaloid, and a feature function that lifted the data to an ellipsoid. The resulting training and validation accuracy on the 8 feature functions is found in Figure 14.

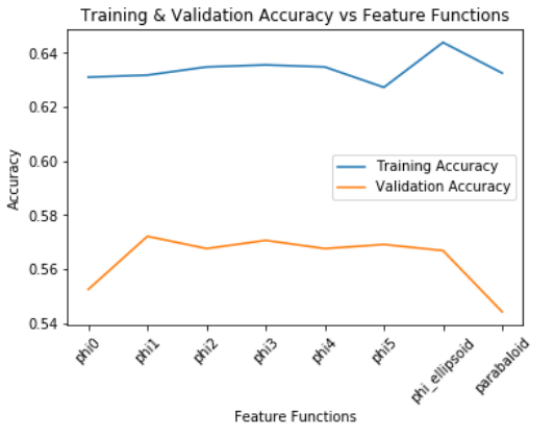

Figure 14. Training and Validation Accuracy on various feature functions.

Figure 14 shows that the best feature function was phi1 which applied a tanh function to the continuous features.

The hyperparameters returned by GridSearchCV and the ones returned by tuning max depth and number of trees individually were both tested as a final model using phi1. It was determined that the hyperparameters returned by GridSearch CV performed better after feature engineering. Thus, the final hyperparameters used were 110 and 6 for number of trees and max depth respectively.

As a final check, the religion feature was added back into the design matrix to see if it would improve validation accuracy. The validation accuracy decreased after to .5577 adding religion as a feature, which confirmed the hypothesis that religion had poor predictive power. The religion feature was left out in the final model.

The final model for the three contraceptive class case was the random forest model with feature function phi1 applied to the data. A max depth of 6 with 110 trees was used. The minimum number of samples in a leaf was left as the default value of 2, as tuning this value did not seem to improve accuracy.

Table 5 shows the training, cross validation, and test accuracy on the final random forest model used to predict the three contraceptive classes.

	Training	Validation	Test
Accuracy	0.6316	0.5720	0.6013

Table 5. Training, Validation and Test Accuracy for fine tuned Random Forest on three class case.

The precision and recall was calculated on the test data for the fined tuned Random Forest model on the three classes of contraceptives. The results are shown in Table 6.

	Precision	Recall
Long Term	.65	.37
No Use	.80	.62
Short Term	.45	.75
Average	.63	.58

Table 6. Precision and Recall values evaluated on the test set of the final random forest model for the three class contraceptive problem.

5. Two Class Case

5.1. Motivation

Analyzing the precision and recall in Table 6 showed some unexpected results in terms of precision when comparing the Long Term class to the Short Term class. Looking at the number of samples belonging to each class in Figure 4, it's expected that the Short Term class would have a higher precision then the Long Term class, since the algorithm had seen more examples of the Short Term class. It is worth noting that the recall performed as expected, since the values for recall were greater for the classes with greater sample representation. However, when taking a closer look at the feature distributions of wife education, husband education, and standard of living in Figures 7, 8, and 9 respectively, it can be hypothesized that these features yield stronger predictions for the Long Term class in comparison to the Short Term class. This is because the features have fairly distinct proportions for the Long Term class in comparison to the Short Term and No Use class. In fact, the proportions for the Short Term and No Use class are extremely similar for these features. Due to these similarities, it was hypothesized that women who used no contraceptives were very similar to those who used contraceptives short term. As a result, the model was transformed from three classes to two, where the Short Term and No Use classes were combined into a single class. The identifying numbers for the two classes were 1 (Long Term) and 0 (No Use/Short Term). A count plot shown in Figure 15 was produced showing the number of samples in each class.

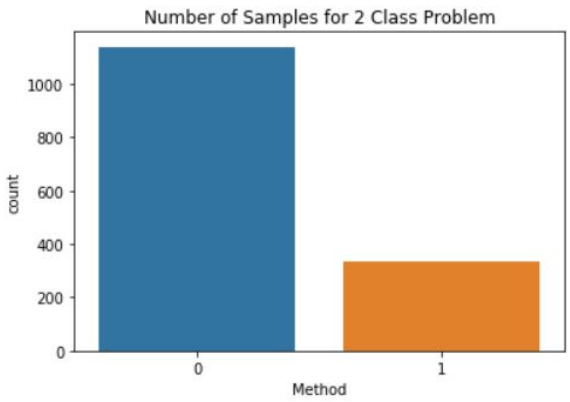

Figure 15. Number of samples in each class. Class 0 made up 77.4 % of the data while class 1 made up 22.6 % of the data.

5.2. Performance

The previous best random forest classifier with value 6 as the maximum depth parameter and value 110 as number of trees parameter was used to classify the two class model. The results are shown in Table 7.

	Training	Validation	Test
Accuracy	0.7750	0.7886	0.7376

Table 7. Training, Validation and Test Accuracy for fine tuned Random Forest on two class case.

As shown in the table above, the validation and test accuracy significantly increased. Collapsing the No Use and Short Term classes into one due to similar feature values resulted in a more accurate model.

Table 8 shows the precision and recall values for the two class case.

	Precision	Recall
0	0.78	0.98
1	0.25	0.02
Average	0.51	0.50

Table 8. Precision and Recall values evaluated on the validation set of the final random forest model for the two class contraceptive problem.

The results show that optimal random forest classifier for the two class problem achieved high precision and recall for class 0 while low values for class 1. This was expected since based on Figure 15, most data points belonged to class 0.

5.3. Undersampling, Oversampling, and SMOTE

To improve the precision and recall of class 1 (Long Term), three techniques were carried out to evenly distribute the data between classes. The first was undersampling, which randomly deletes portions of the over saturated class until it is equal in size to the class that is least represented in the data. The second technique used was oversampling, which makes copies of random sample points belonging the class with the least representation until the proportion matches the majority class. The final technique used was the Synthetic Minority Oversampling Technique **(SMOTE)**, which employs the K Nearest Neighbors algorithm to generate synthetic samples of the minority class. Cross validation was performed on the fine tuned random forest model using the resulting datasets from the three sampling techniques. The validation accuracy for undersampling, oversampling, and SMOTE were 0.6527, 0.6819, 0.7119 respectively. SMOTE seemed to perform the best, which was expected since it is the most popular of the three techniques due to its effectiveness. The precision and recall were calculated on the validation data from SMOTE sampling. The results are shown in Table 9.

	Precision	Recall
0	0.86	0.78
1	0.42	0.57
Average	0.64	0.68

Table 9. Precision and Recall values were evaluated on the fine tuned random forest model for the two class model using the balanced data generated from SMOTE.

While applying SMOTE to the data more evenly distributed the precision and recall values in comparison to the model trained on the original unbalanced data, the unbalanced data still performed better in terms of accuracy. For the purposes of this paper, accuracy was the most important metric. Thus, SMOTE was not used to train the final model referenced in Table 7. However, there may be an application of the data that would prefer to have a higher precision and recall for class 1. For example, if a specific application placed high importance on predicting Long Term contraceptive use, applying SMOTE to the data would likely be a better option then using the original data.

It was initially hypothesized that the improvement from the 3 class problem to the 2 class problem was mainly attributed to the over saturation of data belonging to class 0. However, this was likely not the case since the model trained on evenly distributed data from SMOTE was similar in terms of validation accuracy in comparison to the model trained on the original, unbalanced data.

6. Conclusion

This paper aimed to predict contraceptive use in women based on the *Contraceptive Method Choice Data Set* from the University of California Irvine's Machine Learning Repository [4]. The performance of three types of classification models - logistic regression, decision trees, and random forest - were evaluated on completing this task. The

initial three classes of contraceptive use were No Use, Short Term Use, and Long Term use. The most successful model of the three proposed was the random forest model. The model was fine tuned to optimize its performance. The final model used 110 learners and had a maximum depth of 6. The model was trained on data with various feature functions applied to it. All of the feature functions standardized the numerical features in addition to one-hot-encoding the categorical features. The functions varied in the transformations that were performed on numerical features. The best performing feature function applied a tanh function to the numerical features. The resulting training, validation, and test errors on the final model applied to the three class problem were 62.72 %, 56.91 %, and 56.60 % respectively. The models performance models was relatively low. This was likely attributed to a variety of factors. One issue was that the dataset was relatively small. It contained only 1473 samples in total. Furthermore, it was difficult to add variance to the data, since there were only 2 non categorical features in the entire dataset. A deeper analysis of the features showed that the No Use class was very similar to the Short Term class. As a result, the these classes were combined into a single class of Short Term or No Use.

The two class problem performed much better on the fine tuned random forest model. The training, validation, and test accuracy were 77.50 %, 78.86 %, and 73.76 %, respectively. The precision and recall values for class 1, shown in Table 8, were very low. This was likely attributed to a lack of representation in the data for class 1. It was hypothesized that the model was mainly performing better due to the over representation of class 0. Undersampling, oversampling, and SMOTE sampling were separately performed to balance the data, which improved the precision and recall for class 1. SMOTE performed the best out of the three sampling techniques. While the validation accuracy for the random forest model trained on the balanced data was less accurate then the unbalanced data, its validation accuracy was still a significant improvement over the two class case.

7. Future Work

Future work to improve the models performance may include collecting more data with added features. Specifically, things such as insurance type, location, and race may add some useful information to the model. Collecting data with more numerical features would be beneficial, as it may allow for more complicated decision boundaries as a result of feature engineering. However, ethical concerns may arise based on how the data is collected and interpreted. Results of a model based on socioeconomic factors may negatively reinforce stereotypes. Furthermore, the data that is being collected is very personal. Any new data should be collected anonymously to limit any response bias that may occur. Additionally, data that spans multiple countries would likely reduce the chance of any selection bias that may occur. After all, one of the biggest issues in the data was that there was simply not enough variance in the data.

References

[1] Islam by country. https://en.wikipedia.org/wiki/Islam_by_country. Accessed: 2020-05-07.

[2] Top 10 machine learning algorithms. https://www.dezyre.com/article/top-10-machine-learning-algorithms/202. Accessed: 2020-05-07.

[3] Jennifer J. Frost. Factors associated with contraceptive choice and inconsistent method use, united states, 2004. https://www.guttmacher.org/journals/psrh/2008/factors-associated-contraceptive-choice-and-inconsistent-method-use-united. Accessed: 2020-05-07.

[4] Tjen-Sien Lim. Contraceptive method choice data set. https://archive.ics.uci.edu/ml/datasets/Contraceptive+Method+Choice. Accessed: 2020-05-05.

[5] Claire Scrivani. Attitudes towards and use of contraception in the muslim refugee population. https://med.virginia.edu/family-medicine/wp-content/uploads/sites/285/2018/10/Claire-Scrivani_Attitudes-Towards-and-Use-of-Contraception-in-the-Muslim-Refugee-Population_Final.pdf. Accessed: 2020-05-07.