Predictors of divorce and performance of various ML models in predicting divorce status

1. Introduction

This project aims to predict divorce status based on Divorce Predictors Scale (DPS) developed by Yöntem and İlhan (2017, 2018) on the basis of Gottman couples therapy (Gottman, 2014). Gottman listed four most important reasons for divorce: Criticism, Contempt, Stonewalling and Defensiveness, along with the "failure to repair" (Gottman, 1999; Gottman and Silver, 2015). Over the course of this project, we will explore important DPS factors that are related to these four reasons.

2. Dataset and methods

The divorce dataset (UCI Machine Learning Repository: Divorce Predictors data set Data Set) includes the data of 170 couples from Turkey, 84 of whom were divorced and 86 were married. The answers to the questionnaire based on DPS consists of 5 subscales (0=Never, 1=Seldom, 2=Averagely, 3=Frequently, 4=Always).

I will build prediction models using the following algorithms: Decision tree algorithm, Random Forest, Linear SVC, K-nearest neighbors for prediction models; Principal Components Analysis combined with Linear SVC. The dataset will be split into a training set (80%) and a test set (20%), and the performance of each algorithm will be evaluated using accuracy scores. Since this relatively small dataset includes only 170 data points over 54 attributes, it's prone to high variance in performance metrics so the accuracy score for some models might vary from one run to another. So for all models, I computed the average accuracy scores over 100 runs, and repeated this process 10 times. Variable importance plots and decision trees are also used to identify important predictors of divorce.

All analysis is done using Python, and machine learning algorithms are fitted with the scikit-learn package. All 54 explanatory variables are ordinal, so no preprocessing or conversion of data types are necessary. There are also no missing values. For a support vector classifier model, principal components analysis was also used to reduce the data to two components, which explain ~77% of the variance.

3. Building prediction models

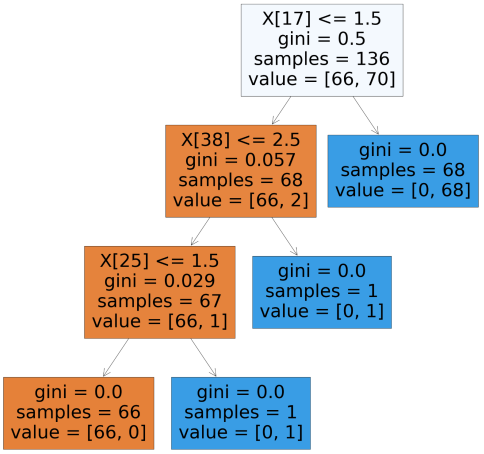
a. Decision Tree Classifier

Decision Tree Classifier was fitted without changing any of the default settings (using Gini coefficient for node impurity, choosing the best splitter, not limiting the maximum number of leaf nodes). Six different examples of decision trees are as follows, all of which have accuracy scores of at least 94% (Figure 1). The orange node consists mostly of class 0 (not divorced), the blue node consists mostly of class 1 (divorced). In all the trees, we see that the root node is always Attr18. Then two trees determine the next best splitter to be Attr26, while the other three trees have Attr25, Attr39, Attr40. The fifth tree has Attr3 to be the next best splitter after Attr26.

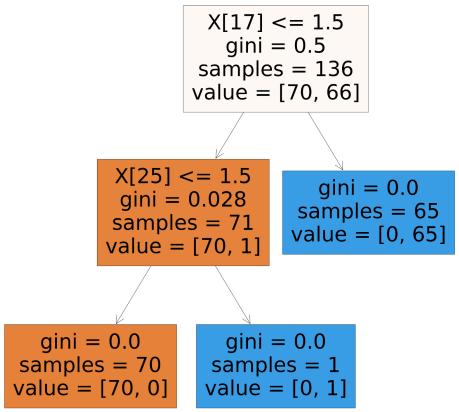
Attr3	When we need it, we can take our discussions with my spouse from the beginning and correct it.

Attr18	My spouse and I have similar ideas about how marriage should be.
Attr25	I have knowledge of my spouse's inner world.
Attr26	I know my spouse's basic anxieties.
Attr39	Our discussions often occur suddenly.
Attr40	We're just starting a discussion before I know what's going on.

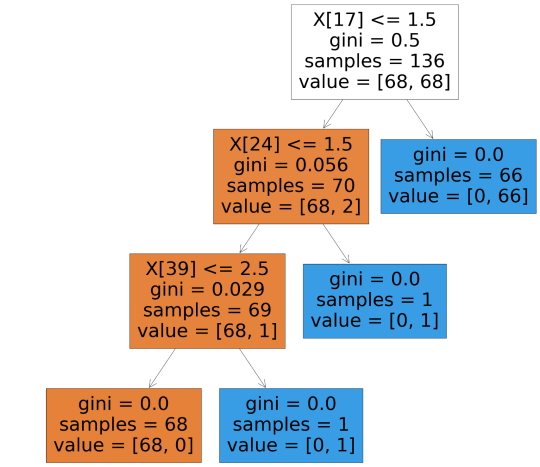
Accuracy: 97.06%



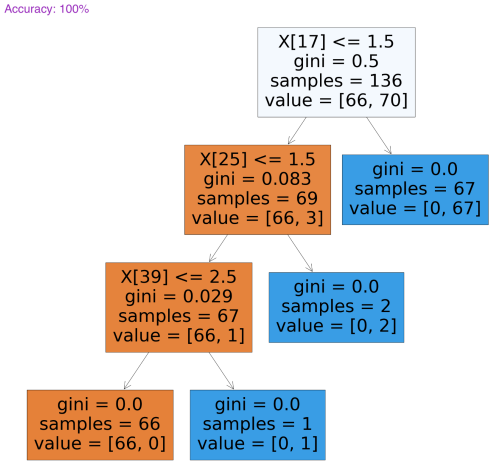
Accuracy: 97.06%



Accuracy: 94.12%



Accuracy: 100%



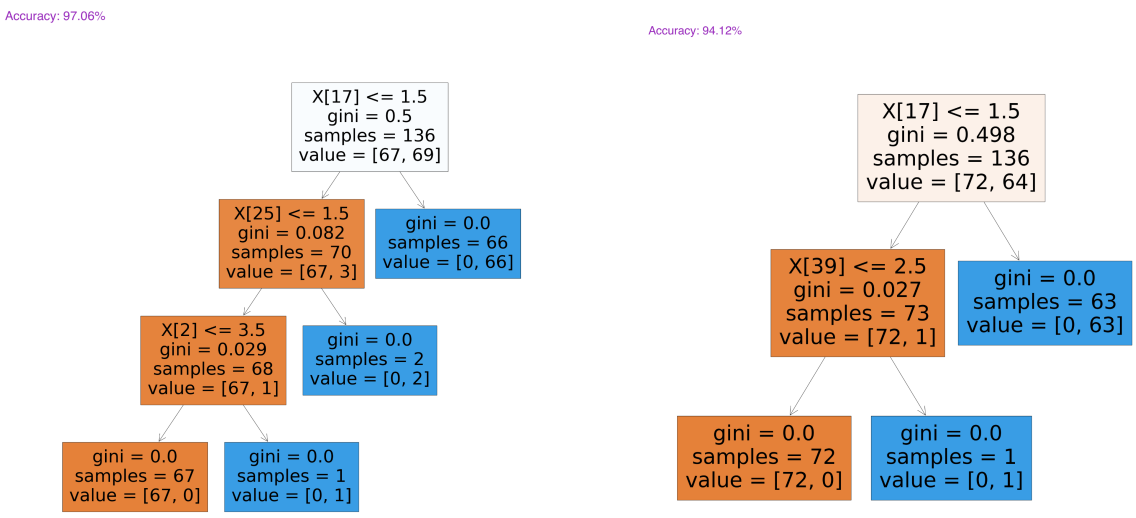


Figure 1: Six different decision trees with default settings. The orange node denotes mostly of class 0 (not divorced), the blue node denotes mostly of class 1 (divorced). Depth of these trees is at most 3.

Attr3 and Attr18 are related to failure to repairs and finding a common ground. Attr25 and Attr26 are related to understanding your partner's anxieties and loving your partner. Attr39 and Attr40 are related to negative conflict behaviors. There is an overlap of the above features with the most significant features found using the correlation-based selection method (Yontem et al., 2019), where Attr18, Attr26 and Attr40 are among top 6 most significant features. It is important to note that decision tree algorithms are easy to interpret, but are prone to instability.

b. Random Forest Classifier

We can also use a random forest classifier to generate 200 random trees, and the tree with the most votes is chosen. Accuracy scores are computed to access the model's performance. This method has slightly better predictive performance than decision trees, but it is difficult to interpret the effect of each predictor. Due to the stochastic nature of random forest and the variables' close values of importance, no stable variable importance plot can be achieved. Six examples of variable importance plots were produced using random forests of 1000 trees (Figure 2).

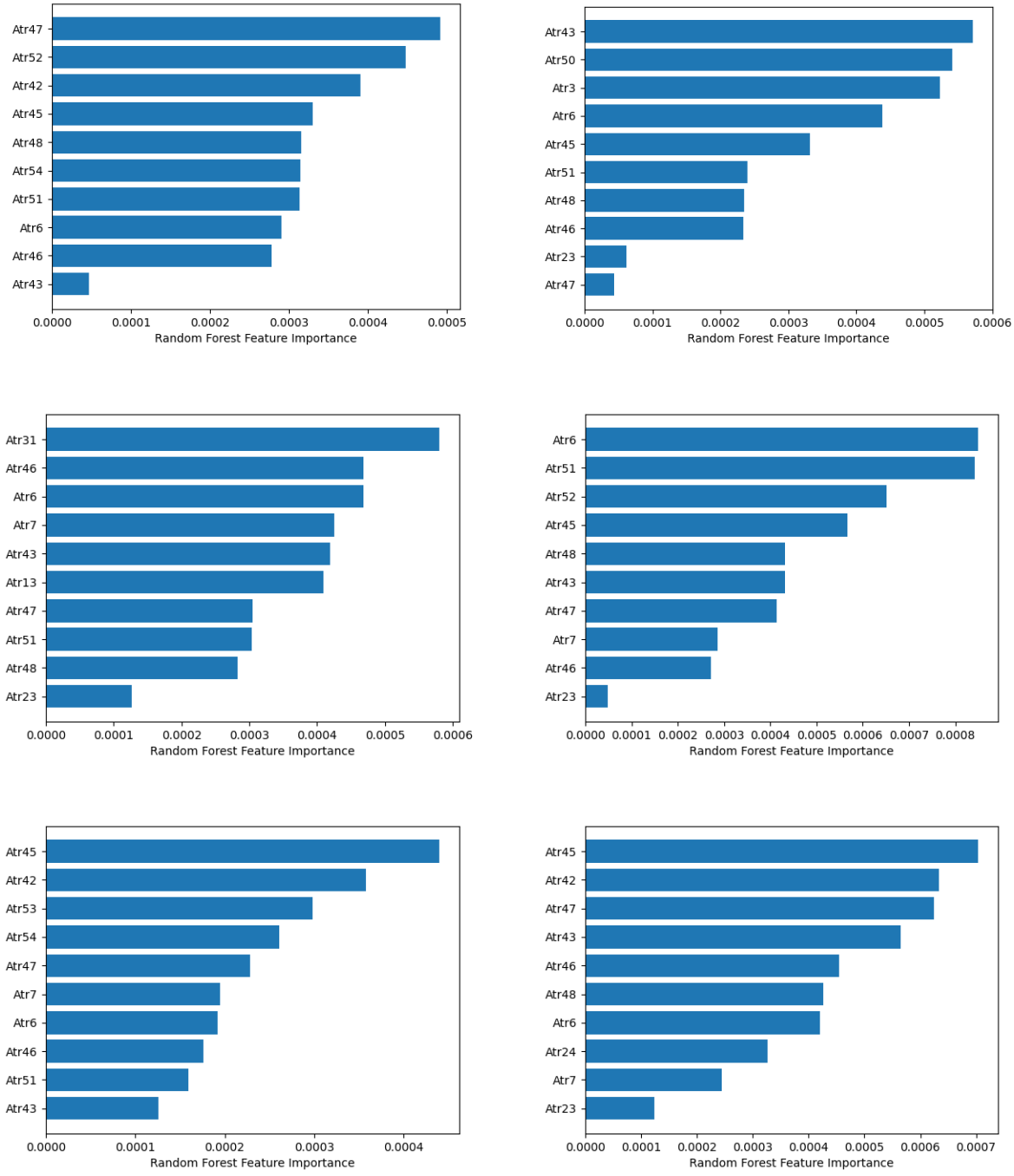


Figure 2: Six variable importance plots from random forests of 1000 trees. Only top 10 variables are shown.

We notice that Attr6, Attr43 Attr46, Attr47 appear in all of the plots.

Attr6	We don't have time at home as partners.
Attr43	I mostly stay silent to calm the environment a little bit.

Attr46	Even if I'm right in the discussion, I stay silent to hurt my spouse.
Attr47	When I discuss with my spouse, I stay silent because I am afraid of not being able to control my anger.

Attr6 refers to failure to repairs and spending quality time together, while Attr43, Attr46 and Att47 are all related to staying silent during a conflict, which is a negative conflict behavior.

c. Support Vector Classifier (SVC)

The SVC model was fitted against the data with a linear kernel. Its performance is better than random forest classifier. The following plot is produced with a linear SVC model fitted against the dimensionally reduced data of two principal components.

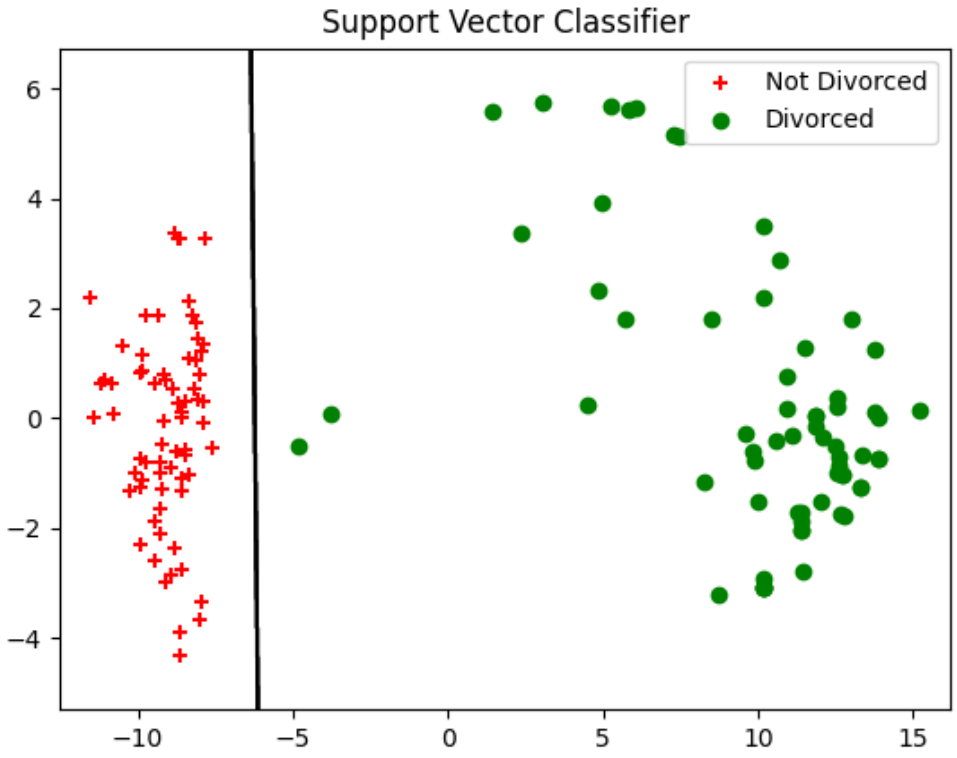


Figure 3. Linear SVC fitted against reduced data of two principal components.

d. K-nearest neighbors Classifier

This model was built using 5 nearest neighbors due to the small dataset.

e. Logistic Regression Classifier

This model was fitted with the default L-BFGS solver due to the small dataset. The L-BFGS solver makes use of the classic Newton's Method (Gradient descent using quadratic loss function), but uses an estimated inverse Hessian matrix which saves memory.

3. Comparison of performance:
For each model mentioned above, I computed the 10 mean accuracy scores over 100 runs, each of which split a new random training set and test set, then computed the overall average accuracy score. The linear SVC model has the best performance in terms of accuracy, followed by the logistic regression classifier.

Model	Mean accuracy scores over 100 runs	Average
Decision Tree Classifier	[0.9679, 0.9694, 0.9688, 0.9741, 0.97, 0.975, 0.9709,0.9738, 0.9679, 0.9726]	0.971
Random Forest Classifier	[0.9765, 0.9797, 0.9753, 0.9774, 0.9821, 0.9729, 0.9726, 0.9774, 0.9732, 0.9782]	0.9765
Linear SVC	[0.9835, 0.9832, 0.9824, 0.9797, 0.9794, 0.9821, 0.985, 0.9803, 0.9829, 0.9838]	0.9822
K-nearest neighbors Classifier	[0.9788, 0.9768, 0.9774, 0.9759, 0.9774, 0.9788, 0.9771, 0.9762, 0.9803, 0.9774]	0.9776
Logistic Regression Classifier	[0.9762, 0.9765, 0.9818, 0.9788, 0.9821, 0.9774, 0.9832, 0.9824, 0.9803, 0.9759]	0.9795

4. Discussion and Conclusion

a. Predictors of divorce

The decision tree and random forest classifiers identify the following reasons of divorce:
i. Failure to repair or find a common ground: Attr3, Attr6, Attr18.
ii. Lack of understanding and love, or not having the same values: Attr25, Attr26.
Both of these might lead to criticisms and even contempt towards each other due to the lack of understanding and unwillingness to fix anything that went wrong.
iii. Negative behavior during conflicts: Attr39, Attr40 (getting into a conflict without knowing); Attr43, Attr46 and Att47 (staying silent, which is related to “Stonewalling”). These behaviors make sure that conflicts could never be resolved because either partners are not aware of what they are fighting about or don’t want to talk to resolve their conflicts.

b. Evaluation of ML models performance

The models are all subject to instability due to a small data set that leads to a small test set. Variables are also closely related to each other in groups, so there are a few equally good splits for the tree models. Depending on how the training set and test set are split, different but similar-in-meaning predictors can be chosen. For this dataset, I also found that tuning hyperparameters does not improve the models by much. Within the same training set and test set, changing the maximum number of leaf nodes for the decision trees, or changing the number of neighbors in K-nearest neighbors did not improve or worsen the model performance. So a maximum number of leaf nodes were not set for the decision tree model, and 5 nearest neighbors were set for the K-nearest neighbors model.

For tree models, there is not much need to tune maximum number of leaf nodes, maximum depth or even minimum samples to split as long as the best splitter is chosen. The maximum depth of trees is at most 3, which is a pretty reasonable size. More complicated trees are produced using the stochastic random forest algorithm, but maximum depth was purposely not tuned so as to identify potentially important predictors that are ignored in decision trees.

Overall, there are no major differences among the performance of the ML models, with the linear SVC model being slightly better than others.

5. References

Yöntem, M , Adem, K , İlhan, T , Kılıçarslan, S. (2019). DIVORCE PREDICTION USING CORRELATION BASED FEATURE SELECTION AND ARTIFICIAL NEURAL NETWORKS. Nevşehir Hacı Bektaş Veli University SBE Dergisi, 9 (1), 259-273.

Yöntem, M.K. and İlhan, T. (2018). Boşanma Göstergeleri Ölçeğinin Geliştirilmesi. [Development of the Divorce Predictors Scale]. Sosyal Polika Çalışmaları Dergisi. 41, 339-358.

Gottman, J. M. (1999). The Marriage Clinic: A Scientifically-Based Marital Therapy. New York: WW Norton and Company.

Gottman, J. and Silver, N. (2015). Evliliği Sürdürmenin Yedi İlkesi.(trans. Gül, S.S.). [The Seven Principles for Making Marriage Work. 1999] İstanbul: Varlık Yayınları.

Gottman, J. M. (2014). What Predicts Divorce? The Relationship Between Marital Processes and Marital Outcomes. New York: Psychology Press.