KAGGLE-DIVORE-PREDICTION

Sanel Štein, Karl-Markus Hannust

Github link: https://github.com/Sanel-Stein/IDS-Divorce-Prediction

Business understanding

Background

Divorce is a social issue that impacts individuals and families. Identifying the factors that contribute to the dissolution of marriages can help professionals develop better support systems and prevent relationship breakdowns.

In this project, we aim to explore the root causes and key predictors of divorce through data analysis.

Business goals

The primary business goal of this project is to gain a deeper understanding of the factors that contribute to divorce.

A secondary business goal is to develop predictive tools that can estimate the likelihood of divorce based on available data. Accurate prediction models can help professionals recognize early warning signs and allocate resources more effectively.

Together, these goals help turn data into useful insights that can support healthier and more stable relationships.

Business success criteria

The project will be considered a success if our predictive model achieves at least 85% accuracy on the test set.

Inventory of resources

Resources available for the project:

- Primary dataset with 5000 records,
- Secondary dataset with 170 records,
- Two HP laptops.

Requirements

- Access to the datasets,
- Team collaboration.

Assumptions

- The datasets are accurate, complete, and representative of real-world scenarios,
- Features included in the datasets are sufficient to identify patterns and predict divorce.

Constraints

- Limited number of datasets,
- The project must be completed by 8.12.2025.

Risks and contingencies

- Time constraints: The project must be completed by the submission deadline.
  - Create a project schedule.
- Limited dataset size: Small datasets may reduce model accuracy or generalizability.
  - Use appropriate techniques.

Terminology

- Divorce: The legal dissolution of a marriage between two individuals.
- Predictive Modeling: Using statistical or machine learning techniques to forecast outcomes based on input data.
- Dataset: A collection of structured data containing records (rows) and features (columns) for analysis.
- Accuracy: A metric for evaluating model performance, representing the proportion of correct predictions.

Costs

Irrelevant to our project

Benefits

Educational value: Team members gain practical experience in data analysis, machine learning

Data-mining goals

- Identify key features in the dataset that most strongly correlate with divorce outcomes,
- Preprocess the data by handling missing values and correcting inconsistencies,
- Explore the data,
- Train predictive models capable of classifying whether a couple is likely to divorce based on the available features,
- Evaluate model performance using accuracy,
- Select the best-performing model,
- Interpret model results in a way that provides meaningful insights into which factors influence the likelihood of divorce.

Data-mining success criteria

- The final predictive model achieves at least 85% accuracy on the test dataset.
- Missing values, duplicates, and inconsistencies are identified and handled properly.
- All analysis steps (preprocessing, modeling, evaluation) can be reproduced using the project code and documentation.

Data understanding

Gathering Data

The data that is needed for making an accurate prediction model are numeric and/or categorical features and the label. The features should represent different aspects of the relationship between the couples, such as how good the communication is between the couple, their conflict resolution, shared hobbies, etc. The label should be binary, true meaning that they divorced and false meaning that they did not. The data format should be CSV, and the time range is not important, as we use the same features for couples who divorced and who did not. The sample size should be at least several thousand couples and features should be consistent and there should be no null values.

If a dataset does not include certain features that we consider important, we will substitute it with another Kaggle dataset, that does include those features. We verified that there are several datasets available in Kaggle that satisfy our requirements.

We plan to use one dataset that contains 5000 samples and 20 features. The second dataset that we plan to use contains 170 samples and 54 features.

Describing data

As mentioned, we got our data from Kaggle in CSV format. The first dataset contains 5000 samples and 20 features, some of which are numeric and a few of which are categorical. It includes all features that we consider necessary, such as information about couples' conflicts, similar hobbies and whether their backgrounds are similar. The second dataset contains 170 samples and 54 features, all of which are numeric. The sample size is quite small, relative to the feature size thus it requires some processing due to the curse of the dimensionality. We have several options to deal with that problem. We can reduce the size of the features, increase the size of the samples or do a mix of the two to solve the dimensionality problem.

Exploring data

The dataset that contains 5000 samples and 20 features satisfies all our requirements, so we use it as is. The second dataset that contains 170 samples and 54 features has a lot of ambiguous and almost identical features. The features can also be divided into around five to six different categories. We plan to combine the identical features, drop the unambiguous features and for each category, we choose one to three features that we do not drop.

Verifying data quality

Although there are some problems with the data quality, we believe that the data quality is sufficient to support our goals. In the event that the data quality is not sufficient, we plan to substitute our second dataset with another dataset from Kaggle.

Project plan

Task 1 - Data Collection & Initial Review

Download datasets, check structure, identify missing values, understand attributes. Both members: 3 hours each.

Task 2 - Data Cleaning & Preprocessing

Handle missing values, remove duplicates, encode categorical variables, scale features if needed. Both members: 5 hours each.

Task 3 - Exploratory Data Analysis

Generate plots, correlations, statistical summaries to identify patterns. Both members: 4 hours each.

Task 4 - Model Training & Evaluation

Try multiple machine learning algorithms, compare metrics, tune hyperparameters. Both members: 6 hours each.

Task 5 - Interpretation of Results

Identify key predictors, analyze feature importance, check model bias or overfitting. Both members: 3 hours each.

Task 6 - Final Presentation

Finalize project, create poster, rehearse presentation. Both members: 3 hours each.

Methods

- Data cleaning techniques: dropping rows with missing values, encoding, normalization
- Machine learning models:
  - Linear Regression
  - Decision Tree
  - Random Forest
- Model evaluation metrics: accuracy, precision, AUC

Tools

- Python 3
- Jupyter Notebook
- Python libraries:
  - pandas
  - numpy
  - scikit-learn
  - matplotlib
  - seaborn
- GitHub
- Canvas