Divorce prediction

Using Logistic regression and decision trees Models

Group – 46

- ***Dias B.R.S.T – EG/2020/3893***
- ***Dissanayake D.K.R.C.K – EG/2020/3910***

IEEE
Advancing Technology for Humanity

Content

- *Introduction*
- *Background*
- *Ethics*
- *Data Preprocessing*
- *Models used*
- *Model Performance*
- *Conclusion*
- *References*

IEEE Advancing Technology for Humanity

The professional home for the engineering and technology community worldwide

Introduction

Importance

- This system would be put forward to clearly predict what would be the outcome that could happen after a marriage between couples.

- This could help people understand their relationships better and give professionals useful insights to help couples stay together. Ultimately, our goal is to find ways to strengthen marriages and lower the chances of divorce based on what we discover.

IEEE
The professional home for the engineering and technology community worldwide

Introduction

Objective

- The main objective of this project is to apply machine learning concepts to predict whether there will be a divorce occur between couples in a certain population. The difficulty that could be faced in here is to read the couples minds which will led to do uncertainty in their relationships due to factors such as economical, emotional, social and mental circumstances.

- As this is a supervised, Logistic regression and decision trees are used as the training models.

IEEE
The professional home for the engineering and technology community worldwide

Background

Source & Description

- Our dataset originates from Kaggle, a popular platform for data science and machine learning datasets. Specifically, it's a compilation of labeled data created at Stratosphere labs, offering insights into harmful or potentially malicious network activities.
- Link - https://www.kaggle.com/datasets/andrewmvd/divorce-prediction
- Key features:

1.Understanding
2. Acceptance
3. Openness
4. Responsiveness
5. Volubility
6. Time management
7. Connection
8. Time allocating
9. Priority
10. Similarities

IEEE
The professional home for the engineering and technology community worldwide

Data Preprocessing

Preprocessing methods

- Missing Values Handling
- Checking Null values
- Checking numerical and categorical data
- Dimensionality Reduction using Principal Component Analysis

IEEE
The professional home for the engineering and technology community worldwide

Data Preprocessing

Data Preprocessing Using Principal Component Analysis

Why we need to apply Principal Component Analysis (PCA) for preprocessing in this project?

- Dimensionality Reduction:
  - When the dataset has many features.
  - PCA reduces dimensionality by creating uncorrelated variables (principal components).
  - Captures essential information.
  - Benefits: Improves model efficiency and lowers overfitting risk.
- Multicollinearity:

  PCA is effective in handling multicollinearity, which occurs when features are highly correlated with each other. High multicollinearity can lead to instability in model estimation. PCA addresses this issue by creating orthogonal (uncorrelated) principal components.
- Visualization:

  PCA can be used for visualizing high-dimensional data in a lower-dimensional space, making it easier to explore and understand the structure of the data.

IEEE
The professional home for the engineering and technology community worldwide

Data Preprocessing

Data Preprocessing Using Principal Component Analysis

Before Preprocessing

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	...	Q45	Q46	Q47	Q48	Q49	Q50	Q51	Q52	Q53	Q54
12	3	4	3	4	3	0	1	4	3	4	...	4	4	4	4	4	4	4	4	4	4
150	0	3	2	0	0	0	0	0	0	0	...	4	4	4	3	1	1	1	2	0	1
149	0	1	0	0	0	0	0	0	0	0	...	4	4	4	4	4	2	2	0	0	0
28	3	4	3	2	3	0	1	4	3	2	...	4	4	4	4	4	4	4	4	4	4
156	0	0	1	1	0	0	0	0	0	2	...	0	2	1	2	1	2	2	1	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
133	1	2	0	0	0	0	0	0	0	0	...	0	2	1	2	2	2	2	2	1	0
137	0	0	1	0	0	0	0	1	1	0	...	3	3	3	3	0	1	3	3	3	1
72	3	3	3	3	3	1	1	3	3	3	...	3	3	3	3	3	3	3	3	3	3
140	0	2	0	0	0	1	0	0	0	0	...	1	1	2	2	1	0	1	3	2	2
37	3	3	2	3	3	1	1	3	3	3	...	3	3	4	4	4	4	3	3	4	4

118 rows × 54 columns

After Preprocessing (10 components)

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
12	14.066466	-0.065186	-0.810092	1.391413	0.603331	-1.032619	-1.856151	-0.928209	-0.119226	-1.371791
150	-8.582718	-3.543475	-2.860693	-0.718076	0.227951	-0.035169	-2.338792	-1.715401	1.697381	0.379319
149	-8.548589	-4.145983	-4.055091	-2.128593	2.438771	0.097365	-0.728773	1.263579	1.254430	-0.350033
28	12.748873	-1.135939	-0.047615	1.160251	0.003013	-1.158083	-2.667211	0.338645	-0.025645	1.251005
156	-7.870086	1.805395	0.221772	-1.231504	0.141632	-1.062326	-0.063183	-1.393429	0.010069	-0.962790
...	...	...	...	...	...	...	...	...	...	...
133	-8.495490	0.508305	0.541573	1.314817	1.032915	-1.834335	0.609724	0.437751	-0.027999	0.102420
137	-8.845278	-2.717350	0.367302	1.781332	-0.159814	1.269320	-1.127132	-1.365665	1.091692	-0.039301
72	9.032262	1.283010	-1.104992	1.114686	0.073987	-0.099858	0.305795	-0.213731	0.215465	-0.344653
140	-8.453574	-0.518549	0.465545	1.019478	-1.699143	-0.461717	-1.131805	-0.269720	-0.752292	0.339372
37	12.038091	0.128717	0.432466	0.442447	-0.363001	-1.164858	0.138097	-0.172443	0.009610	-0.370224

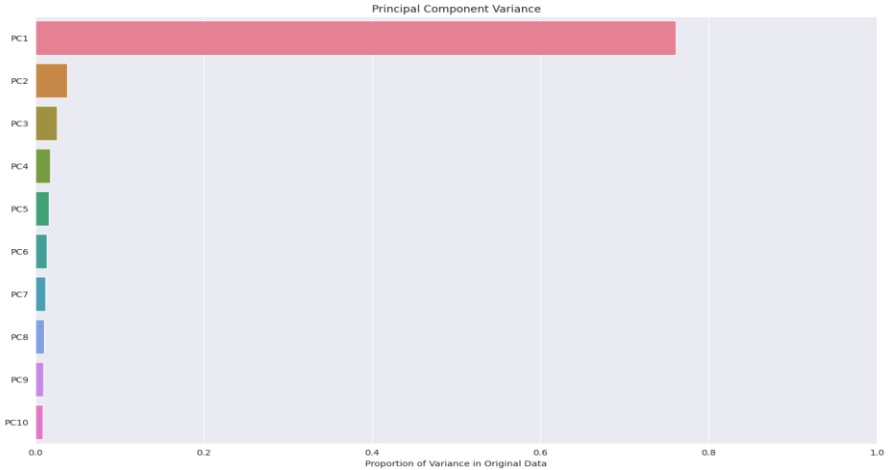
The professional home for the engineering and technology community worldwide

Data Preprocessing

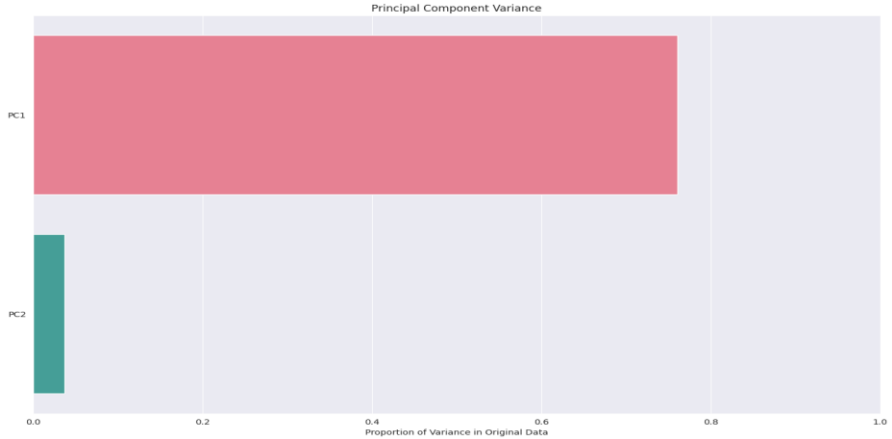
Data Preprocessing Using PCA

- This is horizontal bar plot to visualize the proportion of variance explained by each principal component obtained through Principal Component Analysis (PCA).
- These visuals clearly shows the contribution of each principal component to the overall variance in the original data, aiding in understanding the effectiveness of PCA in capturing essential information.
- Provide insights into the significance of each component in capturing the variability within the original dataset. The larger the proportion, the more important the corresponding principal component is in representing the data.

For 10 Components



For 2 Components



The professional home for the engineering and technology community worldwide

Models Used

Logistic Regression

Definition

- - Used for binary classification in machine learning.
- - It estimates the probability of an instance belonging to a class.
- - Uses the logistic function to transform input features.
- - Produces a probability score between 0 and 1.

How it's done?

- - Adjusts weights to maximize the likelihood of observed class labels.
- - Produces likelihood scores and decides a cutoff for making classifications.
- - Simple and interpretable but assumes linearity in relationships.
- - Essential for predicting binary outcomes with probabilities in machine learning.

IEEE
The professional home for the engineering and technology community worldwide

Models Used

Decision Trees

Definition

- Applicable to both classification and regression tasks.
- Construct a tree-like structure based on feature conditions.
- Good at handling non-linear relationships in complex datasets.

How it's done?

- - Choosing important features to split the data into subsets at each step.
- - Process is recursive, forming a tree structure that shows decision-making steps.
- - They are interpretable, making it easy to understand the decision path.
- - Susceptible to overfitting, capturing noise in the data.
- - Techniques like pruning and Random Forests are used to improve generalization.
- - Known for flexibility, accommodating various types of data.

IEEE
The professional home for the engineering and technology community worldwide

Why Logistic Regression and Decision Tree?

Logistic Regression

➢ **Binary Classification:**

- Used for binary classification problems. In this project, the target variable is binary (divorce or not), making Logistic Regression a natural choice.

➢ **Interpretability:**

- Allowing one to understand the impact of each feature on the likelihood of divorce.

Decision Tree

➢ **Interpretability and Visualization:**

- Easy to interpret and visualize.
- Represent a series of decisions based on features, leading to a final prediction.
- useful in understanding the decision-making process of the model.

➢ **Handling Nonlinear Relationships:**

- Captures nonlinear relationships between features and the target variable.
- Good for complex and nonlinear datasets.

➢ **Feature Importance:**

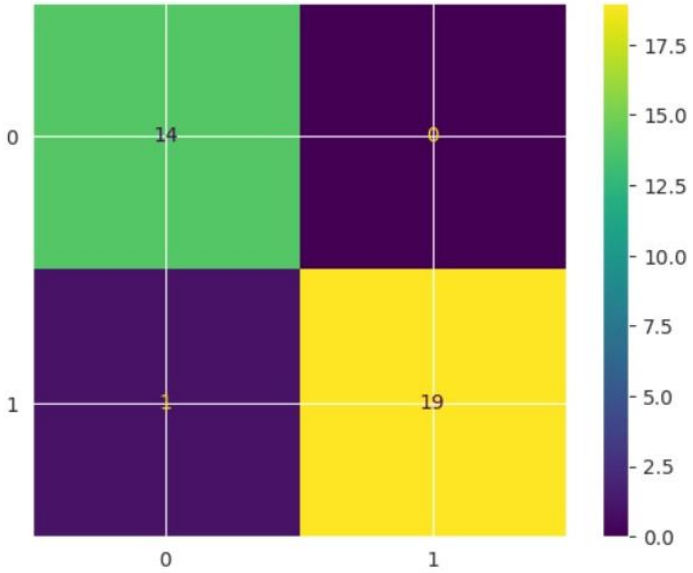
- Valuable in understanding which features contribute most to the decision-making process.

The professional home for the engineering and technology community worldwide

Model Performance

Confusion matrix - Logistic Regression Model

Accuracy : 96.15%

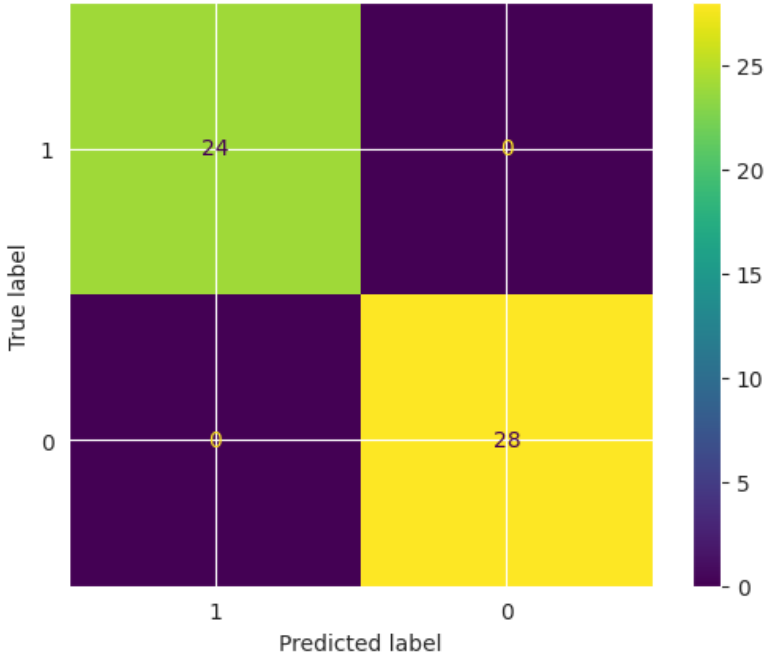


The professional home for the engineering and technology community worldwide

Model Performance

Confusion matrix - Decision trees Model

Accuracy : 100%



The professional home for the engineering and technology community worldwide

Comparison

Logistic Regression

1. **Interpretability:**
    I. **Pros:** clear insights into which factors contribute more or less to the likelihood of divorce..
    II. **Cons:** Assumes a linear relationship between features and the log-odds, which may not capture complex patterns.
2. **Statistical Significance:**
    I. **Pros:** gives us predicted values that help assess how important each factor is in predicting divorce. It's like a guide to figuring out which aspects of a relationship really matter.
    II. **Cons:** Assumes a linear relationship, and may not perform well if the relationship is highly nonlinear.
3. **Stability and Robustness:**
    I. **Pros:** Logistic Regression tends to be more stable and less prone to overfitting compared to Decision Trees, especially with smaller datasets.
    II. **Cons:** May not capture complex relationships as effectively as Decision Trees.
4. **Prediction Probability:**
    I. **Pros:** Gives us probabilities, helping us understand predictions in a more detailed way. It's like getting a closer look at the chances of different outcomes.
    II. **Cons:** The linearity assumption may limit its ability to capture certain patterns.

Decision Tree

1. **Interpretability:**
    I. **Pros:** Decision Trees provide a clear and interpretable structure, representing a series of decisions based on features. This can help in understanding the factors leading to divorce.
    II. **Cons:** Deep decision trees can become complex and may overfit to noise in the training data, making interpretation challenging.
2. **Handling Nonlinear Relationships:**
    I. **Pros:** Decision Trees can capture nonlinear relationships well, which is beneficial if the relationship between certain features and divorce risk is not linear.
    II. **Cons:** Prone to overfitting, especially with complex relationships, leading to poor generalization on unseen data.
3. **Feature Importance:**
    I. **Pros:** Decision Trees provide a feature importance score, indicating which features are crucial in making divorce predictions. This can offer insights into the most influential factors.
    II. **Cons:** Feature importance may be biased towards features with more categories or levels.
4. **Visualization:**
    I. **Pros:** The tree structure can be visualized, providing an intuitive representation of decision-making.
    II. **Cons:** Visualizations can become cumbersome with deep trees, making it challenging to interpret.

IEEE
The professional home for the engineering and technology community worldwide

Conclusion

Considerations for the Divorce Prediction Project:

- **Nature of Data:**
  - If the relationship between features and divorce risk is complex and nonlinear, Decision Trees might capture it better. If relationships are approximately linear, Logistic Regression could be sufficient.(but it is not)
- **Interpretability:**
  - If clear interpretability is a priority, Decision Trees might be preferable for visual representation. Logistic Regression is also interpretable but assumes a linear relationship.
- **Overfitting Concerns:**
  - If the dataset is small or prone to noise, Logistic Regression may be more robust and less prone to overfitting compared to Decision Trees.
- **Exploration of Complex Patterns:**
  - If the project aims to explore complex and nonlinear patterns in the data, Decision Trees could be valuable. Ensemble methods like Random Forests may be considered for improved generalization.

Considering the interpretability, stability, and potential linearity of relationships, Logistic Regression could be a good starting point for the project.

The professional home for the engineering and technology community worldwide

Conclusion

Comparison Between Logistic regression and Decision Trees

According to above results here is the comparison:

	Logistic regression	Decision Trees
Accuracy	96.15%	100%
Computational Time	Lower time	Bit higher

So in the context of this data set,

Logistic regression model outperforms Decision Trees model

IEEE
The professional home for the engineering and technology community worldwide

Conclusion

Considerations for the Divorce Prediction Project:

- **Overfitting:**

Decision trees are prone to overfitting.

- **Small or Homogeneous Dataset:**

For small dataset if it contains instances that are too similar, a decision tree can easily memorize the training data, leading to high accuracy.

Considering the interpretability, stability, and potential linearity of relationships, Logistic Regression could be a good starting point for the project.

IEEE
The professional home for the engineering and technology community worldwide

Ethics

1. Privacy and Data protection:

Protect the privacy of individuals in your dataset. Avoid using sensitive information without explicit consent, and ensure that the data is anonymized to prevent identification of individuals.

2. Transparency:

Make the project transparent by clearly explaining how it works. Users and stakeholders should understand the decisions the model makes and how it impacts them.

3. Security:

Protect your model and data from unauthorized access. Implement security measures to prevent potential misuse or attacks.

IEEE
The professional home for the engineering and technology community worldwide

Ethics

4. Proportionality Do not harm:

When developing ML systems, developers should consider what harm the application could do to individuals, groups and society, especially to vulnerable people. They should seek to eliminate or minimize those harms.

5. Bias Awareness:

Be aware of and actively address biases in your data. Biases can lead to unfair predictions, so it's crucial to identify and mitigate them throughout the development process.

IEEE
The professional home for the engineering and technology community worldwide

References

- Kaggle - https://www.kaggle.com/datasets/andrewmvd/divorce-prediction

- Sckit Learn Documentation
  - https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html
  - https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html

- Google colab for model training
  - https://colab.google

IEEE
The professional home for the engineering and technology community worldwide