# Analysis of Happy DB for Happy Moment Prediction

Kedhara Nethra - 011833640

Mohammed Haroon - 011815479

Lavanya Kadukuri - 011833588

# Introduction

**HAPPY DB:**

- Happy DB is a collection of 100,000+ happy experiences
- Information is gathered from Amazon's Mechanical Turk over a period of four months.

**MOTIVATION:**

- It is necessary to refocus on the everyday things that make us happy. This is what we are trying to do with Happy DB.

**OBJECTIVE:**

- To find the correlation between demographic information of a person and the category of happiness by using classification algorithms and ultimately build a predictive model based on the correlation determined.

# Dataset Description:

- Happy DB is a collection of various files of which we are using only cleaned_hm.csv and demographic.csv.
- cleaned_hm.csv has dimensions 100535*9 and demographic.csv has dimensions 10844*6. The feature that we are considering for our analysis in cleaned_hm.csv is “predicted_category”. We are considering features age, country, parenthood, marital, gender from demographic.csv.
- cleaned_hm.csv contains cleaned happy moments and demographic.csv contains demographic details of a person.

# Solutions

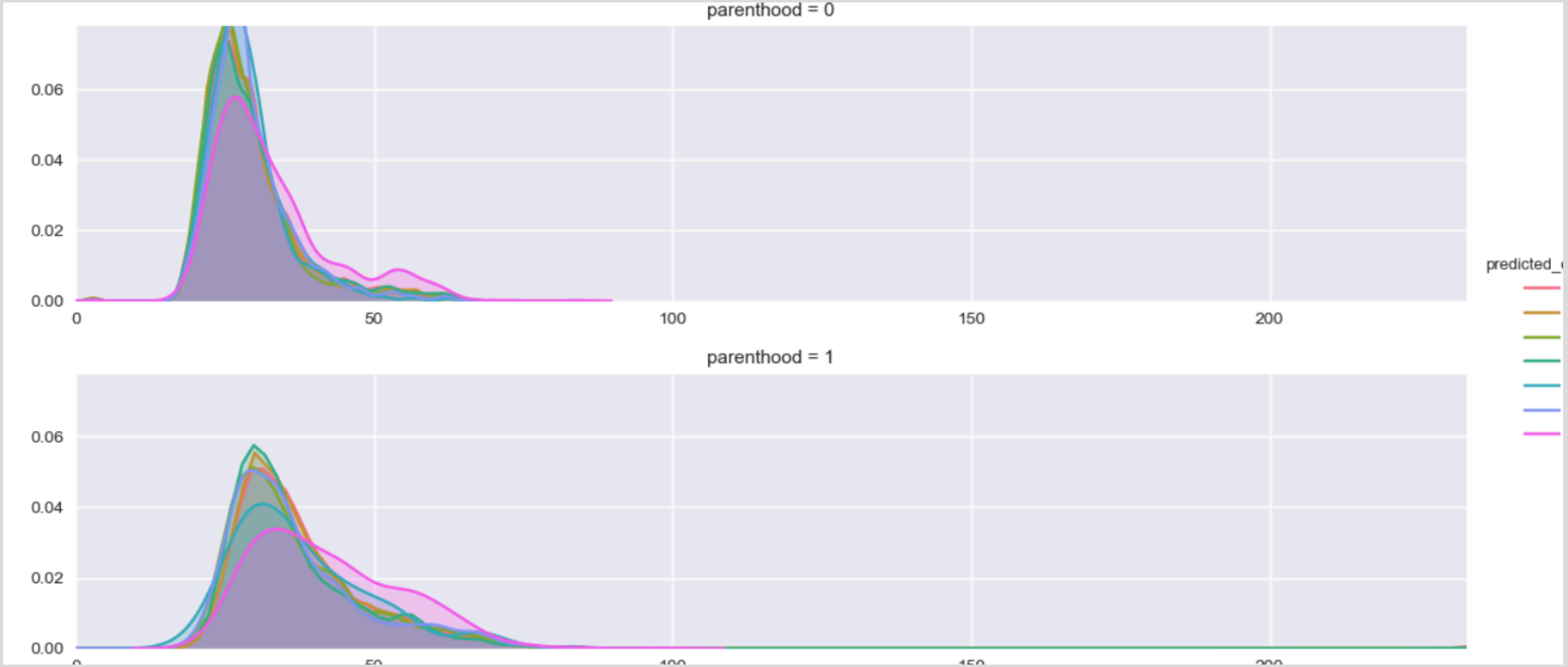
Algorithms used:

- Gaussian Naïve Bayes,
- Decision Trees Classifier,
- Random Forest Classifier,
- Support Vector Machine,
- K-Nearest Neighbor,
- Logistic Regression

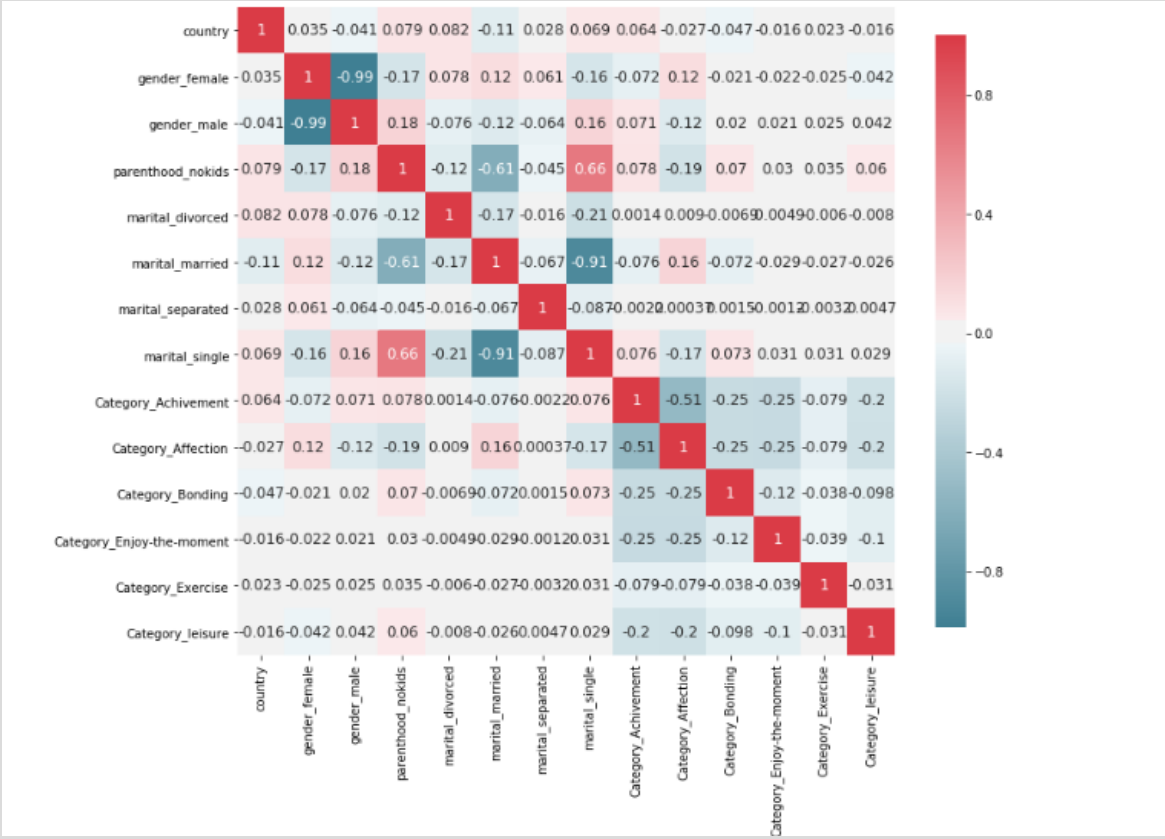
# Data preprocessing

- Dropping features that are not required
- Imputing missing values
- Removing outliers
- Converting categorical to dummy variables
- Preventing dummy variable trap
- Replace NaN with mean for age
- Replace NaN with mode for other categorical values

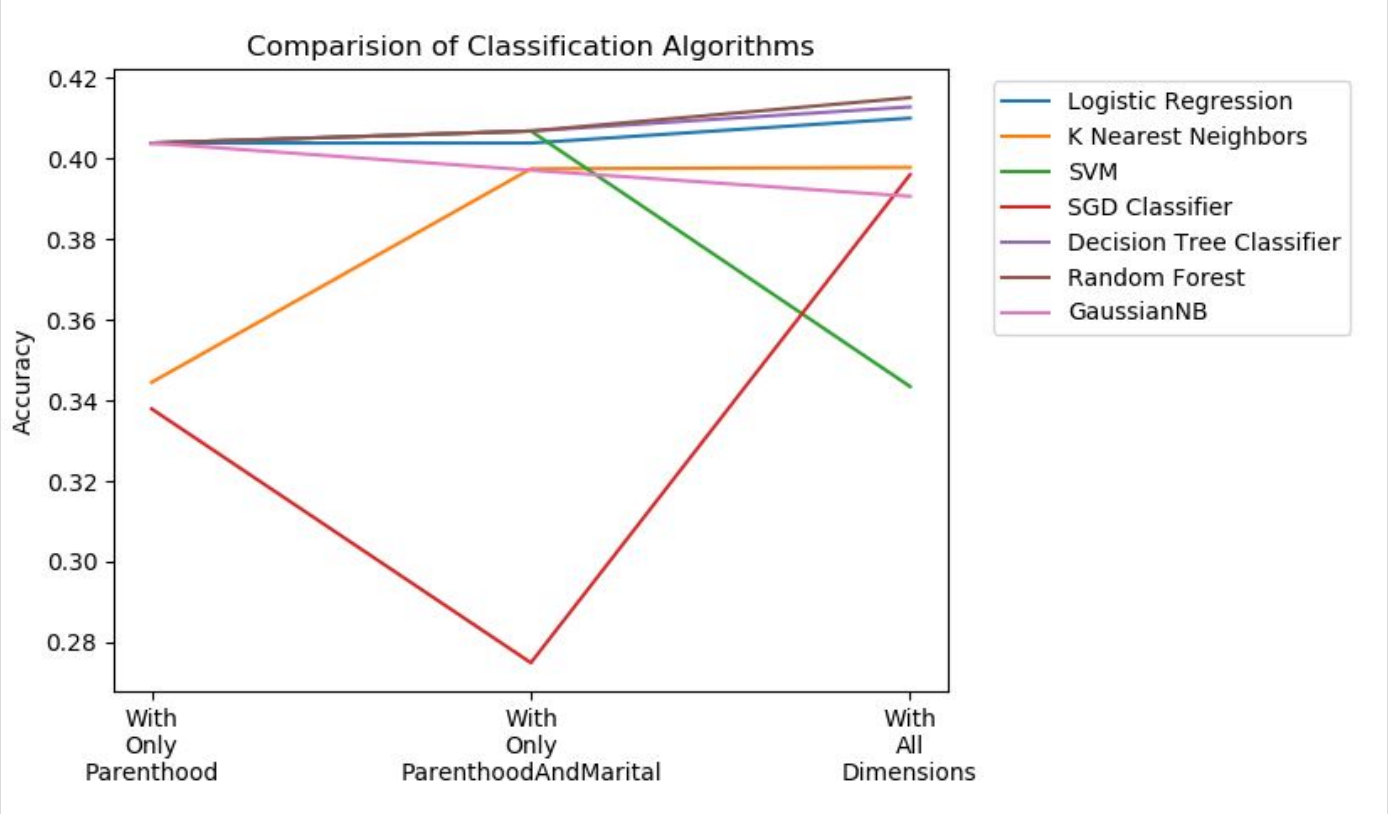
# Data Preprocessing



# Data Preprocessing



# Algorithm Variations:



# Evaluation/Demo

- Accuracy of Knn with Parenthood = 0.34445
- Accuracy of Knn with all features = 0.39775
- Accuracy of Linear Regression with Parenthood = 0.40378
- Accuracy of Linear Regression with all features = 0.40996
- Accuracy of Naïve bayes with Parenthood = 0.40378
- Accuracy of Naïve bayes with all features = 0.39085
- Accuracy of Random Forest with Parenthood = 0.40378
- Accuracy of Random Forest with all features = 0.41504
- Accuracy of SGD with Parenthood = 0.33788
- Accuracy of SGD with all features = 0.39596
- Accuracy of SVM with Parenthood = 0.40378
- Accuracy of SVM with all features = 0.34341
- Accuracy of Decision Tree with Parenthood = 0.403785Accuracy of Decision Tree with all features = 0.412752

# Things that worked:

- Data preprocessing gave us much better results than the raw data. All the measures like data cleaning by removing unwanted features, outliers and imputing missing values improved the accuracy.
- Our decision to convert categorical to numerical values also helped in giving better results.

# Things that didn't work:

- Linear algorithms like SGD Classifier and Logistic Regression didn’t work well in giving satisfactory results as anticipated. Logistic Regression only could predict two values out of a total of six values in predicted_category.
- Initially we thought the features like age and country will contribute towards target variable “predicted_category” which wasn’t the case. Happy moment category is not dependent on demographic data is what we found out.

# Analysis of Results:

- With the help of results and the evaluation graphs we can say that Logistic Regression has good accuracy results but the algorithm failed to predict most of the target categories.
- Tree based algorithms like Decision Tree and Random Forest performs the best in all three conditions, i.e with only Parenthood variable, Parenthood along with Marital and also with all the variables.
- It is also seen that Parenthood variable alone was able to give an accuracy of more than 40% for most of the algorithms except KNN and SGD.
- We can see that adding more dimensions does not improve our accuracy score. With 8 more variables as independent variables, we were able to improve the accuracy by only 1%

# Conclusion:

- Through data exploration and preprocessing of Happy DB, we better understood the data we are dealing with and how features are interdependent. We could apply concepts learnt in the class which was a very good experience.
- Even after applying various algorithms, we couldn’t predict the happy moment category with a good accuracy which shows that this category is independent of demographic information and this is counter-intuitive!
- Future scope is to perform sentiment analysis to determine happiness by specific types of experiences.

# References:


- Akari Asai, Sara Evensen, Behzad Golshan, Alon Halevy, Vivian Li, Andrei Lopatenko, Daniela Stepanov, Yoshihiko Suhara, Wang-Chiew Tan, Yinzhan Xu, ``HappyDB: A Corpus of 100,000 Crowdsourced Happy Moments'', LREC '18, May 2018.
- https://github.com/rit-public/HappyDB
- https://www.kaggle.com/ritresearch/happydb
- http://scikit-learn.org/stable/supervised_learning.html
- http://scikit-learn.org/stable/auto_examples/model_selection/plot_roc.html#multiclass-settings
- https://pandas.pydata.org/pandas-docs/stable/visualization.html

# THANK YOU!