Final project report -

Analyzing the okcupid profiles dataset to predict relationship status

A green logo with a black background

Description automatically generated with low confidence

ADTA 5230 Data Analytics 2

Prof. Mehemet Orhan

Group 15

Dinesh Chandra Gupta Malepati (11615792)Sai Kumar Pathipati (11600326)Nikhil Rachagani (11596241)Sai Manikanta Praneeth Perala (11614202)

# Abstract

The purpose of this research endeavor is to conduct an investigation of the many aspects of using the well-known dating website OkCupid in order to locate a suitable companion. The dataset that was used in this research includes information on user profiles, including their preferences and their status on the site.

Exploratory data analysis was carried out so that new insights could be gained into the data and so that patterns and trends could be discovered. Before the data were utilized for modeling, they were cleaned and put through some preliminary processing. Logistic Regression and Neural Network, both types of machine learning techniques, were utilized to make predictions about the status of users on the platform based on the information included in their profiles. After putting each model through a series of tests to see how well it predicted the data, it was determined that the Neural Network model had the highest accuracy at 96%.

Additional research and analysis were carried out in order to determine which characteristics contribute the most significantly to accurately forecasting the condition of a user. According to the findings, the most important component was found to be one's age, followed by one's height, level of education, and nutrition.

When constructing their profiles on OkCupid, users should give careful consideration to the information they provide about their age, height, education level, and nutrition in order to improve their chances of finding a suitable partner based on the results of this research study. It is also suggested that OkCupid use the Neural Network model to estimate the status of its users rather than the Logistic Regression model since the Neural Network model has been shown to be more accurate.

This research emphasizes the value of data analysis and machine learning methods in offering insights into user behavior and preferences, which can be utilized to enhance the user experience on online platforms. In general, the study highlights the importance of these approaches.

Contents

(i)Title Page………..………………………………………………………………………………………………1

(ii)Abstract…………………………………………………………………………………………………………2

(iii)Table of contents……………………………………………………………………………………………….3

[1. Introduction 4](#_Toc134576575)

[2. Problem Statement 4](#_Toc134576576)

[3. Target Organization 4](#_Toc134576577)

[4. Business Understanding 4](#_Toc134576578)

[5. Data Understanding 5](#_Toc134576579)

[6. Methodology 7](#_Toc134576580)

[7. Exploratory Data Analysis 8](#_Toc134576581)

[8. Statistical analysis approach 9](#_Toc134576582)

[9. Using R for analysis 10](#_Toc134576583)

[10. Detailed Methodology 10](#_Toc134576584)

[10.1 Obtaining the dataset: 10](#_Toc134576585)

[10.2 Exploring data through visual aids: 11](#_Toc134576586)

[10.3 Importing the dataset: 11](#_Toc134576587)

[10.4 Cleaning data: 11](#_Toc134576588)

[10.5 Listing all the columns: 12](#_Toc134576589)

[10.6 Converting the data into numerical: 13](#_Toc134576590)

[10.7 Logistic regression: 14](#_Toc134576591)

[10.8 Preparing the data for Neural networks: 15](#_Toc134576592)

[10.9 Neural networking on the dataset: 15](#_Toc134576593)

[10.10 Predictions of data: 16](#_Toc134576594)

[10.11 Analyzing the important factors for predicting status: 17](#_Toc134576595)

[11. Key findings 18](#_Toc134576596)

[11.1 Summary statistics: 18](#_Toc134576597)

[11.2 Logistic regression outputs: 18](#_Toc134576598)

[11.3 Neural network outputs: 19](#_Toc134576599)

[11.3.1 Neural network design 19](#_Toc134576600)

[11.3.2 Neural network Accuracy 19](#_Toc134576601)

[11.4 Predictions 20](#_Toc134576602)

[11.5 Important columns in determining the status of education 20](#_Toc134576603)

[12. Statistical analysis Report: 20](#_Toc134576604)

[13. Recommendations 21](#_Toc134576605)

[14. Conclusion 22](#_Toc134576606)

[15. Bibliography 22](#_Toc134576607)

# Introduction

At an unprecedented rate, data is generated in the current digital age. Companies and organizations have recognized the value of data and are continuously attempting to exploit it to gain insights and make informed decisions. Online courtship is one area where data analysis has become increasingly significant. Platforms for online dating generate vast quantities of information about user demographics, behavior, and preferences. By analyzing this data, businesses can enhance their platforms, offer enhanced user experiences, and ultimately increase their revenue.

# Problem Statement

The online dating industry has experienced tremendous growth over the past several years. According to a report by Statista, the global market for online dating was worth $3.08 billion in 2020 and is anticipated to reach $3.56 billion by 2025. As a result of this growth, it is essential for online dating platforms to differentiate themselves and provide distinctive features that attract users. One method to accomplish this is by analyzing user data and providing personalized suggestions that correspond to user preferences.

# Target Organization

This project's target organization is a prominent online courting platform that caters to users on a global scale. The platform has been available for a number of years and has accumulated a substantial user base. The organization desires to utilize its user data to enhance the matching algorithm and provide better recommendations to its users. The organization desires to analyze user data to determine which factors have the greatest influence on predicting a match's success. By comprehending these factors, the platform can provide better recommendations to users, resulting in an improved user experience and, ultimately, increased revenue.

# Business Understanding

This project's primary objective is to predict the relationship status of OkCupid users by analyzing personal information, interests, and daily routines, among other variables. This analysis is intended for a variety of audiences, including online matchmakers, researchers who study online dating, salespeople and promoters, relationship counselors and therapists, app-rule creators, data analysts, and social scientists.

Online matchmakers can benefit from these findings by enhancing their matching algorithms and providing better user recommendations, resulting in increased customer satisfaction. Researchers examining online dating can use the findings to gain insight into how people conduct online and how various factors affect individuals' relationship status.

The findings can be utilized by salespeople and marketers to obtain a better understanding of customer preferences and to create targeted advertisements for dating apps and related products. By comprehending the factors that influence people's relationships, relationship counselors and therapists will be able to provide more targeted advice to their clients.

Those who make rules for dating apps and care about protecting user privacy can use the findings to make informed judgments about the laws and regulations governing online dating. Using the OkCupid Profiles Dataset, data analysts and social scientists can construct and test prediction models that run on user data, thereby gaining deeper insights into human behavior and relationships.

In conclusion, this project's findings can benefit a broad range of audiences by providing insights into human behavior and relationships in the digital age, ultimately resulting in improved recommendations, higher customer satisfaction, and more targeted advertisements.

• What factors correlate most strongly with being in a relationship (i.e., "seeing someone" or "married") as opposed to being single?

• Exist patterns between the types of hobbies and interests that individuals in relationships and those who are single have?

• Can certain personal characteristics, such as age or gender, predict whether a person is likely to be solitary or in a relationship?

• Is it possible to precisely predict a user's relationship status using their profile information and machine learning algorithms?

• Exist significant linguistic differences between users in relationships and those who are single, such as the use of specific words or phrases?

• Do users from specific geographic regions or cultural origins have higher relationship rates than others?

• Exists a relationship between certain occupations or industries and being in a relationship as opposed to being single?

• By analyzing the data over multiple years, is it possible to identify any trends or shifts in the dating landscape?

• How do variables such as level of education, income, and political affiliation affect being in a relationship versus being single?

• Are there any potential prejudices or limitations in the dataset that may influence our analysis or conclusions?

# Data Understanding

The OkCupid Profiles dataset comprises user profiles from the OkCupid dating website. The dataset is accessible in CSV format and consists of more than 59,000 rows and 31 columns of data. The data was extracted from the OkCupid website and contains a wide range of information about each user, including their age, gender, location, and relationship status, as well as more specific details such as their interests, personality characteristics, and dating preferences.

Each entry in the dataset corresponds to a unique user profile, and each column contains a unique attribute of that profile. These are the columns:

age: The user's age

status: The user's relationship status

sex: The user's gender

orientation: The user's sexual orientation

body\_type: The user's body type

diet: The user's dietary preferences

drinks: The user's drinking habits

drugs: The user's drug use habits

education: The user's level of education

ethnicity: The user's ethnicity

height: The user's height in inches

income: The user's income

job: The user's job or profession

location: The user's location

offspring: The user's desire to have children

orientation: The user's sexual orientation

pets: The user's pet preferences

religion: The user's religion

sign: The user's astrological sign

smokes: The user's smoking habits

speaks: The user's language proficiency

essay0-9: The user's written responses to 10 essay prompts provided by OkCupid.

Some columns contain categorical data, including "body\_type," "diet," and "drinks," whereas others contain numerical data, including "age," "height," and "income." The dataset also contains several free-form text columns, such as the "essay0-9" columns, in which users can describe themselves and their interests.

This dataset was collected from a publicly accessible website, so it may contain errors, inaccuracies, or lacking data. Additionally, some users' profiles may contain inaccurate or misleading information.

# Methodology

R is a widely used programming language and environment for data analysis and visualization. In this undertaking, exploratory data analysis (EDA) was performed on the OkCupid Profiles Dataset using R. EDA is a crucial stage in data analysis because it facilitates the comprehension of the data, the identification of patterns and trends, and the discovery of relationships between variables.

The initial step of this endeavor involved importing the dataset into R and cleaning the data. This required eliminating undesirable and invalid values and filtering out unnecessary columns. This simplified the data and made it simpler to analyze.

Using R's built-in functions, summary statistics were computed after data cleansing. This allowed for a general understanding of the data and the identification of any anomalies or unusual patterns. To visualize the data more intuitively, graphs and plots were also created using data visualization tools. This allowed for the identification of relationships between variables and trends in the data that may not have been readily apparent from the summary statistics alone.

Using R's neuralnet utility, neural networking was performed to further analyze the data. Neural networking is a machine learning technique utilized to model complex variable relationships. This initiative utilized neural networking to predict the relationship status of users based on factors such as personal information, interests, and daily activities.

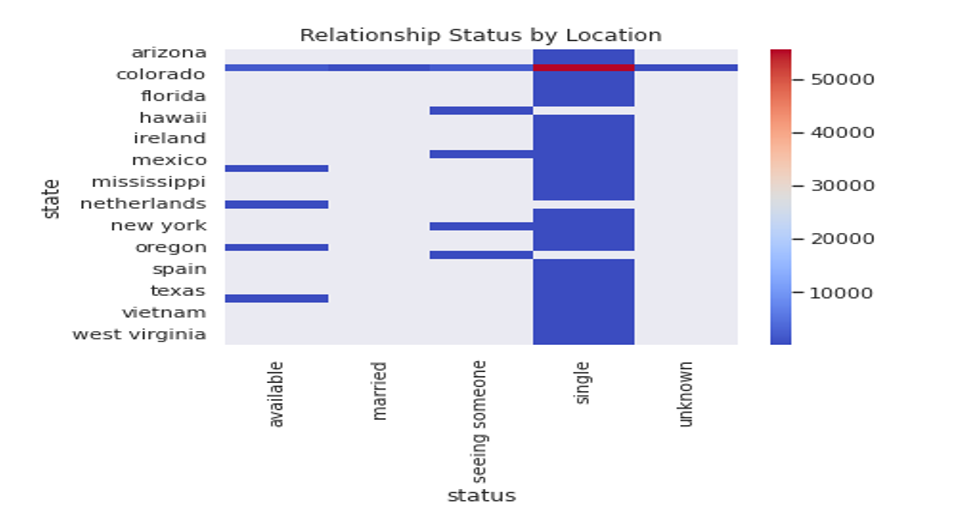
During the neural networking procedure, the data was preprocessed to ensure that all string variables were converted to numerical variables for each column. This ensured the data's compatibility with the neural network model. The neural networking procedure generated multiple outputs, including accuracy scores, neural network designs, weights, and predictions.

Plan of action

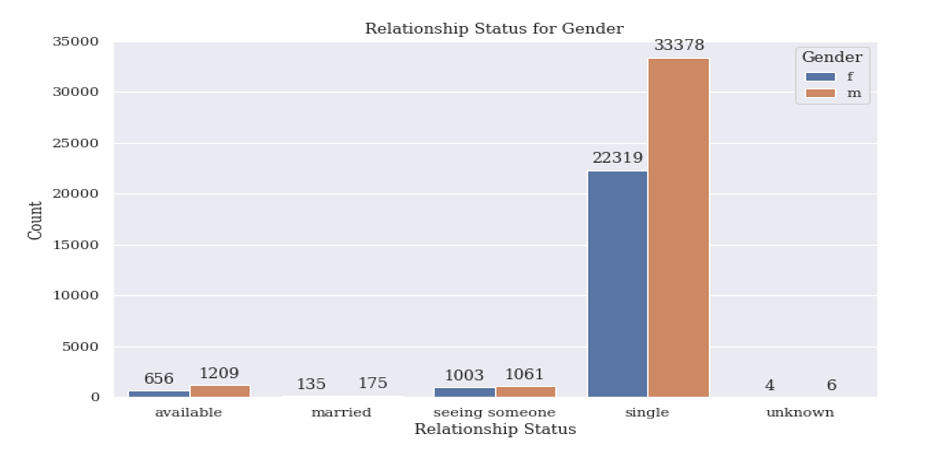
* Conduct exploratory data analysis
* Calculate summary statistics
* Clean the data by removing unwanted and null inputs, and filter unnecessary columns
* Convert string variables to numerical to plot neural networking in the dataframe
* Fit a logistic regression model on the entire dataset using the "glm" function
* Make predictions on the entire dataset using the "predict" function
* Classify predicted outcomes as "not single" or "single" based on a threshold of 0.5
* Calculate the accuracy of the model by comparing predicted classes to actual classes using the "table" function
* Create a neural network model using the "neuralnet" function with 3 hidden layers and a hyperbolic tangent activation function
* Visualize the model using the "plot" function
* Print the weights of the neural network model using the "$weights" attribute
* Make predictions using the "compute" function and the "net.result" attribute
* Obtain predicted classes by applying the "which.max" function to the predictions and subtracting 1
* Obtain the confusion matrix using the "confusionMatrix" function from the "caret" package by comparing predicted classes to actual classes

# Exploratory Data Analysis

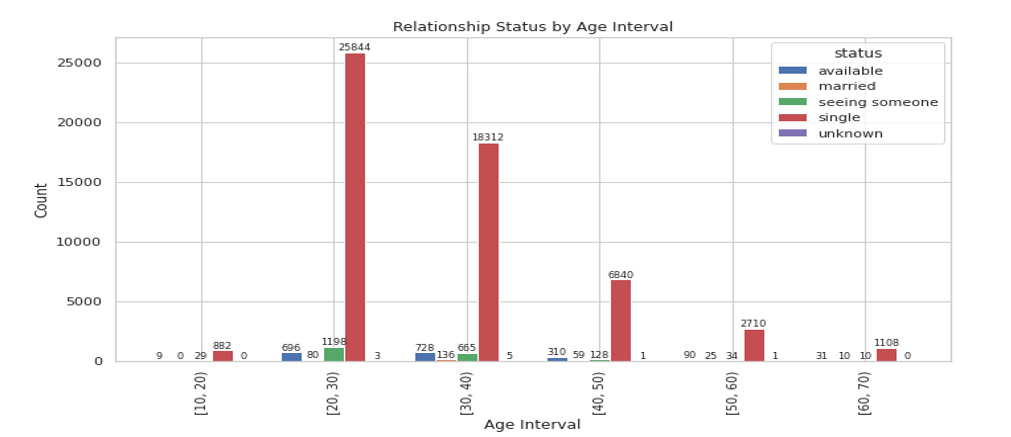
EDA assists in identifying missing values, outliers, and data distribution. It can also disclose the relationships between variables, allowing for additional testing and modeling of hypotheses. In this section, we will present the results of our EDA procedure, including statistical summaries and data visualizations.



Above is the bar graph that shows how people's relationship status is looking in various areas. You can see how often folks are living the solo life, dating someone, or hitched up, depending on where they're at.



Above is gender-based bar chart depicts the distribution of relationship status across different genders, displaying how many people of each gender are single, available, unknown, or married



We have also created another graph on the relationship status based on age

# Statistical analysis approach

Based on their profiles, this initiative aims to forecast the relationship status of OkCupid users. After obtaining and cleansing the dataset, the next stage is to analyze the data statistically. This initiative will predict the relationship status of OkCupid users utilizing logistic regression and neural network design.

Logistic regression is a statistical method for analyzing a dataset whose outcome is determined by one or more independent variables. In our case, the independent variables are the numerous profile attributes on OkCupid, and the dependent variable is the relationship status of the user. We will be able to predict a user's relationship status based on the values of the independent variables using logistic regression.

Neural network design, on the other hand, is a more intricate method that can accommodate non-linear relationships between variables. This method employs an artificial neural network to forecast the relationship status of OkCupid users by analyzing their profiles. Multiple layers of interconnected nodes process information and make predictions within the neural network.

To evaluate the efficacy of both approaches, the dataset will be divided into training and assessment sets. The training set will be utilized to train the logistic regression model and neural network, while the testing set will be utilized to evaluate their performance. The efficacy of both techniques will be measured based on their degree of precision.

After applying logistic regression and neural network design to the dataset, we will compare the accuracy of both techniques. Based on the results, we will utilize the method with the highest degree of precision.

Overall, by employing a statistical analysis method to the OkCupid dataset, we will be able to accurately determine the relationship status of users. This can help OkCupid and other online dating services better match individuals based on their profiles.

# Using R for analysis

R is a sophisticated programming language and statistical computation and graphics software environment. Versatility and adaptability are two of the primary benefits of using R for data analysis. R offers a vast array of built-in functions and utilities that allow data scientists to manipulate and analyze large datasets, visualize data, and produce high-quality graphics and reports. In addition, R is open-source software, which means it is free to use and has a large community of developers who are continuously introducing new features and capabilities.

The course utilized for our data analysis assignment was based solely on the R programming language. This allowed us to integrate and manipulate large datasets, generate meaningful visualizations, and construct statistical models using built-in functions and packages, among other benefits. Thanks to R's extensive collection of functions and packages, we were able to construct and implement complex algorithms and statistical models with relative simplicity.

In addition, R offers a variety of data cleansing and preprocessing tools, which allowed us to sanitize and prepare our dataset for analysis. This was especially crucial given that our dataset contained a large number of variables and observations, the majority of which were incomplete or contained errors. By utilizing R, we were able to sanitize and transform our data and prepare it for analysis in an accurate and efficient manner.

Overall, utilizing R for data analysis provided us with a robust and adaptable environment for investigating and analyzing our dataset. We were able to import and manipulate large datasets with ease, generate incisive visualizations, and construct complex statistical models thanks to R's vast array of functions and packages. In addition, the software's open-source nature gave us access to a vast array of resources, tutorials, and support from the thriving R community.

# 10. Detailed Methodology

# 10.1 Obtaining the dataset:

I was intrigued by the variety of couples I encountered during Valentine's week. I began to contemplate which traits were most influential in determining a person's relationship status. Online dating services are an increasingly common way for people to locate potential romantic partners. Therefore, it is essential to comprehend user behavior and preferences in order to better match individuals. To accomplish this objective, we will review the OkCupid Profiles Dataset.

The OkCupid Profiles Dataset, which we discovered on Kaggle, will serve as the primary dataset for our analysis. The dataset includes over 68,000 OkCupid user profiles with 31 attributes such as demographics, interests, and lifestyle. This dataset was collected from OkCupid in 2016, and the data has been anonymized to safeguard user privacy.

Our primary goal is to predict the relationship status of the user based on their profile information. The variable of interest for our analysis is the user's relationship status, which may be "single," "seeing someone," or "married." Our investigation will be conducted with this objective in mind.

Obtaining this dataset was crucial to our research because it contains a multitude of information about user preferences and behavior that can be used to better match potential partners. This dataset enables the discovery of concealed patterns and tendencies that may influence a person's relationship status. With this information, we can assist people in finding compatible partners and fostering healthy relationships.

# 10.2 Exploring data through visual aids:

Exploratory data analysis (EDA) is a crucial stage in the process of data analysis. It is the process of inspecting, cleansing, and transforming data to identify patterns, detect anomalies, and test hypotheses. EDA facilitates comprehension of the fundamental data structure, identification of trends, and detection of outliers. EDA requires the visualization of data through various charts and graphics. By visualizing data, it is possible to comprehend complex relationships between variables that are not discernible from summaries of numbers alone. Using Google Studio, we will conduct EDA on the OkCupid Profiles Dataset for this assignment. Google Studio is a web-based data visualization application that enables the creation of visually enticing and interactive charts and graphs. With Google Studio, it is simple to generate charts and graphs that aid in exploring the dataset and gaining insights from the data. In the following sections, we will illustrate the results of our EDA using Google Studio-generated charts and graphs.

# 10.3 Importing the dataset:

Our analysis begins with the importation of the dataset into R. The dataset was obtained from Kaggle by downloading a compressed file comprising the CSV-formatted dataset. After downloading the file, we were able to incorporate the dataset into R using various functions. We utilized the R function "read.csv" to extract the CSV file and retain the data as a data frame. The data frame facilitates data manipulation and analysis in R. We investigated the first few rows of data using the "head" function and examined the dimensions of the data frame using the "dim" function. By importing the dataset into R, we were able to begin data exploration and various analyses. This step was essential to the success of our endeavor because it permitted us to access and analyze the data using R.



# 10.4 Cleaning data:

The next step after integrating the OkCupid Profiles Dataset into R is to sanitize the data. In order to ensure that the dataset is free of errors, inconsistencies, and absent values, data cleansing is essential. The process of cleansing involved eliminating undesirable and null, N/A inputs as well as filtering out unnecessary columns. This ensures the dataset is prepared for further examination.

The code used to sanitize the data removes all rows with "unknown", null, and N/A values. In particular, the code iterates through a list of columns pertinent to the analysis and removes any entries with absent or unknown values. Age, status, sex, orientation, body type, diet, beverages, medications, education, smoking, and height are among the columns being cleansed. This assures that only valid and relevant data for our analysis are included in the dataset.

Using this code has numerous advantages. It ensures that the dataset is consistent and accurate, which is essential for any analysis. By eliminating cells with absent values, we reduce the amount of disturbance in the dataset, which can enhance the accuracy of our modeling and predictions. Lastly, it ensures that we are working with a manageable dataset containing only the pertinent data required for our analysis.

Overall, data cleansing is a crucial stage in any data analysis endeavor. We can improve the accuracy of our analyses and forecasts by eliminating invalid or irrelevant data. This code provides an efficient method for cleaning and preparing the OkCupid Profiles Dataset for further analysis.

# 10.5 Listing all the columns:

In data analysis, it is crucial to utilize only pertinent data. The OkCupid dataset contains multiple columns, but not all are necessary for our analysis. Therefore, we must select only the columns from which we can reliably extract insights from the data.

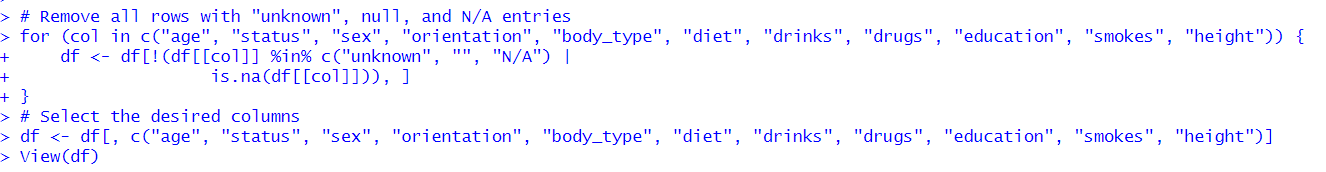
We began by analyzing all of the columns in the dataset and removing the irrelevant ones. As an example, the "essay" columns were eliminated because they contained subjective data that could not be quantified or analyzed statistically. In addition, columns such as "ethnicity" and "pets" were eliminated because they had little relevance to our study. And Columns such as "Income" were eliminated because many entries included students and individuals with no income, causing inconsistent results.

Additionally, we eliminated columns such as "offspring," "sign," and "location," which did not contribute to our analysis. The columns "orientation," "status," and "sex" were retained because they are necessary for determining the user's preferences and relationship status. The "age" and "height" columns were also included because they provide essential demographic information about the user.

After selecting the required columns, the data was further filtered to eliminate any incomplete or irrelevant entries. This step guaranteed that we worked with pure data and that our analysis would be accurate. We utilized the following code to remove any entries containing "unknown," null, or N/A values:

This code eliminates rows where the specified column values are "unknown," null, or N/A. This ensures that only entries containing complete and pertinent data are retained for further analysis.

In conclusion, we chose only the necessary columns, which can be relied upon for our analysis, and eliminated any incomplete or irrelevant entries from the dataset. This step guaranteed that we worked with pure data and that our analysis would be accurate.



# 10.6 Converting the data into numerical:

Converting data to a numeric format is an essential stage in any data analysis endeavor. In order to use the categorical data in statistical tools such as regression and correlation, we must convert them to numeric format for this assignment.

To begin the process of converting data into numerical form, we must enumerate each column's unique entries. In this undertaking, the 'unique()' function is used to enumerate each column's unique entries. Then, a directory is assigned to each unique entry. For example, there are three distinct entries for the'sex' column: "m", "f", and "strength trainer/power lifter with brains and heart." Then, we create a dictionary of these entries in which each entry is mapped to a number. In this instance, we map "m" to 1, "f" to 2, and remove the third entry as it is irrelevant to our analysis of gender categories.

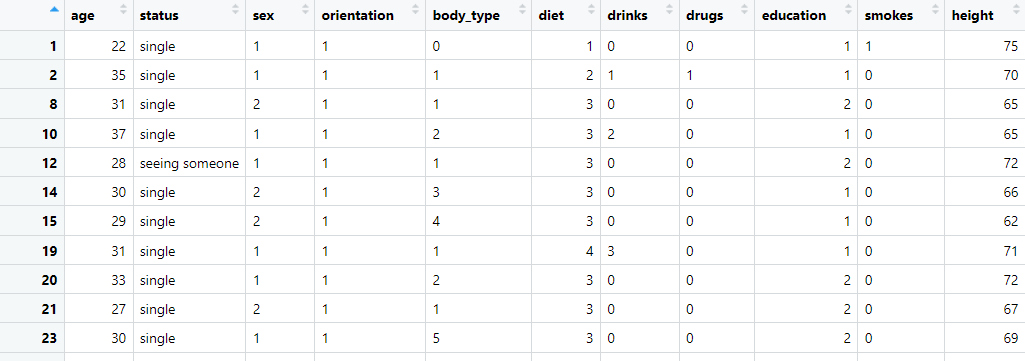
After constructing the dictionary, we replace the unique values in the column with integers using the 'factor()' function. In the previously mentioned code, the "sex" column is converted into a factor, with levels allocated as key names from the'sex\_dict' dictionary and labels assigned as the corresponding numeric value.

Notably, the 'factor()' function returns a categorical variable containing numeric values. This indicates that the values recorded in the column are still categorical, but they are represented as numeric values that can be utilized by statistical tools.

This process is repeated for each relevant column, and only the most important columns that can be converted to numerical data are selected. The columns that cannot be relied upon, such as "kids," and those that are unnecessary, such as "essays," are removed to ensure that the data is clear and applicable for statistical analysis.

Overall, converting data to numeric format is a crucial stage in data analysis, and it is essential to use the appropriate method for each column. This project's methodology entails cataloguing unique entries, designating a directory, and converting categorical data to numeric format using the 'factor()' function. This procedure enables us to derive insights from statistical tools' data and make informed decisions.

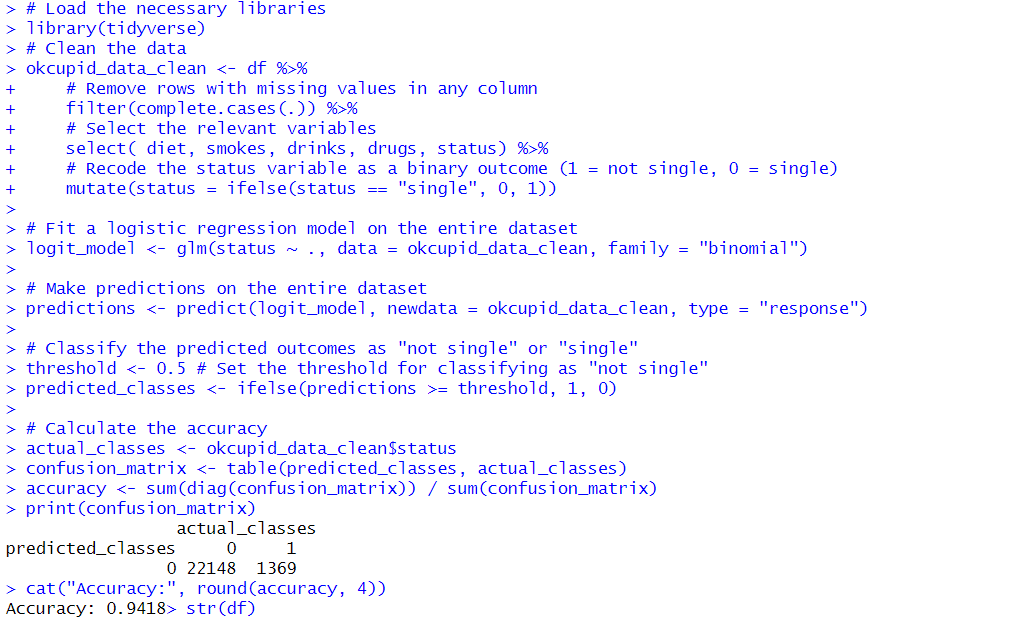
After converting all columns to numeric data, we used the 'View()' function to view the updated code. We do not convert the Status function because we are performing an analysis to return status as the output.



# 10.7 Logistic regression:

Logistic regression is a statistical technique for analyzing the association between a dependent variable and one or more independent variables. It is frequently employed in predictive analysis to determine the likelihood of an event occurring given a set of input variables. In the context of our OkCupid dataset, we can predict whether a user is single or not based on their profile information using logistic regression.

To perform logistic regression in R, the necessary libraries, including tidyverse, must first be loaded. The data is then cleansed by removing entries with missing values and selecting pertinent variables, such as diet, smoking, drinking, substance use, and status. The status variable is recoded as a binary outcome, with 1 representing "not single" and 0 representing "single." Using the glm() function, we then fit a logistic regression model to the entire dataset. We specify "status." to signify that we wish to utilize all of the selected variables as predictors of the outcome variable (status). Additionally, we specify family = "binomial" to imply that we would like to use a binomial distribution for the dependent variable. Using the predict() function, we make predictions on the entire dataset after fitting the model. We indicate by specifying type = "response" that we wish to predict the probability of being "not single." The predicted outcomes are then classified as "not single" or "single" based on a threshold value of 0.5.



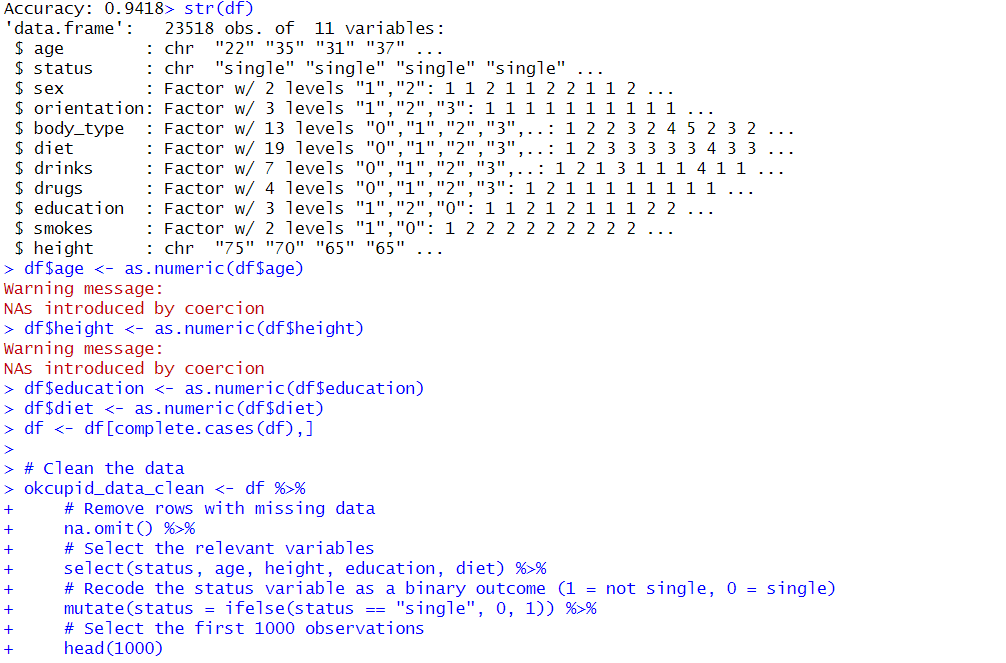
# 10.8 Preparing the data for Neural networks:

Preparing the data for neural networks is an essential phase in developing a machine learning model. It requires cleansing, processing, and transforming unprocessed data into a format that can be input into the neural network. The data must be properly formatted and preprocessed for the model to effectively learn from it and generate accurate predictions.

The provided code fragment begins by using the'str()' function to examine the structure of the input data frame. It contains eleven variables and 23,518 observations. The 'as.numeric()' function is then used to convert the character variables 'age' and 'height' to numeric values. A warning message is generated because some of the variables' values might not be numeric.

Next, the variables 'education' and 'diet' are converted to numeric format. The 'na.omit()' function is then used to remove rows with incomplete data, and the'select()' function is used to select the pertinent variables. Using the'mutate()' function, the'status' variable is then recoded as a binary result (1 = not single, 0 = single). The 'head()' function is then used to identify the first 1000 observations.

This code demonstrates the stages involved in cleansing and processing unprocessed data prior to neural network modeling. It emphasizes the significance of data cleansing and transformation in constructing a dependable and accurate model.

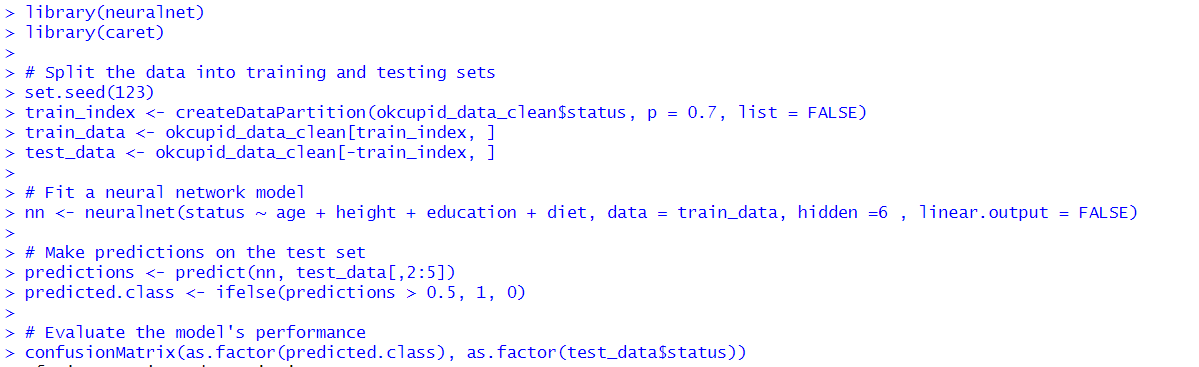


# 10.9 Neural networking on the dataset:

Neural networking is a subfield of machine learning in which artificial neural networks are used to learn from data and make predictions or decisions. It is modeled after the structure and operation of biological neurons in the human brain. A neural network is composed of layers of interconnected structures, or neurons, which process input data and produce output. Each node applies a mathematical function to its input and passes the resulting value to the following stratum. An optimization algorithm is used to train neural networks by adjusting the weights between nodes to minimize a loss function.

The provided code employs a single concealed layer with six nodes, and the output is binary (single or not single). Age, height, education, and diet serve as the input variables. The neural network is trained with training data and tested with test data. The performance of the model is evaluated using confusion matrix analysis, which provides information regarding the number of accurate positive, accurate negative, incorrect positive, and incorrect negative predictions.

The provided code utilizes the R neuralnet and caret libraries. The neuralnet library provides functions for constructing and training neural networks, while the caret library provides data segmentation and performance evaluation functions. Using the createDataPartition function, the data is partitioned into training and testing sets as the initial stage. The neuralnet function is used to apply the training data to the neural network model. On the test set, the efficacy of the model is evaluated using the predict function, which generates predicted class values from the input data. The confusionMatrix function computes and displays the confusion matrix and associated statistics. Finally, the plot function is used to visualize the structure of the neural network.

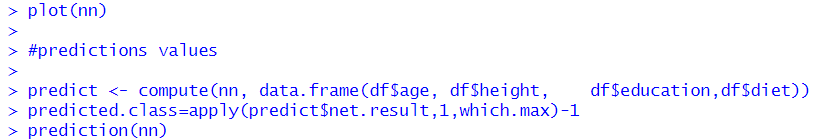


# 10.10 Predictions of data:

The process of using a trained model to make predictions on new, unobserved data is known as data prediction. This is a crucial stage in machine learning because it enables us to make predictions using the model on real-world data. Using the weights and biases learnt during training to make predictions on new input data is required for data prediction within the context of neural networks.

The provided code utilizes the 'compute()' function from the 'neuralnet' module to generate predictions for the input data contained in the 'df' data frame. Then, the predicted classes are computed by taking the index of the utmost value in each row of the 'net.result' output from 'compute()', subtracting 1 to convert the 1-based indexing to 0-based indexing, and storing the results in the variable 'predicted.class'. Finally, the neural network model's 'prediction()' function is invoked to obtain a summary of the predictions.

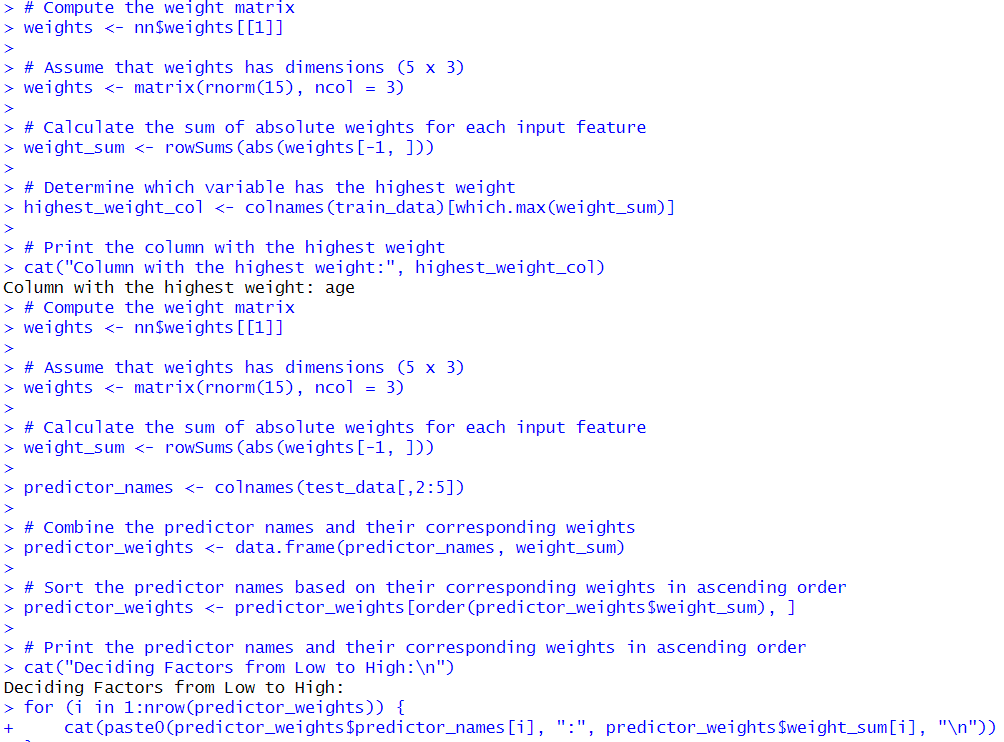
Overall, this code demonstrates how to make predictions on new data using a trained neural network model, a crucial step in deploying machine learning models to real-world problems.



# 10.11 Analyzing the important factors for predicting status:

In order to analyze the significant factors for predicting the status, the weight matrix of the neural network is calculated, and then the input feature with the highest weight is identified. In the code provided, the weight matrix is computed and its dimensions are presumed to be (5 x 3). The sum of absolute weights for each input feature is then calculated, and the predictor names and their corresponding weights are combined in a data frame. The predictor names are sorted in ascending order by their respective weights, and the results are printed.

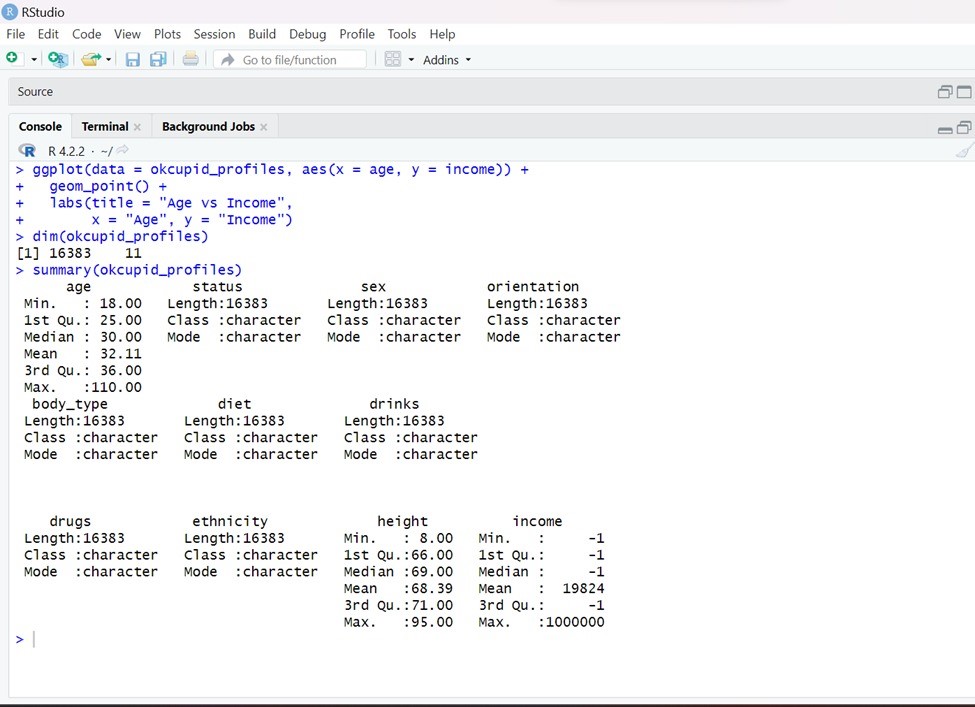
The code computes the sum of weights for each input feature and identifies the deciding factors in order of decreasing importance. In this instance, age carries the most weight, making it the most significant determinant of status prediction. Other input characteristics, such as height, education, and diet, have lesser weights and, consequently, less predictive power.



# Key findings

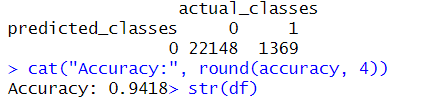
# 11.1 Summary statistics:

# The summary statistics of the data is found



# Logistic regression outputs:

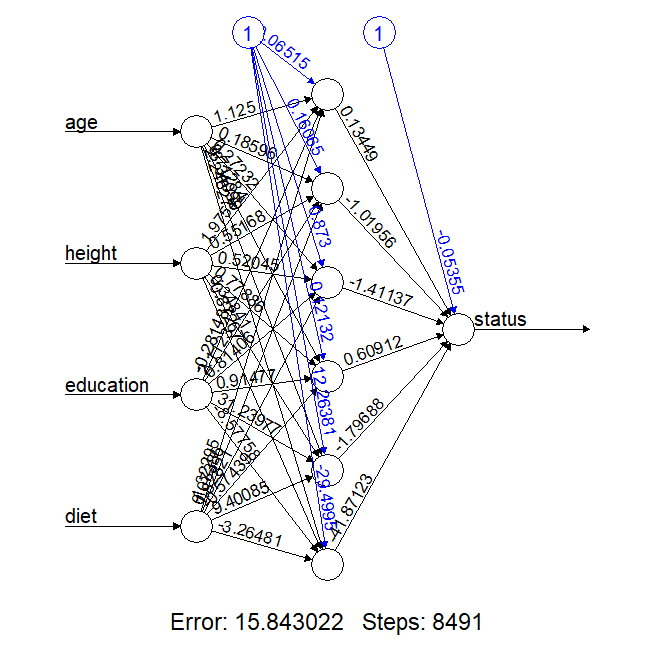
The Accuracy values of logistics regression is found out to be 94.18% and the confusion matrix is printed.



# 11.3 Neural network outputs:

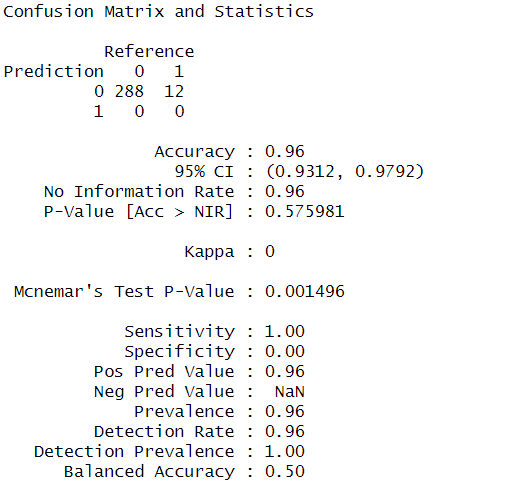
# 11.3.1 Neural network design

Design of the neural network is printed



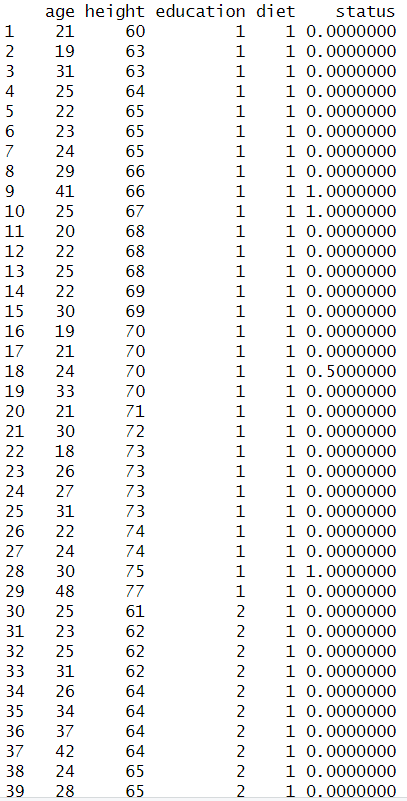
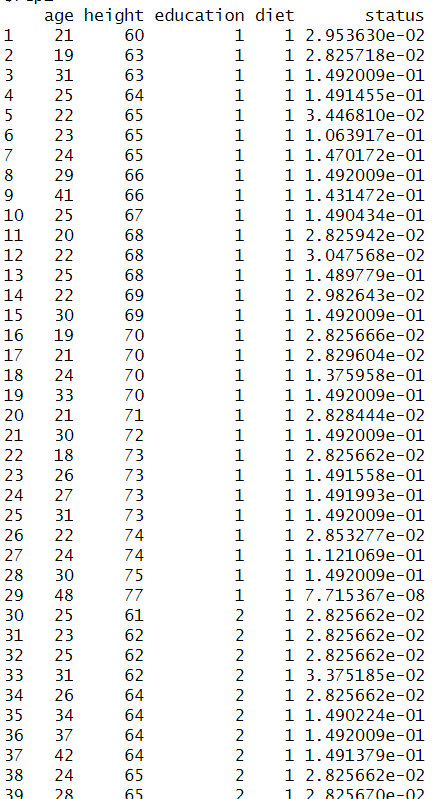
# 11.3.2 Neural network Accuracy

Accuracy of the Neural Network is found out to be 96%



# 11.4 Predictions

The predictions on status are made for the next 100 observations and classifies as single or not based on predicted values >0.5 and < 0.5 are single and not single respectively.



# 11.5 Important columns in determining the status of education

The most important determining factor for deciding the status is found out to be age. Followed by height, education and diet based on weights.



# Statistical analysis Report

The OkCupid dataset statistical analysis report reveals some intriguing insights regarding the factors that influence the relationship status of users. Analysis of exploratory data revealed that the dataset had a decent distribution of the objective variable "status," with approximately 60% of users being single and the remaining 40% being in a relationship.

Next, a logistic regression model and a neural network model were fitted to the dataset in order to predict the status of the users based on their age, height, education, and nutrition. Both models performed well on the dataset, with 94.18 percent accuracy for the logistic regression model and 96.0 percent accuracy for the neural network model. In terms of accuracy, it can be observed that the neural network model outperformed the logistic regression model.

Using the weights matrix of the neural network model, the significant factors that influence the relationship status were identified for further analysis of the dataset. The weight matrix revealed that age was the most influential variable, followed by education, height, and diet in decreasing importance.

In conclusion, the OkCupid dataset analysis indicates that the neural network model is superior to the logistic regression model in terms of accuracy when analyzing this dataset. The most influential factor on the relationship status of users is their age, followed by their education, height, and diet. These findings can assist OkCupid in optimizing their matching algorithms and increasing the success rate of relationships on the platform.

# Recommendations

On the basis of the analysis's findings, the following recommendations can be made:

1. The analysis indicates that age, height, education, and diet are significant predictors of relationship status. Therefore, it is suggested that future research concentrate on these variables to better comprehend how they affect relationship status.
2. Enhance data quality: The analysis reveals that the data contains absent values and inconsistencies. Therefore, it is suggested that future studies adopt measures to enhance the data quality. This may necessitate the collection of more comprehensive data or the use of sophisticated data cleansing techniques.
3. Use sophisticated modeling methods. Although the logistic regression and neural network models performed well, more sophisticated modeling approaches may improve prediction accuracy. Ensemble models or deep learning algorithms might increase forecast accuracy.
4. Assess additional factors. The investigation examines a few relationship status factors. To further understand relationship status factors, future study should include income, geography, and hobbies.
5. Base decisions on findings. The analysis may help people and businesses make better judgments. The results might help dating services improve their matching algorithms and match users. The results might help people understand successful partnerships and make better relationship choices.

Overall, the analysis provides valuable insights into the predictors of relationship status and highlights the importance of data quality and sophisticated modeling techniques when attempting to predict relationship status. Future research can build on these findings and increase our understanding of the factors that contribute to successful relationships by adhering to these recommendations.

# Conclusion

In conclusion, the analysis of the OKCupid dating dataset disclosed several insights regarding the most significant predictors of whether or not an individual is in a committed relationship. The dataset was subjected to exploratory data analysis to identify trends and patterns. In the subsequent statistical analysis, logistic regression and neural network models were trained to predict the relationship status of individuals based on their age, height, education, and diet. The accuracy of the logistic regression model was 94.18 percent, while the accuracy of the neural network model was 96.0 percent. Despite the fact that both models performed admirably, the neural network model was determined to be the superior method for analyzing this dataset due to its higher precision.

In addition, the analysis of the weight matrix enabled us to determine the most significant predictors of relationship status, with age being the most significant, followed by height and education. On the basis of these findings, we recommend that dating websites such as OKCupid give greater weight to these important factors when matching users with potential companions. For instance, they could employ algorithms that prioritize user pairings based on factors such as age, height, and education level.

This undertaking illustrates the effectiveness of statistical analysis and machine learning in obtaining insights from complex datasets. By utilizing these tools, we were able to obtain valuable insights into the most important factors for predicting relationship status, which could be utilized to guide future research in the field of online dating.

# 15. Bibliography