**CSIS 4495 (APPLIED RESEARCH PROJECT)**

**SUMMER 2019**

**YOUNG PEOPLE SURVEY**

**FINAL REPORT**

**SUBMITTED TO:** **SUBMITTED BY:**

Prof. Stephen Chiong Akhilesh Anand

300283773

Arzoo Sood

300288703

**DOUGLAS COLLEGE – CSIS DEPARTMENT**

<https://github.com/anandakhilesh/CSIS4495/blob/master/Final_Report.ipynb>

**TABLE OF CONTENTS**

[ACKNOWLEDGEMENT 2](#_Toc15845576)

[**INTRODUCTION** 3](#_Toc15845577)

[**OBJECTIVE** 4](#_Toc15845578)

[**SCOPE** 5](#_Toc15845579)

[**ANALYSIS AND TECHNIQUES APPLIED** 6](#_Toc15845580)

[**DATA COLLECTION** 7](#_Toc15845581)

[**DATA CLEANING** 8](#_Toc15845582)

[**Data Mining** 9](#_Toc15845583)

[Importing Libraries 9](#_Toc15845584)

[When the user needs to find out the size of the data file 9](#_Toc15845585)

[Descriptive Statistics 10](#_Toc15845586)

[**ANALYSIS OF DATA USING GOOGLE COLAB** 13](#_Toc15845587)

[Correlation 14](#_Toc15845588)

[Functions to draw Linear Regression Models 16](#_Toc15845589)

[Lmplot 16](#_Toc15845590)

[Regplot 17](#_Toc15845591)

[The Violin Plot 18](#_Toc15845592)

[Factor Plot visualization 22](#_Toc15845593)

[Bar plot visualization 22](#_Toc15845594)

[**MACHINE LEARNING METHODS** 32](#_Toc15845595)

[Regression 32](#_Toc15845596)

[Model 1 32](#_Toc15845597)

[Pattern Identification 37](#_Toc15845598)

[Clustering: 37](#_Toc15845599)

[Model 1 37](#_Toc15845600)

[Model 2 40](#_Toc15845601)

[Classification 41](#_Toc15845602)

[Model 1 41](#_Toc15845603)

[Model 2 45](#_Toc15845604)

[**CONCLUSION** 47](#_Toc15845605)

[**REFERENCES** 48](#_Toc15845606)

# ACKNOWLEDGEMENT

We would like to offer our sincere gratitude to our project supervisor Prof. Stephen Chiong for guiding us on how we should approach the problem. His input and guidance have been of tremendous value throughout our applied and research work. This project has taken a considerable amount of time period. And, multiple resources as well, we would like to acknowledge the help of all of those who have made the project possible. We are grateful to acknowledge numerous people who shared their ideas and work at Kaggle discussion forum which helped in approaching the analysis of this project.

# **INTRODUCTION**

Young people survey project is aimed to achieve a working model based on youngster’s data. The end product will be self-answerable to different questions which are raised from this dataset that can benefit stakeholders to decide and come up with products and services as per youngsters’ need. What makes this project unique is that different technologies are integrated like data analysis using Google Colab and passing that data to BigML for performing machine learning methods. Also, making this solution available online for others to give recommendation and modify it.

While working on this project, we learned coding python and its libraries such as pandas (provides high-level data structures), matplotlib (better visualization), NumPy (to manage multi-dimensional arrays), seaborn (the visualization of statistical models). Along with machine learning algorithms such as decision tree, regression and association.

Technologies used:

* Google Colab
* BigML
* Python 3 with following libraries:
* Pandas
* Matplotlib
* NumPy
* Seaborn
* Microsoft Excel

# **OBJECTIVE**

Day by Day the choices of youngsters are changing quite significantly. With the passage of time, young people’s liking has revolutionized. All the business industries and analysts must deal with the upcoming liking and disliking effectively and efficiently. This project would give them strategic preferences and help analyzing quickly. With the analysis of our data we try to find answers of the arising interrogatory questions. Some of the questions are:

To know what is most liked among youngsters

* To compare what is preferred more, movies or music
* To identify the most common type of fear among people
* To find if there are any dependencies in the data
* To figure out is there any correlation between phobias and gender
* To analyze whether people will spend more wealth on health or not

# **SCOPE**

The scope of the project states that this product will be self-answerable to different questions which are raised from this dataset that can benefit stakeholders to decide and come up with products and services as per youngsters’ need.

Who would benefit from this project and how?

* Youngsters: Would take the advantage of numerous choices.
* Business organizations: Can increase revenue by targeting most liked products
* Health industries: Can ameliorate the health services for the youth
* Entertainment industries: Can tweak up their genre
* Fear specialists: Can help the youth to eradicate their fears
* Researchers: Can make use of our findings, experiences and challenges to improve their results.

# **ANALYSIS AND TECHNIQUES APPLIED**

Following is list of some extended data analysis techniques and methods applied on young people survey dataset:

1. Data Collection
2. Data Cleaning
3. Analysis of data using Google Colab
4. Machine learning methods

* Regression
* Pattern Identification using clustering
* Classification

1. Visualization
2. Predictions

# **DATA COLLECTION**

It is a process of gathering and measuring information on targeted variables in an established system, which enables one to answer relevant questions and evaluate outcomes. Data is collected from a variety of sources. The requirements were communicated by Slovakians.

The data was collected by the downloads from online sources (kaggle.com) and reading documentations. The dataset was collected by surveying the students of the Statistics class at FSEK UK in 2013. It has been updated in 2017. The dataset is about the preferences, interests, habits, fears of young people. The survey was conducted in both written and electronic form. The participants were of Slovakian nationality aged between 15-30 years and later data was translated to English.

A screenshot of a cell phone

Description automatically generated

# **DATA CLEANING**

Data Cleaning is the process of detecting and correcting inaccurate records from a record set, table, or database and refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing, modifying, or deleting the dirty or coarse data.[1] Data cleansing may be performed interactively with data wrangling tools, or as batch processing through scripting.

After cleansing, a data set should be consistent with other similar data sets in the system. The inconsistencies detected or removed may have been originally caused by user entry errors, by corruption in transmission or storage, or by different data dictionary definitions of similar entities in different stores. Data cleaning differs from data validation in that validation almost invariably means data is rejected from the system at entry and is performed at the time of entry, rather than on batches of data.

A screenshot of a cell phone

Description automatically generated

# **Data Mining**

Data mining is the process of exploring data and finding patterns in it using machine learning, statistics, and database systems. The end goal of data mining is to derive useful information from data, which can be utilized to increase revenue, reduce costs, or even save lives through some of its applications. When you have a dataset that needs to be mined, it is not feasible to use all the data-mining techniques that are available on every column field of the data to derive insights. This will be a cumbersome task and will take a long time to derive any useful insights. To speed up the process of mining data, knowledge of domains is a great help. With this knowledge, one can understand what the data represents and how to analyze it to gain insights.

## Importing Libraries

Python modules can get access to code from another module by importing the file/function using import. The import statement is the most common way of invoking the import machinery, but it is not the only way.

A screenshot of a cell phone

Description automatically generated

## When the user needs to find out the size of the data file

This command is run to know the size of the dataset, that how many rows and columns are there. This shows that we have 1010 rows and 150 columns in the response dataset.

A screenshot of a cell phone

Description automatically generated

**Head command**

This is used to display the names of the columns in the dataset. If we do not set any value in the. head (0), it shows all the columns. Otherwise it will display the number of columns we have defined.

**Tail command**

This is same as head command but is used to display the column names from the end rather than the beginning.

A screenshot of a social media post

Description automatically generated

## Descriptive Statistics

Descriptive statistics is a term given to data analysis that summarizes data in a meaningful way such that patterns emerge from it. It is a simple way to describe data, but it does not help us to reach a conclusion on the hypothesis that we have made.

Like here, we used variable AGE to display all the summarized data for that field. It shows mean, count, standard deviation, minimum value, maximum value, Quartiles (25%- first quartile, 50% median, 75%- third quartile)

A screenshot of a cell phone

Description automatically generated

**Setting a theme for all the visualizations**

FiveThirtyEight, sometimes rendered as 538, is a website that focuses on [opinion poll](https://en.wikipedia.org/wiki/Opinion_poll) analysis, politics, economics, and sports blogging. One of the good websites to use for the theme visualizations.

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generated

A screenshot of a social media post

Description automatically generated

# **ANALYSIS OF DATA USING GOOGLE COLAB**

Question 1

**How can we map all the similar/homogeneous data to one group?**

This is the command used for mapping columns to one category. Then later using that category to analyze the data and answer some of the hypothesis.

A screenshot of a cell phone

Description automatically generated

Question 2

**What is the mean value of the Spending Category?**

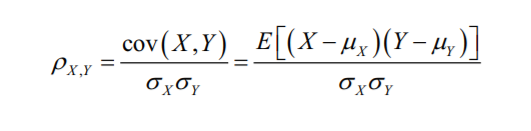
Now, we have picked one category and we are using to calculate the mean of each column in one graph. Here, we have used bar graph to show the differences between the same types.

A screenshot of a cell phone

Description automatically generated

### Correlation

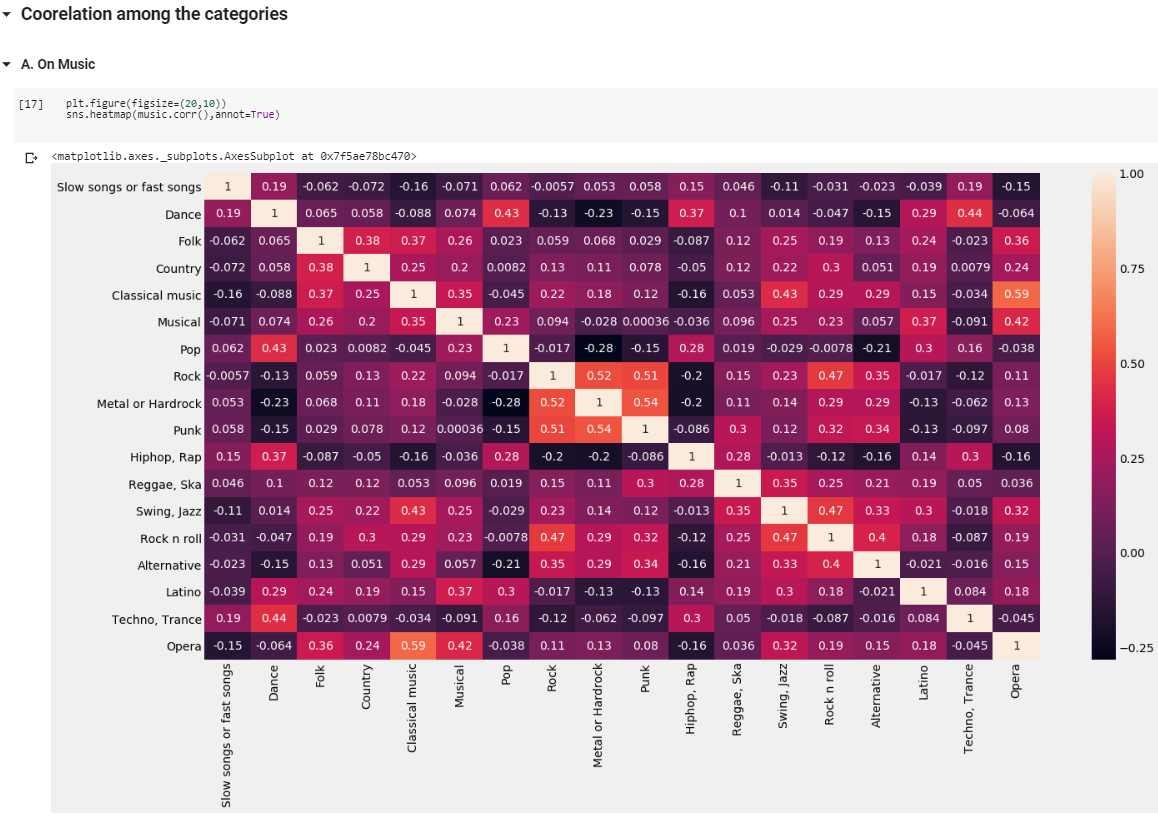
In statistics, correlation defines the similarity between two random variables. The formula defines the Pearson correlation as the covariance between X and Y, which is divided by the standard deviation of X and Y, or it can also be defined as the expected mean of the sum of multiplied difference of random variables with respect to the mean divided by the standard deviation of X and Y.



Question 3

**What is the correlation between the music types? What can you analyze from it?**

This correlation shows that whichever values are greater than 0.05 are considered in good correlation.

The categories that go together are Opera and Classical songs, Metal or Hardrock and Punk 

The categories that go together for Interests are Physics and Mathematics, Biology and Chemistry. A screenshot of a video game

Description automatically generated

### Functions to draw Linear Regression Models

In the simplest invocation, both functions draw a scatterplot of two variables, x and y, and then fit the regression model y ~ x and plot the resulting regression line and a 95% confidence interval for that regression:

Question 3

**What values for Height and Weight make the cluster?**

### Lmplot

Here, the visualization shows the cluster for Height and Weight. Most youngsters have weight in the range of 50-100 kg and the height ranges

A close up of a map

Description automatically generated

### Regplot

It shows the same height and weight cluster but using the scatterplot. It shows us the dot on the coordinated values. This is an easy way to see the individual values.

A screenshot of a cell phone

Description automatically generated

### The Violin Plot

It describes the visualization among two columns where one should be a numeric value and other can be the string or can’t be. Here it displays the values of all types of current smoker, former smoker, tried smoker, never smoked.

Question 4

**What is the distribution of Smoking based on the different age categories among various classes?**

A screenshot of a social media post

Description automatically generated

Question 5

**Calculating the ratio of female and male in the dataset.**

A screenshot of a cell phone

Description automatically generated

Question 6

**What is the quartile for the fear of spiders for each gender?**

The bp variable in the boxplot function is a Python dictionary with key values such as boxes, whiskers, fliers, caps, and median. The values in the keys represent the different components of the box plot and their properties. The properties can be accessed and altered appropriately to style the box plot to your liking. This shows that the females are more afraid of spiders.

A screenshot of a social media post

Description automatically generated

Question 7

**What are the outliers for the weight for each gender? How do you visualize the quartile of each gender?**

A screenshot of a cell phone

Description automatically generated

Question 8

**What is the count of Workaholism among youngsters for various ratings?**

A screenshot of a cell phone

Description automatically generated

Question 9

**What is the count of Fear of public speaking among youngsters for various ratings?**

A screenshot of a cell phone

Description automatically generated

Question 10

**How often do the youngsters use the internet? Analyze it on basis of gender and the age.**

The male and female ratio for the internet usage is described here, and age is another factor that is taken into consideration.A screenshot of a social media post

Description automatically generated

Question 11

**How often do the youngsters consume alcohol? Analyze it on basis of gender and the age.**

A screenshot of a cell phone

Description automatically generated

Question 12

**Visualize how often do young people smoke? Analyze it on basis of gender and the age.**

### Factor Plot visualization

A screenshot of a social media post

Description automatically generated

### Bar plot visualization

A screenshot of a cell phone

Description automatically generated

Question 13

**Is there any relation between the left-handed people and right-handed people? Do they have any differences?**

As per research, there are differences between left-handed people and right-handed people. Research says that left-handed people are more intelligent, wealthier, are healthier and are more dedicated towards their work.

Keeping that in mind, we tried some analysis and found that the research is valid for this dataset as well.

Question 14

**Display the count of left-handed and right-handed young people.**

A screenshot of a cell phone

Description automatically generated

* Left-Handed Vs Right-Handed on basis of wealth

A screenshot of a social media post

Description automatically generated

* Left-Handed Vs Right-Handed on basis of Workaholism

A screenshot of a social media post

Description automatically generated

Question 15

**What can be analyzed from the Height and Weight of the user?**

The height and weight of the user can be used to calculate the BMI (Body Mass Index).

In this, we have divided all BMI into different categories. Categories are:

* Very severely underweight
* Severely underweight
* Normal Weight
* Overweight
* Moderately obese
* Severely obese
* Very severely obese

A screenshot of a social media post

Description automatically generated

A screenshot of a cell phone

Description automatically generated

Question 16

**Visualize the different groups of BMI as per their count and categories.**

A screenshot of a social media post

Description automatically generated

Question 17

**What is the distribution of Bmi Category with respect to Age and Gender?**

A screenshot of a cell phone

Description automatically generated

Question 18

**Do you find any relation between the BMI of a youngster and spending on healthy eating habits for an age?**

A screenshot of a cell phone

Description automatically generatedA picture containing screenshot

Description automatically generated

Question 19

**What are the ratings given by youngsters to Empathy? What is the count of each rating?**

A screenshot of a cell phone

Description automatically generated

Question 20

**To test the hypothesis whether people living in village have more empathy as compared to city living people.**

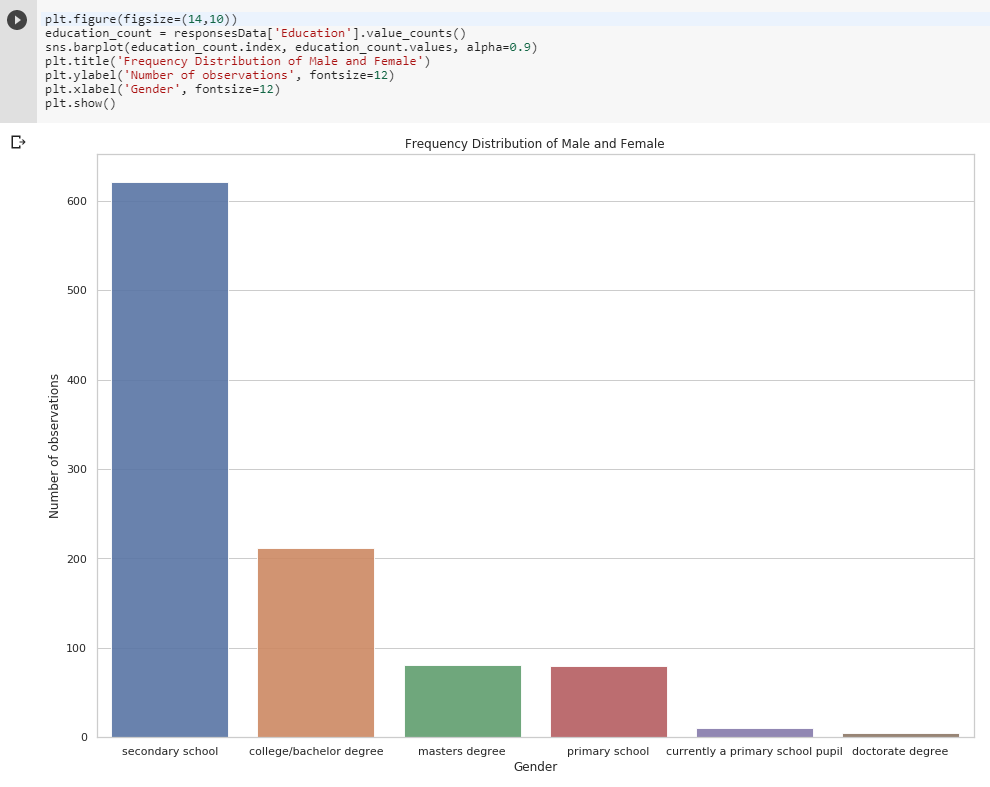
A screenshot of a cell phone

Description automatically generated

Question 21

**To count the number of individuals for categorized education.**

Using the bar plot, we are implementing this. We can set the title with plt.title.



Question 22

**Count the number of only child and number of observations for not only child.**

A screenshot of a cell phone

Description automatically generated

Question 23

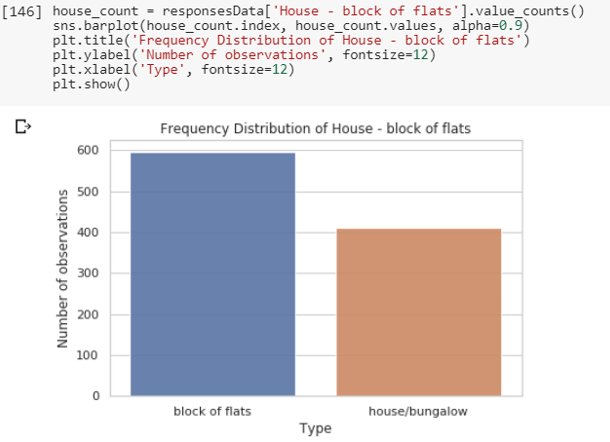
**To count the number of observations for young people who are living in City or in the village.**

A picture containing screenshot

Description automatically generated

Question 24

**To know the value of people living in flat or house/bungalows**



Question 25

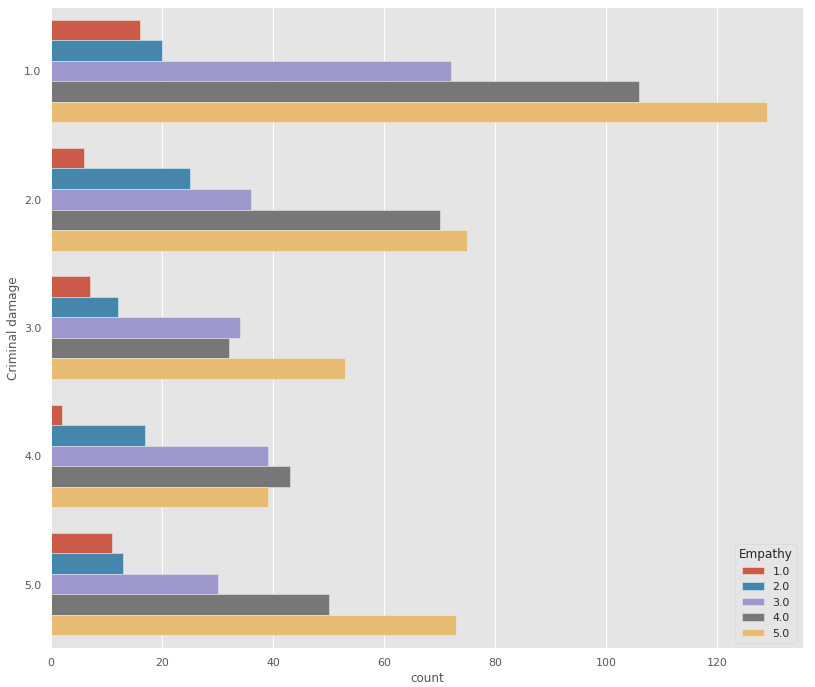
**What is the number of youngsters who like Mathematics?**

A screenshot of a cell phone

Description automatically generated

Question 26

**What is the relation between empathy and criminal damage?**



# **MACHINE LEARNING METHODS**

Machine learning is a data analytics technique which enables computer programs to learn from data and improve from the experience of its own. It basically involves continuous improvement of PC programs that can get to data and use it to learn for themselves.

Following are list of machine learning analysis we have performed:

## Regression

Regression analysis is a predictive modelling algorithm used to find pattern between the outcome (dependent variable Y) and input (independent variables X). It determines a mathematical equation to predict the dependent variable Y based on independent variables X.

Following are types of regression analysis:

1. Linear regression - the outcome variable is continuous.

Example: income (Dependent) ~ age and education (Independent)

1. Logistic regression - the outcome variable is categorical.

Example: two valued outcomes true/false, yes/no or pass/fail etc.

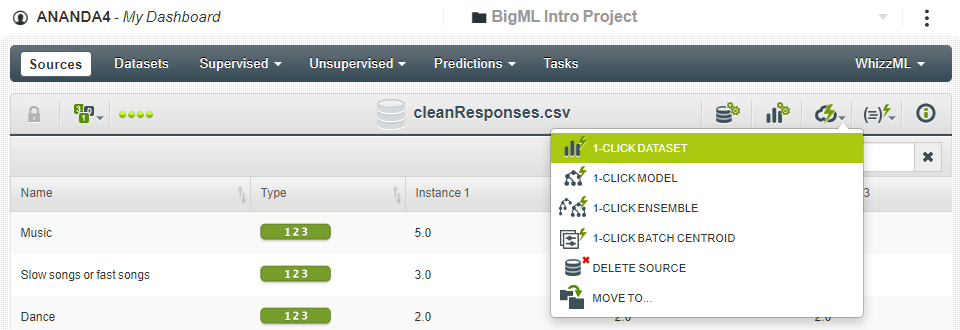
### Model 1

First model we created is the gender prediction regression, in this model the independent variables are height, weight, romantic and life struggles. We wanted to see if regression modelling can predict the gender based on sets of different personality and physical traits.

Following are the steps performed to get desired output.

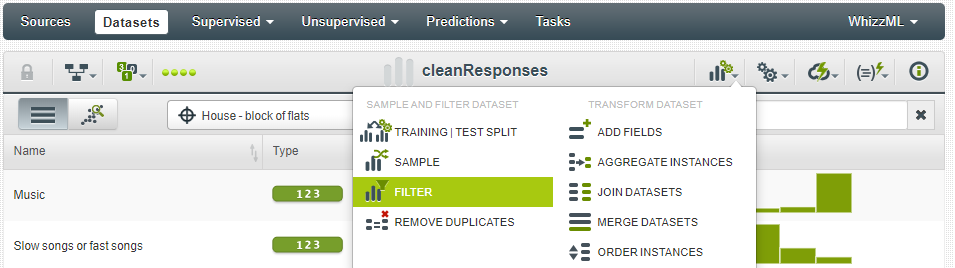
Step 1: Creating Dataset

The clean responses file that was created earlier using Google colab is uploaded to bigML. Once uploaded, open that file and click on following option “1-Click Dataset”:

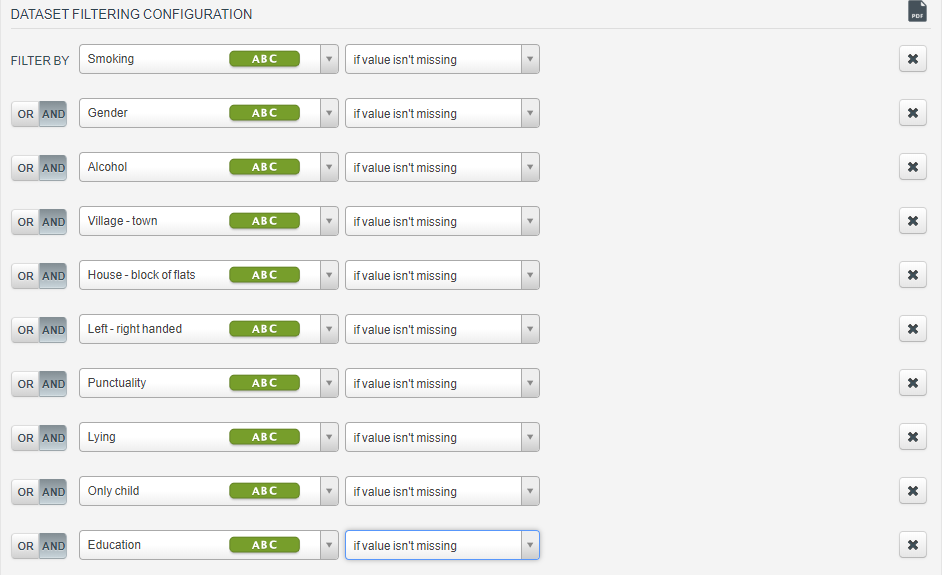


Step 2: Filtering the NULL values

Once created, open the dataset and click on filter option as shown below:

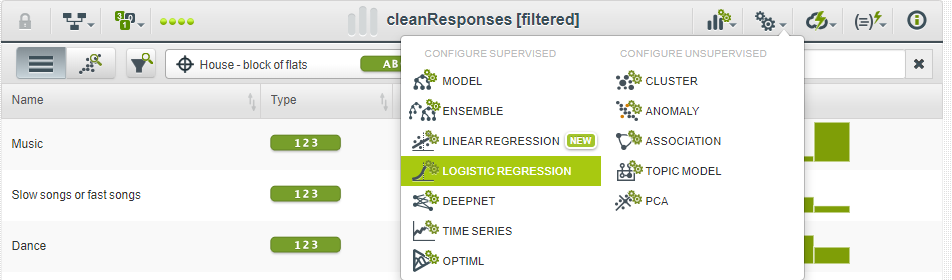


Add following fields, select the option “if value isn’t missing” and click “Create Dataset”:

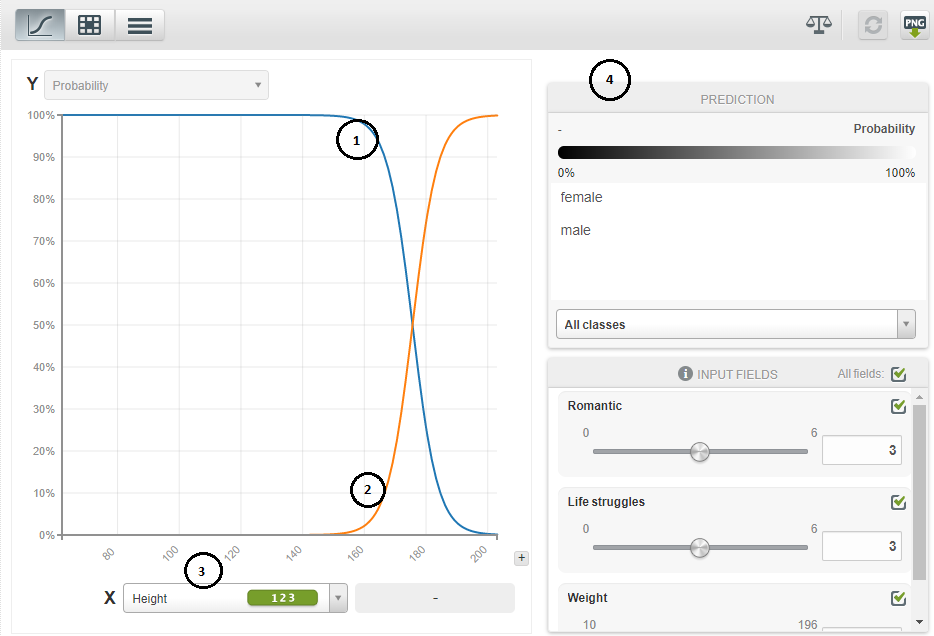


Step 3: Creating Logistic regression model

Once the dataset is filtered



In configuration screen fill the objective field as “Gender” and check the options height, weight, romantic and life struggles. After clicking create regression following will be displayed:

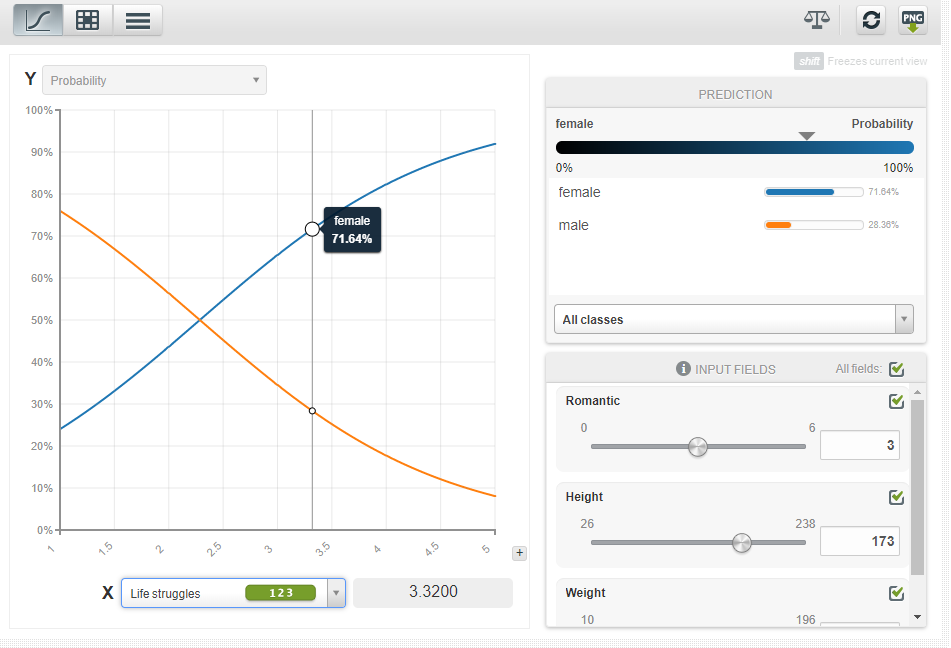


This is our regression model based on the selected variables. There are some observations that can be made here as follows:

1. Point 1 represents female prediction. When X (height) is below 160 the model is 100% confident that the independent variables describe female.

Note: The point 3 is representing X – axis. We can change it to any of the independent variable we selected. This is explained in following steps.

1. Point 2 represents male prediction. We can note that probability of male increases when height is more than 160.
2. Point 3 can be changed to either weight or romantic or life struggles. And based on that the point 1 and point 2 changes accordingly as shown below:

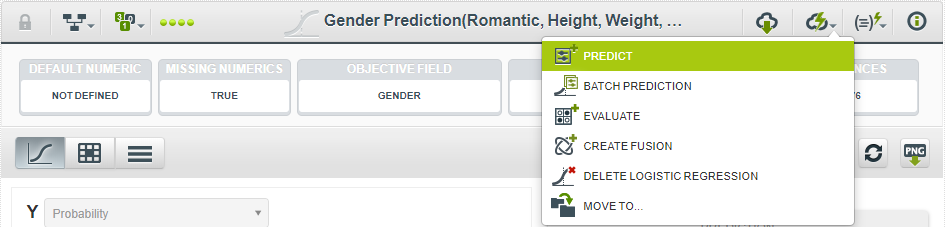


Here, we changed the X to Life struggles and accordingly the probability of male and females’ changes.

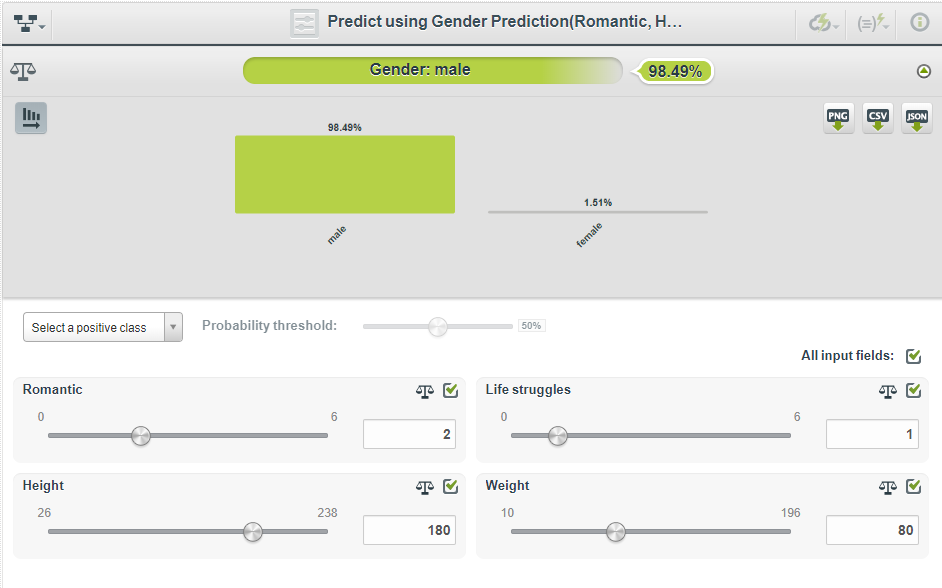
1. Point 4 provides more detailed values about regression model. It changes in real time as we select any point in graph.

Step 4: Predicting with regression model

Once regression model is created successfully. Click on the Predict option as shown below:



We can predict the gender by giving entering values for independent fields. Fill the following fields as shown:



Here, we can see that the model is predicting gender of person as Male. We successfully created our first regression model.

## Pattern Identification

### Clustering:

Cluster analysis or clustering creates groups of objects. Every object in a group or cluster behaves or contains attributes which are like other objects belonging to same group. But every cluster has different attributes to compared to other clusters.

Example: Image analysis, pattern recognition, computer graphics etc.

### Model 1

Clustering can be very useful when it comes to deciding which object can be best relate to the current one. These days everyone uses Spotify, Apple music and many other music streaming services. It’s crucial for the companies to generate relevant suggestion as per persons liking. In this scenario clustering comes in place. Clustering creates cluster of similar songs or genres to create better suggestion for the user.

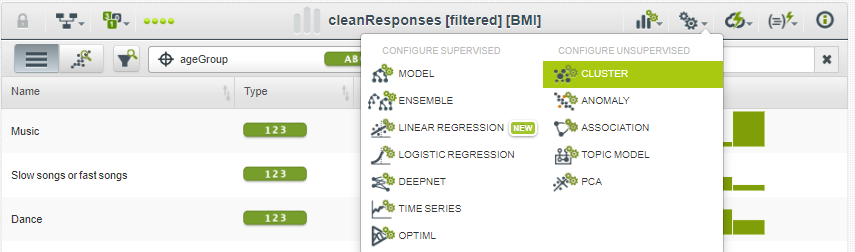
From our dataset we have applied clustering on the music preferences and then created a separate file (also available in excel format) which contains the all genres, their cluster number and distance of each object (genre) from the centre of the cluster. Following are steps to get the desired output:

Step 1: Creating Dataset and filtering

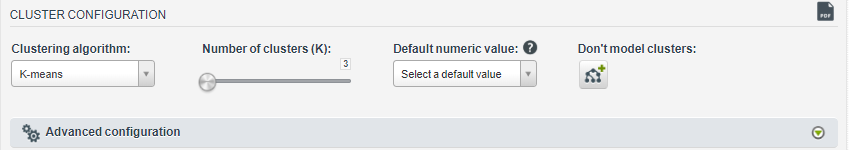
Follow same steps as mentioned in regression analysis Model 1 to get the dataset and perform filtering. Alternatively, the dataset created earlier in regression analysis Model 1 can be used.

Step 2: Creating classification modelling

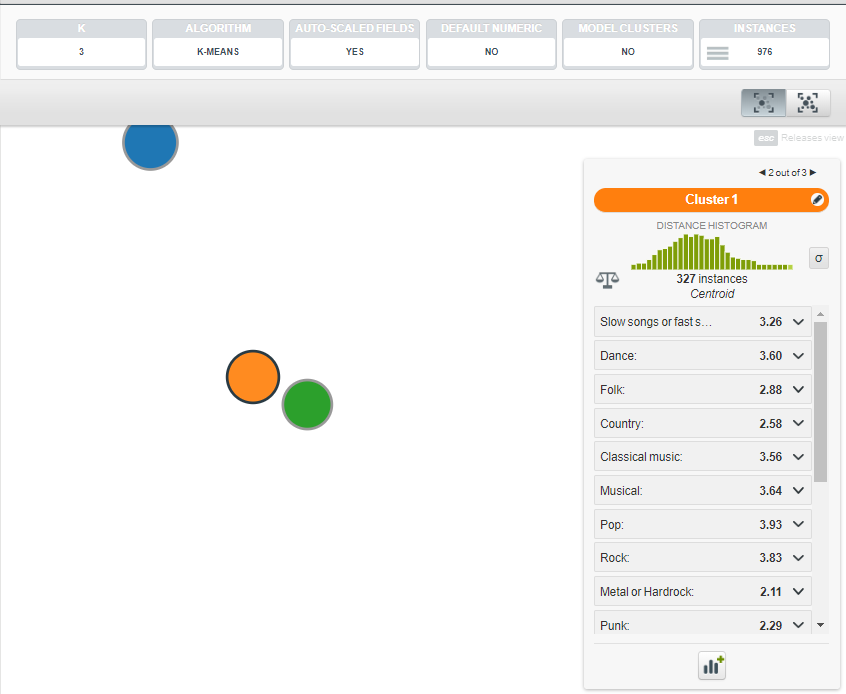
Click on Cluster option from the dataset as shown below:



In the cluster configuration form fill the number of clusters (K) as 3 and select these fields: Slow songs or fast songs, Dance, Folk, Country, Classical music, Musical, Pop, Rock, Metal or Hardrock, Punk, Hiphop, Rap, Reggae, Ska Swing, Jazz, Rock n roll, Alternative, Latino, Techno, Trance and Opera. After that click on Create Cluster.



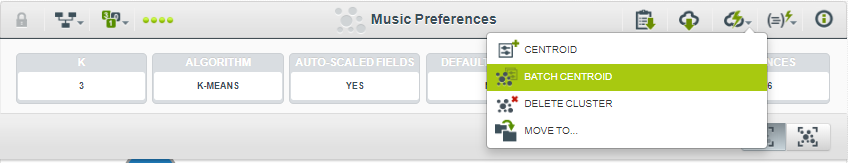
After some processing following would be displayed:



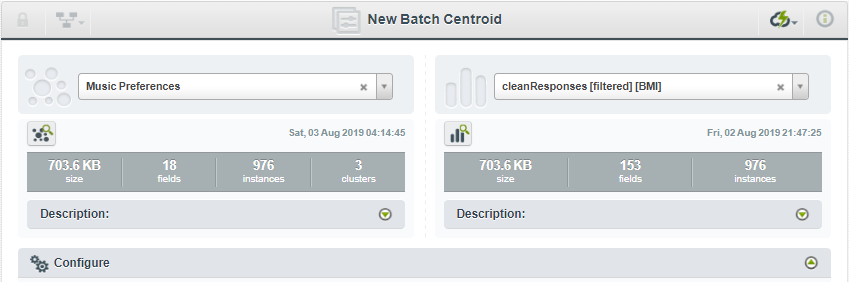
We have created 3 clusters. On the right side, detailed analysis of the selected cluster is shown including number of instances and mean values of every field.

Step 3: Creating batch centroid

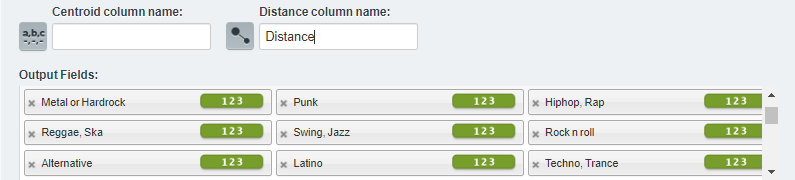
Click on Batch centroid option as shown below:



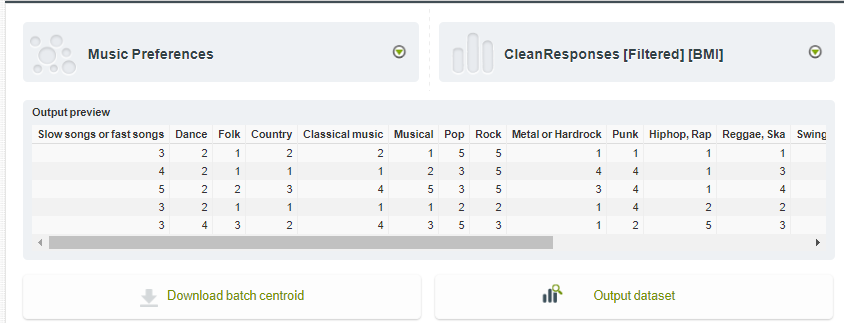
Then fill the following details:



Click on Configure -> Customize centroid output settings -> Output fields: Only keep the music preferences field and remove all other. And fill the distance column name as “Distance” as shown below:



Then click on Centroid button at the bottom of the page. Following is the output of batch centroid:



We can download the batch centroid. The downloaded file contains all the cluster details. Their values and distance of every object from the centre. This can be used for making suggestion for genres.

### Model 2

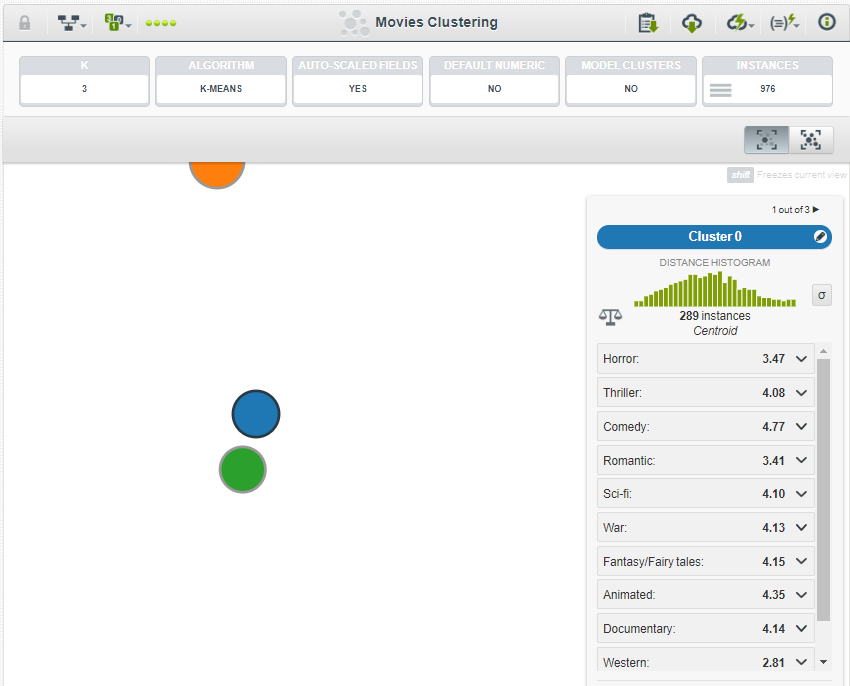
Like the model 1, in model 2 we are creating clusters on movie preferences.

The steps for movie preference clustering is similar as model 1. Apart from the following mentioned differences:

* Selecting the fields

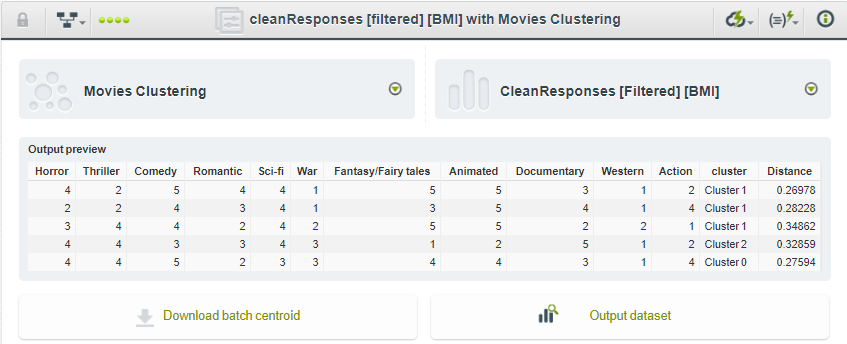
In cluster configuration window. Select Horror, Thriller, Comedy, Romantic, Sci-fi, War, Fantasy/Fairy tales, Animated, Documentary, Western and Action options only.

Output:



We have created 3 clusters. The right side lists detailed analysis of the selected cluster. It shows number of instances and mean values of all fields.

After getting the correct output, we can create the batch centroid cluster like we did earlier.



## Classification

Classification is a supervised method which can be used to make prediction based on historical data. There are different ways to implement classification. We have used decision trees on interests of youngsters based on place they live i.e. City or Village.

### Model 1

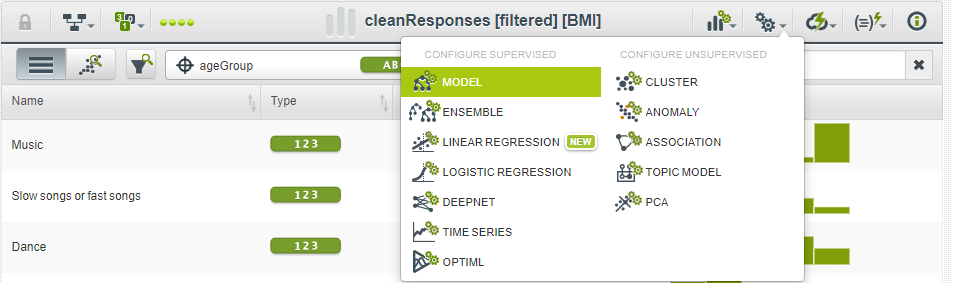
In this model we performed classification on interests of youngsters based on place they live i.e. City or Village. Following are the steps performed to get desired output:

Step 1: Creating Dataset and filtering

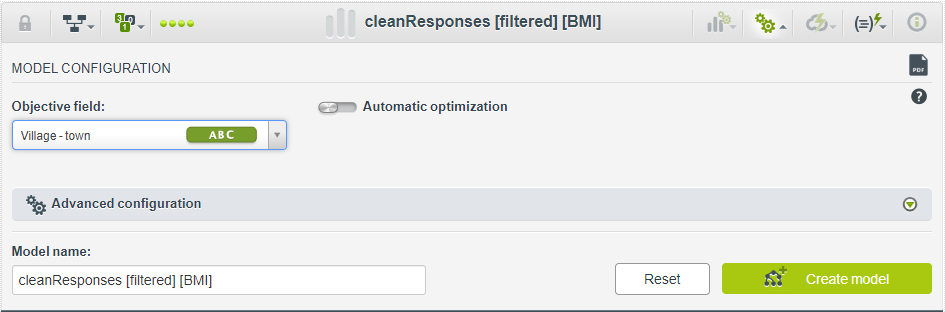
Follow same steps as mentioned in regression analysis Model 1 to get the dataset and perform filtering. Alternatively, the dataset created earlier in regression analysis Model 1 can be used.

Step 2: Creating tree model

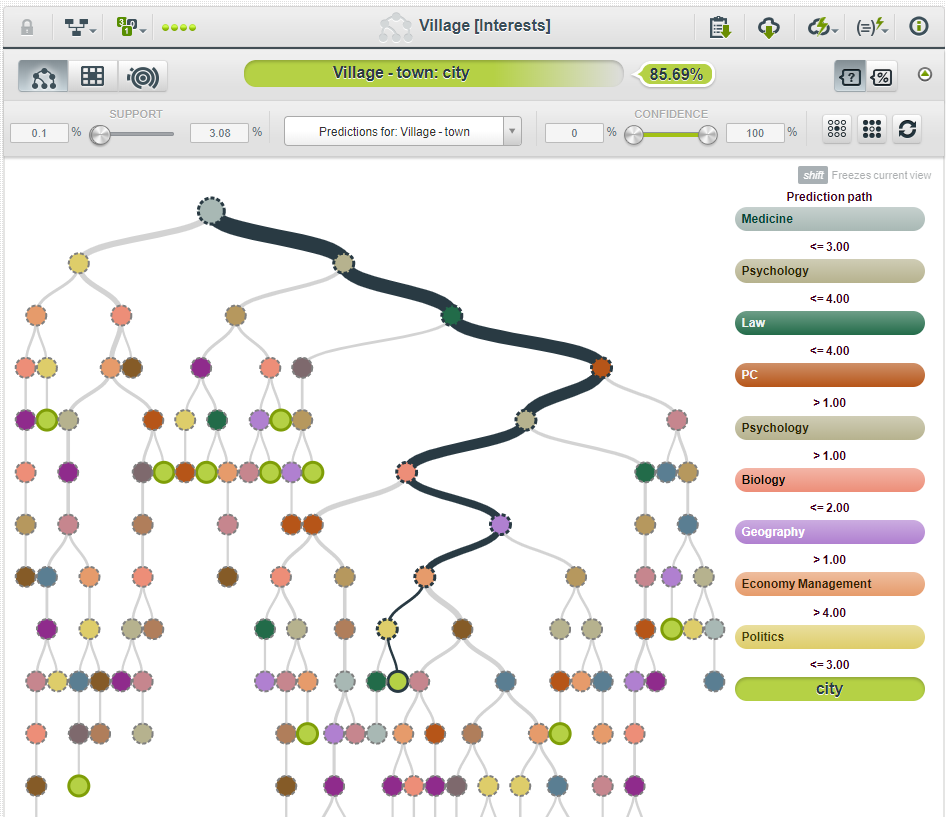
Click on Model option as shown below:



After that click enter the objective field as “Village – Town” and check the interests which are: History, Psychology, Politics, Mathematics, Physics, Internet, PC, Economy Management, Biology, Chemistry, Reading, Geography, Foreign languages, Medicine and Law.



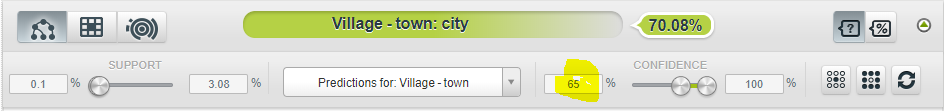
Then click on “Create model”. After some processing, tree model would be shown as below:



Here, every branch is decision represented in form of a tree or can be said as series of if else statements. The root node for all decision is P-Path.

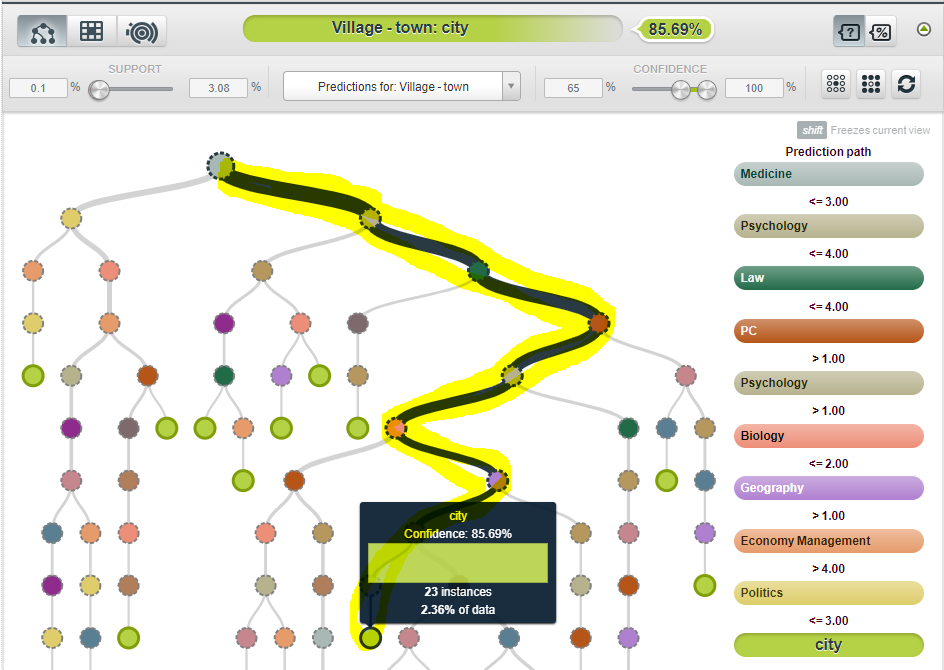
Step 3: Selecting Confidence Interval:

In previous step, default confidence interval was from 0 to 100%. For the analysis we need good confidence like more than 65%. Change the lowest confidence value to 65 as shown below:



Step 4: Understanding branches:

Selecting any of the branch from the tree, a change in prediction path can be seen. Try and select the yellow highlighted node as shown below:



* Point 1: This is the prediction path, here all the condition and their responses are shown. Example Medicine <= 3.00.
* Point 2: Shows the output as city predicted from the selected prediction path.
* Point 3: Shows the 85.69% that shows how confident our model is in the prediction.

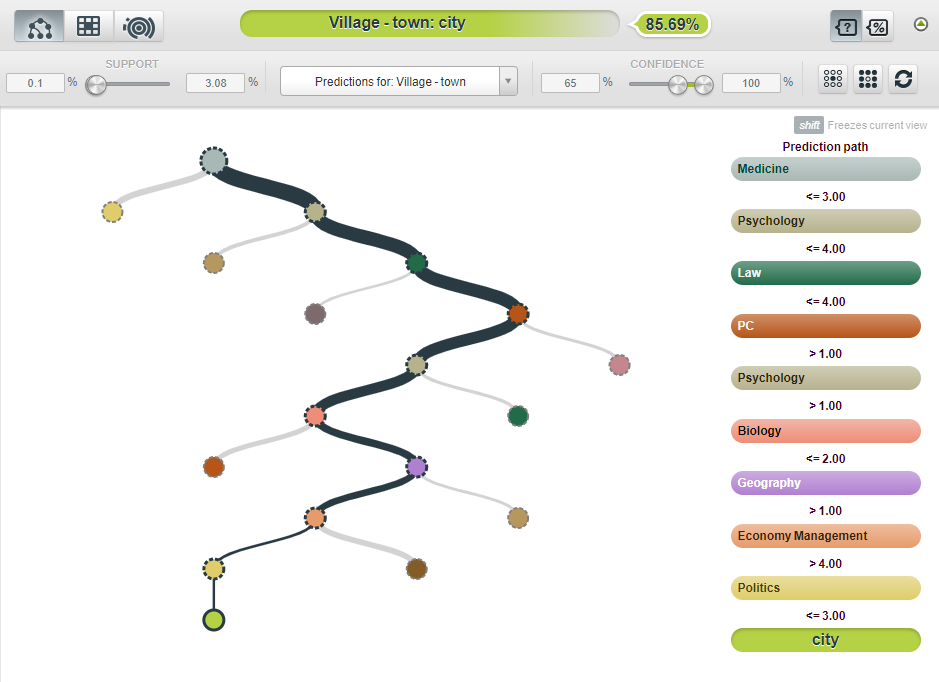
Result:

Expected Village - town = City

Confidence = 85.69%

Step 5: Understanding Subbranches:

On clicking a node, more detailed branch is shown along with the prediction path used to reach to selected node. Select the same node from Step 4. Following is shown:



The predicted path is same as Step 5. But as one can see the branch is more descriptive.

Above steps helps to create a tree model for prediction. This model clarifies every step taken by the model to predict the output in form of tree branches. User have control over the selection of important parameters such as confidence interval and support.

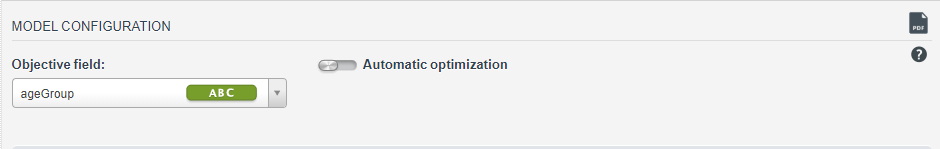
### Model 2

In this model we performed classification on spending of youngsters based on their age group. The steps performed are like Model 1:

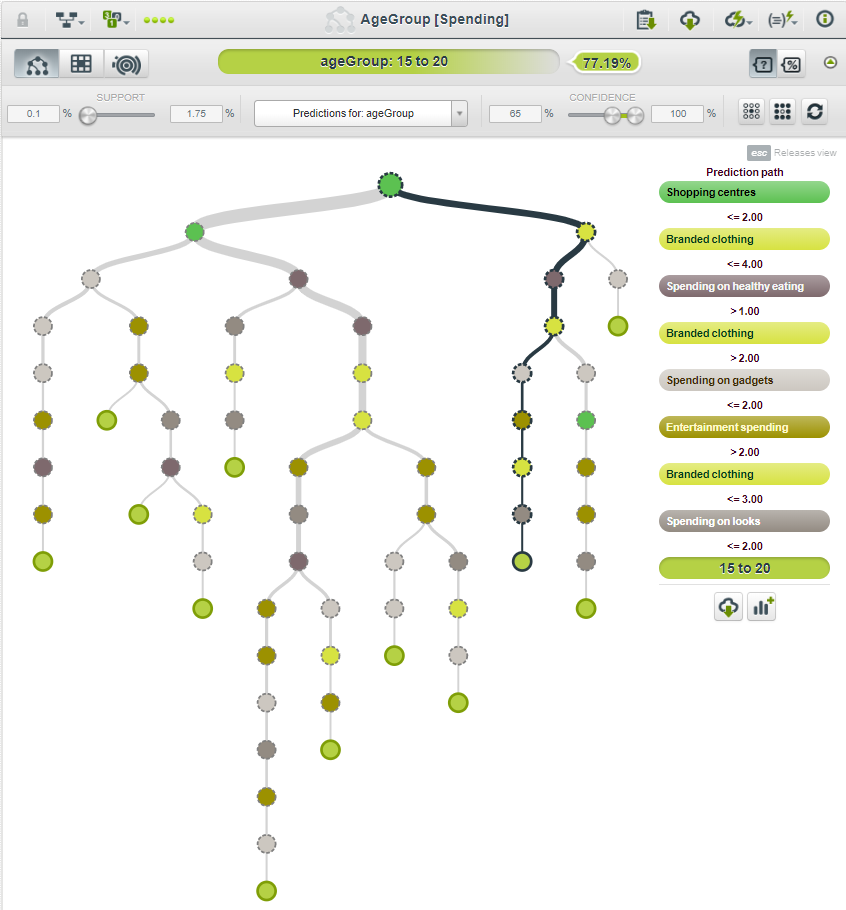
Changes to make:

* Creating tree model

Select the objective field as “ageGroup” and check the spending fields which are: Finances, Shopping centres, Branded clothing, Entertainment spending, spending on looks, Spending on gadgets and Spending on healthy eating.



Output:



# **CONCLUSION**

To conclude, we can say that the analysis that we have done in the project were found to be same as the actual findings by the researchers.

* Left-handed founded to be healthier, wealthier and merrier than right-handed.
* Women are more afraid of fear rather than men.
* Youngsters living in city spend more on their healthy eating habits, looks, gadgets as compared to village youngsters.
* Empathy and criminal damage ratings don’t go in spiral.
* Internet usage is high among males.

# **REFERENCES**

* [https://bigml.com](https://bigml.com/)
* [https://colab.research.google.com](https://colab.research.google.com/)
* <https://github.com/anandakhilesh/CSIS4495>
* <https://matplotlib.org/>
* <https://pandas.pydata.org/>
* <https://www.youtube.com/watch?v=ICQ4H2ms0sU>
* <https://towardsdatascience.com/getting-started-with-google-colab-f2fff97f594c>
* <http://nuovolabs.fauser.edu/~valeria/materiale-didattico/python/Packt.Mastering.Aug.2015.ISBN.1784390151.pdf>
* <https://scikit-learn.org/stable/auto_examples/model_selection/plot_learning_curve.html>