Individual Attributes and Social Network Predictors for Sexual Behaviour Profiles Among Adolescents in a Rural Area

Isabella Rodas Arango

MSc. Degree Thesis, Department of Industrial Engineering, Universidad de los Andes,
Bogotá, Colombia
Email: i.rodas@uniandes.edu.co

Decision-makers in the school Institución Educativa de Santa de Ana, located in a small municipality of Barú are interested in evaluating the sexual behaviour profiles found in the teenagers attending their institution. This information is useful for improving sexual health within the community and lower the number of teenage pregnancies expected in the future. Using data collected from a self-reported survey administered to the students from the sixth to the eleventh grade of the school, this study investigates the individual and social network predictors for the different sexual behaviours profiles found in the community. In the first place, 5 different sexual behaviour profiles using Factor Analysis and classic clustering methodologies were found and the age at which every participant touche someone else's private parts, what contraceptive method are they using when being intimate, the reason why they are not using any contraceptive, how close they are to their peers and how influential they are within their group of friends seem to be the most influential variables to predict the sexual behaviour profiles of an individual in this particular population.

Keywords: Social Networks Analysis, Adolescent Sexual Behaviour Profiles, Predictive Models, Multinomial Ridge Regression, Factor Analysis, Clustering

Received 29th November, 2021; Revised 07 December 2021

1. INTRODUCTION

Risky sexual behaviour among teenagers is noted as a public health care concern in different countries across the globe, including Colombia. This behaviour includes participating in sexual acts prematurely, engaging in sexual activity with multiple partners in a short amount of time, and drug and alcohol abuse prior, during or after sexual acts [1]. These activities are considered risky due to their deleterious health and psychological consequences including contracting sexually transmitted diseases (STD), unplanned pregnancies, dating violence and sexual assault in countries like the United States [2].

Several studies indicate risky sexual behaviour to be a prevalent occurrence among many adolescents depending on their friendships and peers influence [2, 3, 4]. These explain risky sexual behaviour given the teenagers social interactions and social attributes in high income countries but very few have explored the different predictors for these behaviours in low income countries or rural areas. Therefore, given the many negative consequences the high rates of occurrence of risky sexual behaviors among teenagers and the lack of studies regarding this topic in rural areas, this papers aims to use individual and social network attributes to predict sexual behaviour profiles among teenagers in a rural area of Colombia.

The main objective of this paper is achieved by analysing the data from a survey administered to a group of teenagers from the ages between 13 to 17 in the rural island of Santa Ana, Barú in Colombia. The individuals in this survey will be classified in groups using a classic clustering techniques to determine the different groups of behaviours that this particular population is having regarding their sexual activity. After that, the most relevant survey questions are selected using the Chi-Squared correlation index and network centrality measures are estimated in order to use them both as parameters in the three predictive models that are developed. These models proves to segment the participants in the most susceptible profiles, characterize them and compare them to the initial classification. They are also included to estimate the effects of each attribute or question on the response variable set as the sexual behaviour classification each subject received.

We found 5 main profiles for different sexual behaviours in this population. These profiles outline that the reason why subjects do not use birth control, the fact that they already have touched someone else's private parts, how old where they when they did, how influential they are in their friendship network and how close they are to their friends has as significant predictors for the sexual behaviour profiles. This paper is organized as follows. In the next, Section 2, we discuss the related literature. Section 3 includes the methodology used: data, subject classification, variable selection and predictive models. Section 4, provides the results from the study case. Section 5 discusses the findings and possible recommendations for the decision-makers in the near future and the last Section(6) concludes and outlines future work.

2. RELATED WORK

Most of the studies found that relate social networks to the sexual behaviors of adolescents use statistical methods to define the correlation between the variables evaluated. Authors such as Asrese, K., & Mekonnen, A. (2018) in their study *Social network correlates of risky sexual behavior among adolescents in Bahir Dar and Mecha Districts, North West Ethiopia: an institution-based study* [5] they have already explored the relationships between social networks and risky sexual behavior in adolescents. In this study, a survey of 806 individuals was used to collect the necessary information and subsequently hierarchical logistic regressions were used for the analysis. With these tools, it was possible to reach the conclusion that social networks predict adolescent sexual behavior more accurately. The findings indicate that the context and relationships of each individual can have both a risky and a protective effect on adolescent sexual behaviors.

Likewise, the authors Treboux, D., & Busch-Rossnagel, N. A. (1990). in his study *Social Network Influences on Adolescent Sexual Attitudes and Behaviors.* [4] They used the *LISREL* methodology to analyze whether the influence of social networks by both parents and peers affects knowledge of contraceptive methods and sexual behavior. After conducting a survey of 361 adolescents, they concluded that this influence changes significantly according to the adolescent's gender and whether or not they had already started their sexual life.

Other authors, such as Nicholsona, Marcumb, and Higgins (2020) in their study *Predictors of Risky Sexual Behavior among High School Students in the United States* used the data of the Youth Risk Behavioral Survey[6] together with bivariate regressions to find predictors of risky sexual behavior in adolescents in the USA. The authors conclude that the use of psychoactive substances before sexual activities and having multiple sexual partners are the best predictors for risky sexual behavior in adolescents.

3. METHODOLOGY

3.1. Data

The data used in the present study was gathered in the ongoing research project *isBaru: the importance of social behaviors in the health of adolescents in rural areas from Colombia* carried out by Florida International University, Universidad de Los Andesstudents and in partnership with NGO Amor por Barú. This particular survey was administered to 242 teenagers between the ages of 13 to 17 that attended the school Institución Educativa de Santa de Ana, located in a small municipality of Barú, Colombia, on April of 2019.

The subjects completed several self-report measures about their self-perception, families and peers, as well as their behaviors related to sexual activity and substance use. Around 50% of the population surveyed identify themselves as female, 49% as male and the remaining 1% as other gender. Additionally, the 51% were 14 or 15 years old and in the eight, ninth, or tenth grade at the time of the survey. 76% identify themselves as Afro-Colombian, while only 8% identify themselves as Caucasian. From the participants, 24-68% of them answered they lacked food in their houses, stating that: in the last month, 1 or more times someone in their family could not eat the 3 meals because of food shortage. Additionally, 24.2% of the participants stated that they worked and 17% of them answered that their family expected them to do so, in order to contribute to the household's expenses with their income. The aformentioned characteristics make of this population somewhat homogeneous for the subsequent analysis that took place.

Regarding sexual conduct and behaviours, 67% of the participants were or had been in a romantic relationship, 76% had held hands with a significant other, 55% had already been kissed for the first time, 26% had already had a sexual encounter and that same 26% had already had their sexual debut (12.5% before the age of 10).

On the other hand, the subjects where also asked to name up to 10 different peers from their school who they consider their friends, understood as, those who they talk with about what is important for them. This to recreate the positive relationships' network between the participants of the survey (i.e. a friendship network). In Figure 1 a graphic depiction of the network is shown. Every node represents a participant and every connection symbolizes a friendship nomination, the degree and *betweenness* centrality of each node are represented by the size and color of each of them respectively.

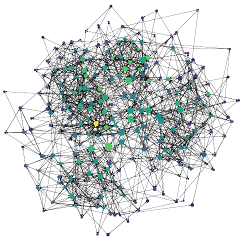


FIGURE 1: Positive relation network for the population in the study.

3.2. Variable Selection

As mentioned before, the survey used for this study had multiple blocks of topics and questions with different purposes. As not all of those variables were related with sexual behaviour, a process of variable selection based on literature review of similar studies was carried out.

The following topics were included in the analysis: age, grade, academic achievements, parents' education and relationship, family poverty, psychoactive substances consumption and romantic relationships, as per mentioned in the John Hopkins 2019 study [7]. Taking into account the age and grade is necessary because teens who have already experienced a their first sexual encounter or relationship are generally older than those who have not [8]. Academic behaviors and psychoactive substances consumption were taken into account because these factors may be associated with friendships selection [9, 10]. Also psychoactive substances consumption before sexual intercourse has been linked to risky sexual behaviour in several studies [11]. Because family disadvantage and family structure have also been associated with risky sexual behaviour and unwanted pregnancy [12, 13], family characteristics were taken into account to indicate parents' education level, and lack of parental figures in the household. Romantic relationships were also considered as having a stable romantic relationship has been linked to fewer sexual partners and safe sex practices [1].

Having selected the topics, 123 questions of the total survey were remaining. However, in the survey there was not any question that directly measured sexual behaviour. This is why out of the 123 questions 14 where selected according to [1] that better represented sexual behaviour and were used to categorize each of the participants in a group according to they answers. These questions are shown in detail in Appendix A. The remaining 109 questions were used to classify the subjects into the groups that were found with the 14 previous ones.

3.3. Data Cleaning

The survey used for this study displayed questions depending on the previous answers of the students. For this reason, some of the questions were never shown to some of the subjects depending on their answers. This produced many missing values in the original data and it is why the variables of interest were observed one by one, eliminating missing data due to questions that were not shown, these cases were assigned to category *0 (Not Applicable)*.

After the missing data associated with the survey's logic was processed, each missing value was tested to identify if the missing data was completely random or had a relationship with the other observed variables. Every single variable was tested against the others using Chi-Square's Test of Independence. This test is used to determine if there is an association between two (categorical) variables. Cramer's Coefficient was also used as this test is used to measure the association between (categorical) variables [14]. After these tests were conducted and the association between the variables was statistically proven the remaining missing data was imputed using the *Multiple Imputation by Chained Equations* method. This method was selected as it is capable of estimating values in such a way that the descriptive parameters of the data set are preserved and, therefore, the analytical methods of subsequent complete data would be minimally affected by the points of missing data [15].

3.4. Subject Classification

After the data was cleaned and completed subjects were classified in different groups according to their answers in the 14 questions outlined in Appendix A. Since this type of variables are highly correlated a transformation was needed to find the underlying structure that the variables had, to group coherently all the subjects [16]. To know if in fact this type of methodology was suitable for the data in question Barlett's Test and Kaiser-Meyer-Olkin (KMO) Test were administered. After the aforementioned test were successful factor analysis was conducted and an optimal number of factors were chosen to describe the data using eigenvector analysis. Finally, the adequacy of the factors was confirmed using the Crohnbach-Alpha Tests and the variance explained by each of the factors [17].

After having the factors outlined, the total of observations was grouped using different clustering methodologies and evaluated using the Silhouette, Calinski Harabasz and Davies Bouldin indices. The Silhouette Index was used to measure how similar are the observations inside each cluster compared to the observations in other clusters. On the other hand, Calinski Harabsz Index represents the ratio between the dispersion found in each of the clusters and the dispersion between the totality of clusters. Davies Bouldin Index provides information about the distance between clusters compared with the the size of the clusters themselves [18]. The different methods used to cluster were:

- Affinity Propagation
- Agglomerative Clustering
- Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH)
- Ordering points to identify the clustering structure (OPTICS)
- K-Means
- Mean Shift
- Spectral Clustering
- Gaussian Mixture Model

After evaluating all the different clustering models a grouping was chosen given the results obtained. This clusters were considered the predictor variable in the rest of the study.

3.5. Network Analysis

Different studies suggest that teenagers tend to mimic sexual behaviours of their peers [7, 11, 4, 19]. This is why, in this study the friendship network of the participants is analyzed in relation to their sexual behaviours. With this in mind, the friendship network outlined in Section 3.3 was used for this analysis. The resulting network is a direct network, as receiving a nomination and making one are considered two different aspects of each node in this study.

From the aforementioned network the more relevant measure are obtained and later used in the Predictive Models (Section 3.6). The first set of measures that were obtained from the network are the *degrees* and *motifs*. The *indegree* of each node refers to the number of nominations received from other subjects. While the *outdegree* is the number of nominations sent from each node. Depending on the proportions of each, the students will have distinct roles in the network, this role will be refer as the *motifs*. A higher *indegree*, or also called sociometric popularity reflects how likeable the adolescent is in his network. There are two main motifs taken into account in this study, the *bottle necks* and the *sinks*. The former are participants who receive more friendship nominations than they do, and the latter are those who only receive nominations but do not nominate other participants. However, other *motifs* exist but are not as relevant for the study in this network. These are: the *path*, when the nodes have the same number of *indegree* and *outdegree*, the *amplifiers* or *admirers*, nodes who receive fewer nominations than the nominations they send, the *source*, nodes who nominate various peers but are not nominated at all and, lastly, the *isolated*, nodes that have zero nominations and do not nominate anyone [20].

Additionally, other measures that are also of importance in this networks are *closeness centrality* and *betweenness* that help identify who are influencers and brokers inside the network. The influencers are those who will have a high *closeness centrality* which means that they can spread information or behaviors very efficiently through the network. The closeness centrality, measures the extent to which a teenager inside the network makes friendships with the other participants. The more central the student is, the closer it is to all the other subjects. The brokers are intermediaries between two students, in the way that a participant is friend with the broker and the broker can connect him or her to another friend. The *betweeness* is the measure that quantifies the level to which a participant is a broker. It quantifies how many times the student comes in the shortest friendship path between 2 other participants [20, 21].

In this study is also of importance to know according to the network, which is the community each of the nodes belongs to. The communities represent subgroups of nodes inside the network, where the participants who share common friends are grouped together. A perfect community is represented by a group of densely connected nodes, as the friends nominate each other and loose connection with other groups. To calculate the community to which each participant belongs, the Louvain Method for Community Detection is used. It extracts the communities by optimizing what is known as the modularity value. The modularity value measures the relative density of the friendship connections inside communities compared to the connections outside the communities. The Louvain method iterates the modularity until the highest value is found, as it is theoretically the best possible grouping of participants that are friends between them [20, 21].

If participants are more densely connected they can be analyzed by looking at the individual or local clustering coefficient. This is a measure that represents if the "friends of my friends are my friends", and it quantifies how close a participants neighbours are to being completely connected, in such a way that all are friends with each other [21]. Individuals with higher clustering coefficient will tend to receive a higher social pressure as the group is more homogeneous and closed up together [21].

3.6. Susceptibility Models

In order to assess students susceptibility to risky sexual behaviours, classification and regression models were used in order to identify the features that better predicted risky sexual behaviours in this population. However, before conducting this type of analysis feature selection techniques were used to reduce the number of variables that are used in this models. To reduce the number of variables Chi-Squared statistic and mutual information statistic were used.

3.6.1. Regression Models

In order to have a baseline model a multi-class logistic regression was used. This model is also used to evaluate the different predictors and the effect they have on sexual behaviour classification. In this case a Ridge regression is chosen to reduce the models complexity and avoid over-fitting that can result from simple linear regression.

3.6.2. Classification Models

The first type of susceptibility models tested in this paper are decision trees as they are used to classify observations into different groups, in this case sexual behaviour profiles found in Section 3.4. This model will be useful to define the characteristics that are most important when classifying the different subjects into groups. Meaning that will be used to determine what characteristics are the most important to predict sexual behaviours. Decision trees are also used as a flowchart structure here each of the decisions used to classify individuals in the sample are visualized which is useful in this case as it provides a clear interpretation of the characteristics used to find the optimal partition.

However, decision trees are a very unstable model as every time it is executed a different result will be achieved. This is why Random Forest was later used and analyzed. Even though, the interpretation is slightly more difficult of this model than the decision tree it is a accurate estimation of the variables that are mainly used for partitioning the participants in the different sexual behaviour classes.

4. RESULTS

4.1. Data Cleaning

The 14 questions used to group the different participants of the survey contained various missing values. Figure 2 outlines the missing values for each of the questions used and each subject in the survey. This was also useful to outline the pattern the missing values might have.

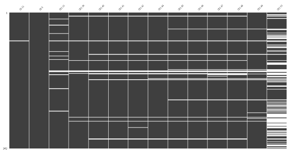


FIGURE 2: Missing values per questions.

There are some participant who have more than 50% of their answers as missing values, this is shown in Figure 3. According to some studies, 5% missingness has been suggested as a threshold to implement Multiple Imputation [15]. However, according to [15] for Missing At Random data, valid Multiple Imputation reduces bias even when the proportion of missingness is large. Due to the size of the sample, a threshold of 50% was chosen to minimize the number of dropped data. In this case 14 students did not reached the missing data threshold so they were dropped from this study. This can be detailed in Figure 3.

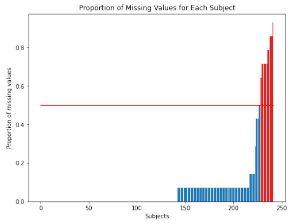
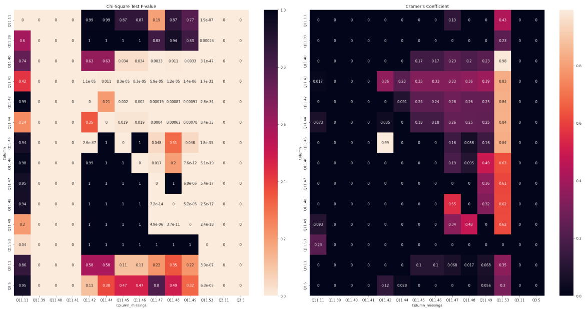


FIGURE 3: Missing values per subject after the initial data cleaning.

After conducting the previous elimination correlation test were estimated to statistically observed the association between missing values and the values of each other variables. The result of this test estimate that the majority of the variables had a correlation between them and their missing values. In Figure 4 the resulting p-value for the Chi-Square independence test is shown as well as the Cramer's coefficient for every combination of variables. It is noted that all the p-values lower than 0.05 represent dependency of the variables. Also, Cramer's coefficient as it quantifies the relationship between variables all the values that are over 0.25 represent a strong effect between them.



	Q11.11	Q11.39	Q11.40	Q11.41	Q11.42	Q11.44	Q11.45	Q11.46	Q11.47	Q11.48	Q11.49	Q11.53	Q3.11	Q3.5
Q11.11	0	0	0	0	0.99	0.99	0.87	0.87	0.19	0.87	0.77	1.9e-07	0	0
Q11.39	0.6	0	0	0	1	1	1	1	0.83	0.94	0.83	0.00024	0	0
Q11.40	0.74	0	0	0	0.63	0.63	0.034	0.034	0.0033	0.011	0.0033	3.1e-47	0	0
Q11.41	0.42	0	0	0	1.1e-05	0.011	8.3e-05	8.3e-05	5.9e-05	1.2e-05	1.4e-06	1.7e-31	0	0
Q11.42	0.99	0	0	0	0	0.21	0.002	0.002	0.00019	0.00087	0.00091	2.8e-34	0	0
Q11.44	0.24	0	0	0	0.35	0	0.019	0.019	0.0004	0.00062	0.00078	3.4e-35	0	0
Q11.45	0.94	0	0	0	2.6e-47	1	0	1	0.048	0.31	0.048	1.8e-33	0	0
Q11.46	0.98	0	0	0	0.99	1	1	0	0.017	0.2	7.6e-12	5.1e-19	0	0
Q11.47	0.95	0	0	0	1	1	1	1	0	1	6.8e-06	5.4e-17	0	0
Q11.48	0.94	0	0	0	1	1	1	1	7.2e-14	0	5.7e-05	2.5e-17	0	0
Q11.49	0.2	0	0	0	1	1	1	1	4.9e-06	3.7e-11	0	2.4e-18	0	0
Q11.53	0.04	0	0	0	1	1	1	1	1	1	1	0	0	0
Q3.11	0.86	0	0	0	0.58	0.58	0.11	0.11	0.22	0.35	0.22	3.9e-07	0	0
Q3.5	0.95	0	0	0	0.11	0.38	0.47	0.47	0.8	0.49	0.32	6.3e-05	0	0

(a) Chi-Square p-value for the independence test between variables taken into account.

	Q11.11	Q11.39	Q11.40	Q11.41	Q11.42	Q11.44	Q11.45	Q11.46	Q11.47	Q11.48	Q11.49	Q11.53	Q3.11	Q3.5
Q11.11	0	0	0	0	0	0	0	0	0.13	0	0	0.43	0	0
Q11.39	0	0	0	0	0	0	0	0	0	0	0	0.23	0	0
Q11.40	0	0	0	0	0	0	0.17	0.17	0.23	0.2	0.23	0.98	0	0
Q11.41	0.017	0	0	0	0.36	0.23	0.33	0.33	0.33	0.36	0.39	0.83	0	0
Q11.42	0	0	0	0	0	0.091	0.24	0.24	0.28	0.26	0.25	0.84	0	0
Q11.44	0.073	0	0	0	0.035	0	0.18	0.18	0.26	0.25	0.25	0.84	0	0
Q11.45	0	0	0	0	0.99	0	0	0	0.16	0.058	0.16	0.84	0	0
Q11.46	0	0	0	0	0	0	0	0	0.19	0.095	0.49	0.63	0	0
Q11.47	0	0	0	0	0	0	0	0	0	0	0.36	0.61	0	0
Q11.48	0	0	0	0	0	0	0	0	0.55	0	0.32	0.62	0	0
Q11.49	0.093	0	0	0	0	0	0	0	0.34	0.48	0	0.62	0	0
Q11.53	0.23	0	0	0	0	0	0	0	0	0	0	0	0	0
Q3.11	0	0	0	0	0	0	0.1	0.1	0.068	0.017	0.068	0.35	0	0
Q3.5	0	0	0	0	0.12	0.028	0	0	0	0	0.056	0.3	0	0

(b) Cramer's coefficient between the all variables taken into account for clustering.

FIGURE 4: Statistical results for the Chi-Squared independence and Cramer's test.

Using the Chi-Squared tests and the Cramer's coefficient as support of the missing at random behavior on the sample, it is possible to proceed to impute the data. Multivariate imputation by chained equations (MICE) was used in this case to impute the data [15] and the totality of the observations was filled to be later clusterized.

A similar process was carried out for the data used for the explanatory variables. Is was also found that the majority of missing values were statistically dependent to other variables so the missing values were imputed using the same technique reducing the proportion of missing values from 14.34% to 0.

4.2. Subject Classification

After exploring and cleaning the data from the survey, this section will provide the results of the subjects' clustering.

4.2.1. Factor Analysis

First, a factor analysis was conducted to simplify the data that will be used to group the participants of the survey. It is also used to find the underlying relationships each of the questions have with each other as with factor analysis a more accurate structure to represent the data can be found. This was motivated by the fact that the variables used for clusterization have a high correlation with one another as it is noted in Figure 5. To ensure the adequacy of the factor analysis in this particular case Barlett's test and KMO test were conducted. Barlett's test showed a p-value close to zero representing the test was statistically significant, and so indicating that the observed correlation matrix is not an identity matrix. This ultimately symbolizes that at least some of the variables are correlated with each other. The KMO test estimated a 0.84 coefficient indicating that this is the proportion of variance among all the observed variable. A value higher than 0.6 is considered in literature adequate for factor analysis. The results obtained in the last two test represent enough statistical proof to say that the data is adequate for factor analysis.

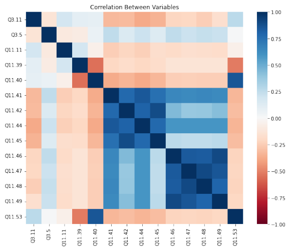


FIGURE 5: Correlation between the variables taken into account for clustering.

To determine the number of factor that would describe best the data a scree plot will be used. The goal is to achieve the most parsimonious (but still interpretatable) factor structure. To calculate a scree plot eigenvalues are necessary as they measure of the amount of variance accounted for by a factor, and so they can be useful in determining the number of factors that would be ideal. In a scree plot, the eigenvalues for all of the possible factors are plotted and where they drop off sharply the ideal number of factors is found. In Figure 6 the scree plot for the data used in this study is shown. It is clear that the optimal number of factors is 3.

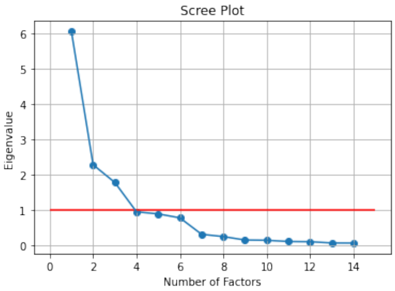


FIGURE 6: Scree plot to determine the number of factor necessary to describe best the data for clustering.

Measure	Factor 1	Factor 2	Factor 3
Cronhbach-alpha	0.42	0.84	0.63
SS Loadings	4.20	3.79	2.14
Portion Var	0.30	0.27	0.15
Cumulative Var	0.30	0.57	0.72

TABLE 1: Crohnbach-Alpha, loadings, portion of variance and cumulative variance explained by each of the factors found with factor analysis.

Additionally, Crohn-Bach Alpha tests were conducted to asses the adequacy of each of the factors. The results are shown in Table 1. According to [17], a Crohn-Bach Alpha value higher than 0.6 demonstrates internal consistency in the factor evaluated. Two of the three factor taken into consideration demonstrate high internal consistency and the three of them explain 72% of the variance in the data which in this case is enough to continue the data analysis with this three factors.

In Figure 7 the interpretation of each of the factors can be detailed. For Factor 1 is mainly represented by questions: Q11.46, Q11.47, Q11.48, Q11.49, which represent questions about sex in their current relationship, worry about sexually transmitted diseases, worry about unwanted pregnancies, and if they are actually using any type of method to avoid getting pregnant. This is why this factor is outlined as the representation of sexual health and save practices. Factor 2 is mainly outlines by questions: Q11.41, Q11.42, Q11.44, Q11.45 that represent the age at which the subjects had their first intercourse, the usage of any contraceptive method in their first intercourse, the usage of psychoactive substances before their first intercourse, and if they have had sex with their current partner. This factor is outlined mostly by the questions that describe the participants first intercourse. Lastly, factor 3 is mainly represented by questions: Q11.39, Q11.40, and Q11.53 that represent if subjects understand the term *sexual intercourse*, if subjects have had sexual intercourse before, and if they think they will have sex in the next year. This is why this factor is outlined by their sex knowledge or experiences. These will be taken into account to group the observations in the next section.

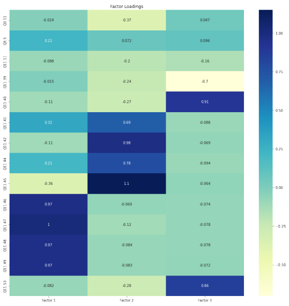


FIGURE 7: Factor loadings that represent which questions are the most representative for each of them.

Method	Silhouette	Calinski	Bouldin
OPTICS	0.45	69.14	1.42
Agglomerative	0.60	205.66	0.71
Mini K-means	0.60	205.66	0.71
DBSCAN	0.74	231.99	1.31
Espectral	0.76	299.61	0.43
Mean Shift	0.77	429.61	0.46
Gaussian	0.80	511.30	0.44
BIRCH	0.80	516.59	0.43
K-means	0.80	516.59	0.43
A. Propagation	0.82	1288.19	0.37

TABLE 2: Index value for each clustering methodology.

4.2.2. Clustering

As it was outlines in Section 3.4 8 different clustering methodologies were used to group the subject in clusters according to the 3 factors found in the previous section that represent sexual behaviour. They were evaluated using the Silhouette, Calinski Harabasz and Davies Bouldin. In Table 2 the results for each of the methods used in every index is presented.

According to the index estimated the best methodologies to cluster the data are the Gaussian Mixture Model, BIRCH, K-means and Affinity Propagation. However, as affinity propagation partitions the data in 8 different clusters and having that amount of groups would produce aggregations with very little observations and would later compromised the interpretation of the clusters it was not taken into account. This represent that the best grouping method would be K-means, BIRCH or Gaussian. As this three methodologies end up grouping the data in the same 5 clusters those are the ones used in further analysis.

Cluster 0 contains 22 individuals, that are predominantly male, are 13 or 16 years old, and are in a relationship with a girl and have had sexual intercourse before. They have had they first intercourse predominantly between the ages of 10 and 13 and they also admit to have not used any type of contraceptive to protect themselves and their partners from pregnancy the first time they had sex. They were also under the influence of alcohol or drugs during their sexual debut. However, the majority of people in this group have not had sexual intercourse with their current partner but believe they will in the next year even though they would not preferred it. This group is then represented as males that had their first intercourse very young, as comparing to the national average for rural areas which is 16 years old [22]. It is also important to note that this group seems to be represented by individuals that are in a relationship but have not had sex with their current partner and do not which to engage in sexual activity in the future even though they think they will. This could represent they feel pressured by their environment to engage in sexual activity with their current or future partners. This seems fitting with the [4] theory in which teenager feel pressured by their peers to have sex as to fit in with their peers and demonstrate love to their partners.

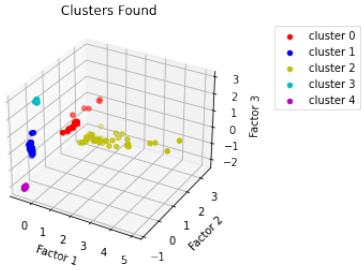


FIGURE 8: Clusters found according to the 3 factors chosen.

Cluster 1 is made up by 126 individuals, containing more than half the students participating in this study. This cluster is predominantly females, between the ages of 13 and 15. They are either in a relationship or have been in a relationship in the past. From this group the majority of the individuals have not had sexual intercourse and they do not expect to have sex within the next year. This group represent those who have note initiated their sexual life and are not interested in doing so in the next year.

Cluster 2 contains 42 subjects, the majority of them identifying themselves as male, being between 15 and 17 years old. They predominantly find themselves in a relationship. They also state that the first time they had sex they were 14 years old or older, used some type of method to protect themselves and their partners form an unwanted pregnancy and were not under the influence of any type of alcohol or drugs during this first sexual encounter. 90% of this group have already had sex with their current partner and feel their have a stronger bond with their partner because of it. This groups also manifested being worried about contracting any type of sexually transmitted disease and of having and unwanted pregnancy. This is why this group tends to use some kind of method to prevent an unwanted pregnancy. Lastly the majority of them state that they expect to have sex within the next year willingly. This cluster is then outlined by the males that have had already had sex and are practicing safe sex with their current partners.

Cluster 3 is made up by 13 individuals from which 51% are male and 49% are female. They also are either 15 or 16 years old and have a girlfriend or boyfriend. They, even though understand what sexual intercourse is, do not which to answer is the have had or not sex and also refuse to answer or do not know if they will have sex in the next year. This cluster is then made up by the individuals who could possibly find the survey to overwhelming or private.

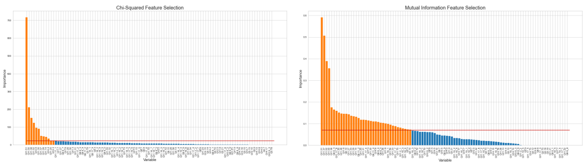
Lastly, cluster 4 has 25 participants that are between the ages 13 and 15. This group is also made up by 55% females, 48% males and the rest identify themselves as other gender. This subjects mostly have a boyfriend or do not which to disclose their relationship status. They also do not understand the term *sexual intercourse* so they are a the group of students that are younger and have zero knowledge about sexual intercourse and what it entails.

This clusters are the response variable that is used in Section 3.6 to estimate sexual behaviour profiles in teenagers.

4.3. Survey Variable Selection

The variables that are used to predict for each of the observations in which sexual classification cluster there are must be extracted form the variables that were carefully selected according to their topic in previous Section 3.3. However, the number of resulting variables may be too large to successfully estimate the groups as the number of variables may introduce to much noise to the model. This is why Chi-Squared and mutual information feature selection were implemented to select the best questions that describe the clusters and to be used to estimate the susceptibility models in Section 3.6.

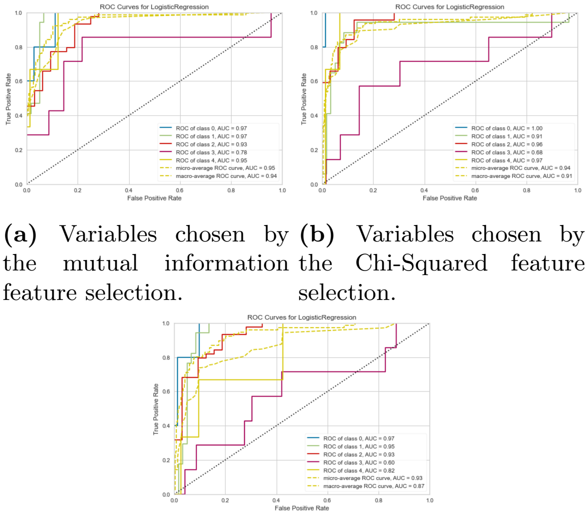
To select the number of variables according to their importance for the susceptibility models the variables that had a value equal or higher to the mean importance were selected. In this case the mean importance was used as a threshold as it seem to produce the most consistent results. In Figure 9 the variable importance according to the both method used are shown. From the figures shown the number of variables selected with Chi-Squared are 12 and with the mutual information methodology is 37.



(a) Chi-Square feature selection methodology. **(b)** Mutual information feature selection.

FIGURE 9: Variable importance for the feature selections method tested.

To correctly select the number of variables that should be chosen simple multi-class logistic regressions were estimated with all the possible variables, the variables selected with the Chi-Squared feature selection and with the variables selected with the mutual information feature selection. In Figure 10 the ROC curve and AUC are shown for the different models tested. Additionally and accuracy test shown that the models containing the complete, Chi-Squared and mutual information variables had an accuracy of 0.75, 0.83, and 0.76 respectively revealing that the best number of variables to select is 12. This variables are detailed in Appendix B. These 12 variables will be the ones included in the susceptibility models in Section 3.6.



(a) Variables chosen by the mutual information feature selection. **(b)** Variables chosen by the Chi-Squared feature selection.

(c) All possible variables.

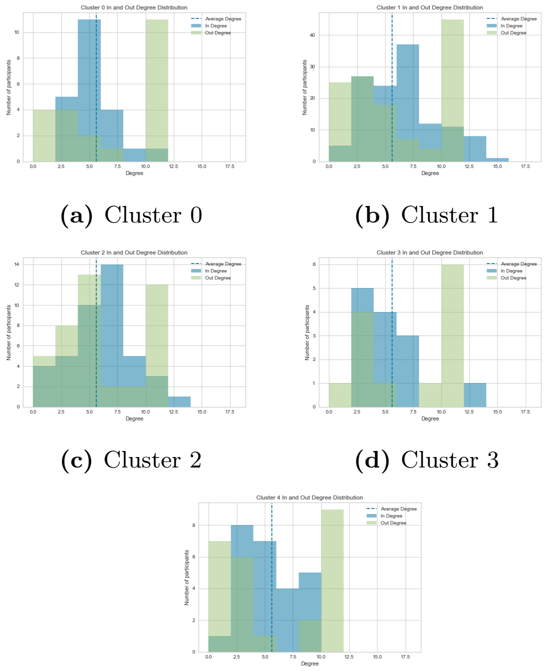
FIGURE 10: ROC curves for the different models tested according to their nomber of variables.

4.4. Network Variables

The friendship school network will be characterized, and analyzed in order to determine how the school social structure is subdivided and to compare the social behavior of the different clusters found in Section 4.2. The network includes the 228 participants and it has 1217 connections, with a density or cohesion of 2%, which means that 1 of 50 possible friendship nominations is made. Additionally, from the relationships outlined by the network only 42% of the the friend nominations were reciprocal. As seen, not all the participants who sent a nomination to someone who they considered his or her friend, got a nomination back from that person. Besides, if 2 people knew the same person it was 36% likely that they were also friends with each other [20]. There are also 23 strongly connected components and 1 weakly connected component within the network.

4.4.1. Degrees and Motifs

To analyze the different type of students and their social behavior at the school, participants were segmented into the different clusters that were found in Section 4.2. This with the main objective to get insights of how the subjects in each of the clusters socialize before entering this as parameters in the susceptibility models.



(a) Cluster 0 **(b)** Cluster 1

(c) Cluster 2 **(d)** Cluster 3

(e) Cluster 4

FIGURE 11: Histograms for the in and out degrees of all the different clusters.

Figure 11 shows the in and out degree distribution for the different sexual behaviour clusters. From this figure a clear difference in friendship structure could be noticed for each cluster. Cluster 0, containing mainly individuals with a very premature start to their sexual life, showed a smaller in degree (lower than the 7.4 mean in degree of the total network) representing they tend to be considered as a friend by very few, while they showed a higher out degree in general, representing they consider many more peers as friends as the number of peers that consider them their friends, making them mostly admirers in the social network. In contrast, Cluster 1 has a broader in degree distribution showing the majority of the subjects have a higher in degree and lower out degree, representing the majority of subjects in this cluster are bottle necks within the network. This shows this cluster, that is represented mainly by female students that have not have had their sexual debut is mainly made up of the popular crowd in the school as they receive more public recognition. Similarly Cluster 2, that is represented by male students who practice safe sex and are in a relationship currently, is also mainly made up of the popular crowd in the school as they receive more public recognition. As [23] states, male teenagers tend to be rewarded in terms of popularity for having sex as their peers. It is common to find that the more sexually experienced teenager may be the role model and most popular in a social network. For teen women, the complete opposite is held accountable as ”girls tend to be labeled and stigmatized and are often blamed for sexual encounters that result in sexually transmitted infection or pregnancy” [23].

Cluster 3 and 4, which are represented by the participants that have not had sex or do not understand what sex means, have a very similar behaviour. They tend to have a low in degree and a high out degree, mostly representing participants that are admirers within the network. The aforementioned conclusions can be observed in detail in Figure 12, where the y axis represents the percentage of students in each cluster and the x axis outlines the motif in which they are categorized. The bars different colors represent the 5 different clusters taken into account. Clusters 1 and 2 showed similar proportions for every motif just as clusters 3 and 4. In this case isolated participants are not found and very little participants are considered sources. On the other hand, clusters 1 and 2 showed the higher percentage of bottle necks (popular students), showing they have some social positioning and recognition within the network. Cluster 0, also shows a higher percentage of a category, this case of sinks, representing popularity as they are recognized within their community it also represents detachment as they do not recognize any of their peers as their friends. This is supported with studies from [23], where early sexual debut in male is linked to lower global self-worth and so difficulty in forming long lasting relationships with their peers.

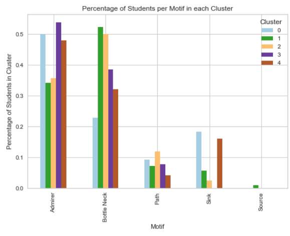


FIGURE 12: Percentage of students in each cluster per motif.

4.4.2. Communities

As mentioned before the density of the school network was low with a value of 2% of participants over the total number of possible connections. One of the main reasons why the network might not be so densely connected is because it is divided in multiple subgroups or communities.

Using the Louvain Method, 8 communities were detected with a maximum local clustering coefficient value of 0.75, which is high considering that the maximum local clustering coefficient is 1. This represents that the participants are closely connected to one another. Therefore, strong communities structures within the school network having high density within can be found. It is important to note that the community detection is influenced by the school grades, as participants tend to associate mostly with peers form their same grade. However, 8 clusters were found while there were 6 different school grades, from 6th grade to 11th grade. In Figure 13 the different communities can be detailed by their different colors.

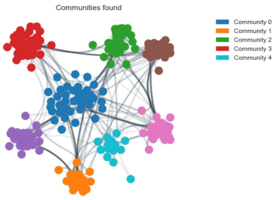


FIGURE 13: Communities found for the schools network.

Since there are more communities than the existing school grades this represents that the communities found are made up of students within the all the grades. In general terms, sixth grade students were mainly grouped in community 1, however some of of them were assigned to communities 0 and 6. At the same time, community 0 contains mainly students form the seventh grade and community 6 contains participants coursing mainly the ninth grade representing that this portion of the students relate more with peers that are not necessarily in their same grade or their same age. Eighth grades are mainly represented by community 4 and ninth graders are equally distributed in communities 2 and 6. Tenth graders were almost equally distributed in communities 1 and 7 and eleventh grades were represented by community number 5.

In Table 3 the distribution of each community can be outlined. The largest communities are 0, 2 and 3. It is also important to note that, communities tend to have a high percentage of participants that belong to cluster 1, which are those individuals that have not had yet their sexual debut and do not expect it to happen in the next year. Communities 1, 3 and 4 also have a high percentage of students that are classified in cluster 2, which is made up by students that have already had their sexual debut and generally speaking practice sex safely. Communities 2 and 7 have high percentages of students classified in cluster 1, which are those subjects that admitted having their sexual debut at a young age and under the influence of psychoactive substances. This can be related to the fact that communities 2 an 7 have a high percentage of students reporting their friends have already been intimate with someone else and had sexual relations. This last could represent that there might be a social influence or/and assortativity that causes cohesion between sexual behaviours.

Cluster / Community	0	1	2	3	4	Total (%)
0	0.08	0.58	0.13	0.00	0.23	0.18
1	0.09	0.35	0.30	0.07	0.17	0.10
2	0.15	0.59	0.12	0.03	0.12	0.15
3	0.10	0.46	0.26	0.10	0.08	0.17
4	0.07	0.60	0.27	0.07	0.00	0.07
5	0.04	0.65	0.19	0.00	0.12	0.11
6	0.04	0.74	0.17	0.00	0.04	0.10
7	0.18	0.50	0.11	0.18	0.04	0.12

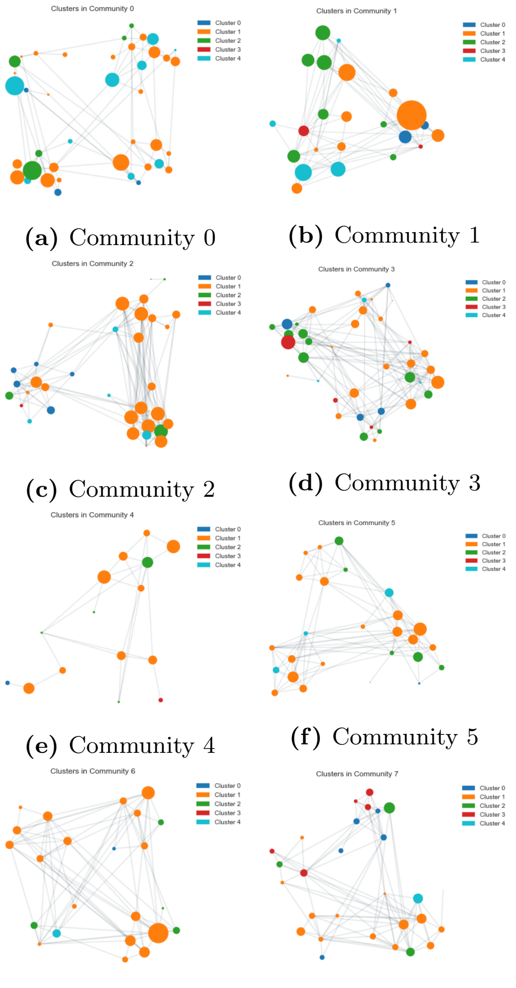
TABLE 3: Communities distributions

As it was mentioned previously, the communities initially found seem to be heavily influenced by the grade the participants are in. This is why for the sub network produced by each of the 8 communities found the louvain algorithm was executed to find the communities within them. In Figure 14 the communities found for each of the initial communities. They are also compared to the sexual behaviour clusters as they are represented by their color and to the popularity of each node as the size of each node is the in degree, the higher its number of nominations received the larger it would be.

Community 0 is the largest group, with a total of 40 participants with an average age of 14 years old, with the majority of teenagers coursing the seventh or sixth grade or sixth. Its members have on average 6 nominations received and 4.5 made. 37% of its members are classified as admirers, 33% as bottle necks and 23% as sinks, showing their popularity. In this community 58% of the participants have not been sexually active and do not think they will be in the next year. Additionally, 4 different communities could be found within were the most popular belonged to clusters 1, 2 or 4.

Community 1 and 6 have the same number of participants, with a total of 23 students with an average age of 15 years old, with the majority of teenagers coursing the tenth or eleventh grade. Its members have on average 5 nominations received and 5 made. While community 1 tends to have more than 57% of their participants as bottle necks, community number 6 has 40% of its members are classified as admirers, showing the majority of the participants in community 1 are considered popular. In community 1 35% of the participants have not been sexually active and do not think they will be in the next year while in community 6 50% of them state the same thing. Additionally, while in community 1 3 different communities were found with a very clear popular node classified in cluster 1, community number 6 had 4 different communities were the subjects shared almost the same degree of popularity and had were classified in the same cluster.

Community 2 has a total of 23 participants with



(a) Community 0 **(b)** Community 1

(c) Community 2 **(d)** Community 3

(e) Community 4 **(f)** Community 5

(g) Community 6 **(h)** Community 7

FIGURE 14: Communities found in the sub-networks created by each of the communities initially found. They are also compared to the clusters and the in degree of each node.

an average age of 14 years old, with the majority of teenagers coursing the eighth grade. Its members have on average 5 nominations received and 5 made. 50% of its members are classified as bottle necks and 33% as paths. In this community 59% of the participants have not been sexually active and do not think they will be in the next year. Additionally, 3 different communities could be found within were the students seem to have almost the same level of popularity.

Continuing with community 3, it is the second largest community, with a total of 39 students that are mostly in eleventh grade. Moreover, 35% of them declared already having had their sexual debut and 36% of them are currently in a relationship. Participants in this community tend to have a low in degree around 4 and a higher out degree of 6 on average. 46% of its members are classified as bottle necks and 44% as admirers and 4 different communities were found within. The majority of them having a subject in clusters 1 or 3 more popular than the others.

Community 4 is the smallest community with only 15 participants that identify themselves around the age of 14 and are coursing the eighth grade. 40% of them are bottle necks and 37% admirers. 3 different communities were identified within the community and the majority of them belonged to cluster 1.

Lastly, communities 5 and 7 had 26 and 28 students, respectively and their participants were mostly from the ninth and tenth grade and were between 14 and 15 years old. In community number 5 3 different communities were found while in community 7, 4 of them were discovered.

4.4.3. Influencers and Brokers

Analyzing the students that are central within the network meaning they can spread information or behaviors efficiently thought the network can be useful to implement policies about sexual health and general knowledge. To do this the the *clossness centrality was used.* It is observed that the top 5 influencers are 3 males and 2 females. The males are 14, 15 and 17 while the female are 16 and 15 years old. One of the males is in sixth grade while the rest are in tenth or eleventh grade. Females are in the tenth or eleventh grade. Two of the males, the oldest ones, do not understand what sex is and have girlfriends. The remaining male and female have been relationship but are not one currently and do understand what sex means, however they stated they have not been sexually active yet. They all represent different communities within the network but belong to mostly the same cluster, number 4. They have an average in degree of 8 nominations received and of 10 nominations made. All these subjects from the top tier are mostly admirers.

Looking at the brokers or those participants who can connect a friend to another friend they got from the network, there were detected that over the top 5 brokers, none of them has had sex and only two of them understand what sex is. Additionally, 60% are male and only two of them share grade, being eleventh. The rest of them are in different grades except seventh and eighth. They all represent different clusters and communities within the network. They mostly have bottle neck structures and are considered popular within the school.

4.4.4. Clustering Coefficient

Determining those participants who are more densely connected in the school, in other words, the students that are more likely to be connected with their friends friends could be helpful to better understand the school's social structure. The top 5 more densely connected participants, were bottle necks who are mostly females. Furthermore, 3 out of the 5 of them had already their sexual debut, belonging to clusters 0, 1 and 3. They are also in ninth, tenth and eleventh grade and are mostly 14 years old. Deeper analysis into how being densely connected with friends could impact sexual behaviour will be tested in the predictive models.

4.5. Predictive Models

Having the social metrics parameters that characterize how are the participants behave in their network, tpredictive models were run to determine the sexual behaviour profiles and identify which of those attributes and individual characteristics are the most accountable for their behaviour.

4.5.1. Decision Tree

A classification tree was run to segment the susceptible profiles among students. 25% of the 228 participants were used as the test set and with them a simple classification tree was executed. The model was run under 10-fold cross-validation to find the depth that produced the best results and the accuracy was tested using the ROC curve, Figure 15 and the AUC measure was 80.70%.

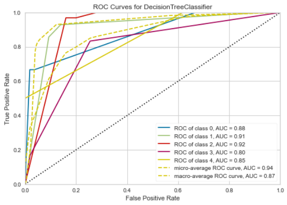


FIGURE 15: ROC curve for the classification tree model.

The decision tree structure obtained can be detailed in Figure 16 that had an optimal depth of 3. Additionally the most important variables can be extracted from this model and can be detailed in **??**. This variables are questions 11.51, 11.36, 11.37, 11.25, 11.26, and the in degree of each node. This variables represent the reason why each subject is not using contraception, if they have already touched someone else's private parts, at what age did they do it, if they ever spend time in private with someone they loved and at what age they did it. Additionally, the in degree represents the number of people that nominated them as their friends.

However, after running this algorithm various times the results were not stable enough to consider this an acceptable model this is why Random Forest models were also implemented.

4.5.2. Random Forest

A random tree was run also to segment the susceptible profiles among students and obtained the variable importance. Just as in the previous section 25% of the 228 participants were used as the test set. The best results and the accuracy was tested using the ROC curve, Figure 15 and the AUC measure was 84.21%

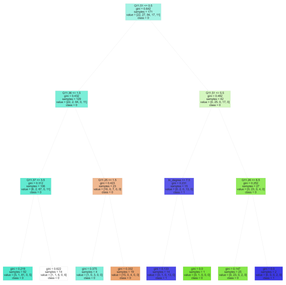


FIGURE 16: Structure for the classification tree model.

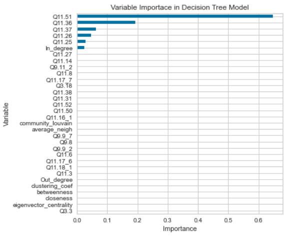


FIGURE 17: Classification tree model variable importance.

The most important variables can be extracted from this model and can be detailed in 19. This variables are questions 11.51, 11.37, 11.50, 11.52, 11.38, betweenness and the closeness. This variables represent the reason why each subject is not using contraception, if they have already touched someone else's private parts, at what age did they do it and what method are they using to avoid contraception. Additionally, the betweenness represents the amount of influence the node has over the flow of information in the school, and closeness represent how close the node is to others.

4.5.3. Lasso Model

As the target variable was the sexual behaviour cluster, multi-nomial regression models can be also useful. Due to the kind of desired analysis, models with the capacity to show the influence of the different attributes (predictors) on the target variable were ideal.

Ridge Regression adds a L2 penalty to the logistic model. This model will shrink the effect of the features, but will still include all the variables that are input. The results of the stronger positive and negative influences can be seen in 4. ROC curve was plotted as well (see Fig 20) AUC (Area Under Curve) for this plot is 80.70%.

5. DISCUSSION

The random forest model outlines the most important variables and predicts with 84.21% of accuracy the sexual behaviour profile the subjects are classified in.

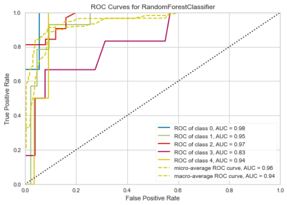


FIGURE 18: ROC curve for the random forest model.

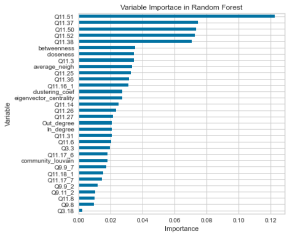


FIGURE 19: Random forest model variable importance.

According to this model the variables that are most important to classify the subjects into sexual behaviour clusters are:

- If subjects have already touched someone else's private parts and at what age they did.
- What method are subjects using to prevent pregnancy.
- What are subjects reasons to not use any kind of contraceptive methods.
- How close are subjects to their peers.
- How influential a subject is in the network.

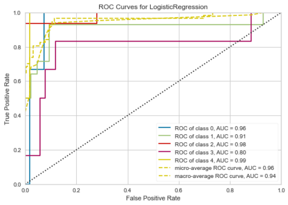


FIGURE 20: ROC curve for the ridge model.

This variables complemented with the results of the ridge regression model can be interpreted to know if they affect positively or negatively the sexual behaviour classification found. Negative effects are recorded for the question variables that are included in the model, while positive effects are found for the network variables. For example, if subjects have already touched someone else's private parts and at what age they did tend to have a negative effect in the classification, representing that the younger the subjects are the more likely they will The younger the subjects are when touching for the first time some else's private parts the higher they are at risk to be classified in higher clusters. The same effect negative can be appreciated with the method that participants are using to prevent pregnancy. In this case the lower coded variables were embarrassment, pregnancy prevention, partner preferences and pleasure interference, while higher rated answers were because they did not know were to obtain contraceptives, do not wanting to seem an expert in the subject, subjects don not believe they can get pregnant, can not afford contraception and have never thought about it. In a similar way methods that are lower rated are condoms, contraceptive pills, contraceptive injection, feminine condom, the rhythm method, interrupted coitus, emergency contraceptive pill, IUD, and implants. Also the closer subjects are with they peers the higher cluster rating they will have and so the more influential over the network they are the higher cluster they will also be.

Variable	0	1	2	3	4
Closeness	0.07	0.11	-0.07	-0.04	-0.07
Betweenness	0.05	0.02	-0.07	-0.01	0.00
Q11.38	-0.01	-0.01	-0.22	-0.18	0.41
Q11.52	-0.02	-0.33	-0.01	0.38	-0.02
Q11.50	-0.02	0.42	-0.00	-0.38	-0.01
Q11.37	-1.02	0.21	-0.18	0.22	0.77
Q11.51	-0.84	1.40	-1.44	1.27	-0.40

TABLE 4: Strongest influences Ridge Regression

In this study the desire would be that subjects belonged to the sexual behaviour cluster 1 or 2 as this two groups tend to either do not have started their sexual life or if they have the majority of them have done it safely and willingly. This is why it would be desirable that the participants had a combination of the variables previously mentioned that kept their cluster on the lower spectrum. In this case it is desired that the participants touched for the first time someone else's private parts when they are older, and use more robust contraceptive methods such as condoms or pills. It is also desired that the they are more influential in their friendship network and they have a close knit group of friends to fall in this two categories that are considered healthier. This results are coherent with the different literature recorded in the topic as having a premature sexual debut can lead to higher rates of sexually transmitted disease infection, emotional detachment, unwanted pregnancies and lower self-worth [24]. Also being more influential in their group of friends represent that they are not as easily influenced by others [25] and so having a close knit group of friend also contributed to having a healthier sexual behaviour profile as having positive friendship relationships have been linked to happier, more confident and better adjusted individuals to their environment [26].

6. CONCLUSION

In this particular case, for teenager in Santa Ana, Barú , a rural area in Colombia, sexual behaviour clustering is associated with:

- If subjects have already touched someone else's private parts and at what age they did.
- What method are subjects using to prevent pregnancy.
- What are subjects reasons to not use any kind of contraceptive methods.
- How close are subjects to their peers.
- How influential a subject is in the network.

among other variables. After this study it is recommended to the decision-makers within the school to improve sexual education policies mostly oriented to the more influential subjects within the network, as the majority of them seem to do not know what sexual intercourse is and so can spread more easily false information on the topic. It is also recommended that the subjects keep positive and tight knit relationships with their friends as that makes them more confident and overall satisfied beings which in turns tends to increase sexual debut age to when they are physically and mentally ready [25].

This study was only a preliminary analysis of the subject within this Santa Ana's particular community, however future work should be considered as to strengthen the weak points found in the paper. In the first place, some possible answers for each of the questions in the survey should be reconsidered to avoid bias and noise in the results. For example, having options such as ”I do not know” or ”I do not want to answer” can be joined in only response. It is also suggested that to avoid large amounts of missing data, the survey should be administered by parts or in a less consuming way. A weighted holistic network can be more useful for analysis (including external peers). To do so, questions such as how long have you been friends with each other or rating the importance of the relation. It is also considered that a multi layer network analysis could be carried on in order to include adverse relations and measure the impact that those have on the sexual behaviour profiles. It would be also interesting to include other type of variables in this analysis such as gender perceptions, physical activity and racial differences. On the other hand, this subjects tend to be a taboo topic within teenagers this is why methods to measure the reliability of the surveys' answers should implemented. This study can be used as a baseline for other projects considering other type of variables and behaviours within this sample population.

REFERENCES

[1] Turchik, J. A. Identification of sexual risk behaviors among college students: A new measure of sexual risk.

[2] Nicholson, J., Marcum, C. D., and Higgins, G. E. Predictors of risky sexual behavior among high school students in the united states. *Deviant Behavior*, **41**, 1052–1064.

[3] Nwagwu, W. E. Social networking, identity and sexual behaviour of undergraduate students in nigerian universities. *Electronic Library*, **35**, 534–558.

[4] Treboux, D. and Busch-Rossnagel, N. A. Network influences on adolescent sexual attitudes and behaviours. *Journal of Adolescent Research*, **5**, 175–189.

[5] Asrese, K. and Mekonnen, A. Social network correlates of risky sexual behavior among adolescents in Bahir Dar and Mecha Districts, North West Ethiopia: an institution-based study. *Reproductive health*, **15**, 61.

[6] Nicholson, J., Marcum, C. D., and Higgins, G. E. Predictors of Risky Sexual Behavior among High School Students in the United States. *Deviant Behavior*, **41**, 1052–1064.

[7] Humberstone, E. Friendship networks and adolescent pregnancy: Examining the potential stigmatization of pregnant teens. *Network Science*, **7**, 523–540.

[8] Durowade, K. A., et al. Early sexual debut: Prevalence and risk factors among secondary school students in ido-ekiti, ekiti state, south-west nigeria. *African Health Sciences*, **17**, 614–622.

[9] Haynie, D., Doogan, N., and Soller, B. Gender, friendship networks, and delinquency: A dynamic network approach. *Criminology*, **52**.

[10] Woodward, D., Drager, N., Beaglehole, R., and Lipson, D. Globalization and health: A framework for analysis and action. *Bulletin of the World Health Organization*, **79**, 875–81.

[11] Jeon, K. C. and Goodson, P. Alcohol and sex: friendship networks and co-occurring risky health behaviours of us adolescents. *International Journal of Adolescence and Youth*, **21**, 499–512.

[12] Maness, S. Maness, s. b., buhi, e. r. (2016). associations between social determinants of health and pregnancy among young people: A systematic review. public health reports 131(1), 86-99. *Public Health Reports*, **131**, 86–99.

[13] Manlove, J., Logan, C., Moore, K., and Ikramullah, E. Pathways from family religiosity to adolescent sexual activity and contraceptive use. *Perspectives on sexual and reproductive health*, **40**, 105–17.

[14] McHugh, M. The chi-square test of independence. *Biochemia medica*, **23**, 143–9.

[15] Madley-Dowd, P., Hughes, R., Tilling, K., and Heron, J. The proportion of missing data should not be used to guide decisions on multiple imputation. *Journal of Clinical Epidemiology*, **110**, 63–73.

[16] Gorman, B. S. and Primavera, L. H. The complementary use of cluster and factor analysis methods. *The Journal of Experimental Education*, **51**, 165–168.

[17] Taber, K. S. The use of cronbachs alpha when developing and reporting research instruments in science education. *Research in Science Education*, **48**, 1273–1296.

[18] Martn, J., Luna-Romera, J. M., Pontes, B., and Riquelme, J. Indexes to Find the Optimal Number of Clusters in a Hierarchical Clustering.

[19] Adegboyega, L. O., Ayoola, V. A., and Muhammed, S. Influence of peer pressure on sexual behavior of undergraduates in kwara state. *Anatolian Journal of Education*, **4**.

[20] Diego, J., Contreras, A., Muoz, N., Advisors, V., Jimenez, F. M., Mara, A., and Bernal, G. Individual attributes and social network predictors of alcohol

consumption susceptibility among adolescents in rural area of colombia: Santa ana, baru.

[21] Hansen, D. L., Shneiderman, B., Smith, M. A., and Himelboim, I. Social network analysis: Measuring, mapping, and modeling collections of connections.

[22] Morales, A., et al. Sexual risk among colombian adolescents: Knowledge, attitudes, normative beliefs, perceived control, intention, and sexual behavior. *BMC Public Health*, **18**.

[23] Cuffee, J. J., Hallfors, D. D., and Waller, M. W. Racial and gender differences in adolescent sexual attitudes and longitudinal associations with coital debut.

[24] Golden, R. L., Furman, W., and Collibee, C. The risks and rewards of sexual debut. *Developmental Psychology*, **52**, 1913–1925.

[25] Widman, L., Choukas-Bradley, S., Helms, S. W., and Prinstein, M. J. Adolescent susceptibility to peer influence in sexual situations. *Journal of Adolescent Health*, **58**, 323–329.

[26] LESKO, N. Denaturalizing adolescence: The politics of contemporary representations. *Youth & Society*, **28**, 139–161.

[27] Billy, J. O. G., Landale, N. S., Grady, W. R., and Zimmerle, D. M. Effects of sexual activity on adolescent social and psychological development.

[28] Asrese, K. and Mekonnen, A. Social network correlates of risky sexual behavior among adolescents in bahir dar and mecha districts, north west ethiopia: An institution-based study. *Reproductive Health*, **15**.

[29] Fenton, K. A., Johnson, A. M., McManus, S., and Erens, B. Measuring sexual behaviour: Methodological challenges in survey research. *Sexually Transmitted Infections*, **77**, 84–92.

[30] Mercer, C. H. Measuring sexual behaviour and risk.

[31] Coley, R. L., Lombardi, C. M., Lynch, A. D., Mahalik, J. R., and Sims, J. Sexual partner accumulation from adolescence through early adulthood: the role of family, peer, and school social norms. *The Journal of adolescent health : official publication of the Society for Adolescent Medicine*, **53**, 91–92.

[32] Graham, C. A., Sanders, S. A., Milhausen, R. R., and McBride, K. R. Turning on and turning off: a focus group study of the factors that affect women's sexual arousal. *Archives of sexual behavior*, **33**, 527–538.

[33] Patton, G. C., Coffey, C., Romaniuk, H., Mackinnon, A., Carlin, J. B., Degenhardt, L., Olsson, C. A., and Moran, P. The prognosis of common mental disorders in adolescents: a 14-year prospective cohort study. *Lancet (London, England)*, **383**, 1404–1411.

APPENDIX A. CLASSIFICATION QUESTIONS

1. Q3.11: Do you identify as female or male?
2. Q3.5: How old are you?
3. Q11.11: Today are you in a relationship or have been in any? Please specify the gender of your partner.
4. Q11.39: Do you understand what we are referring to when we say *sexual relations*?
5. Q11.40: Have you had sexual relations? With this, we mean when a man or boy puts his penis in the vagina of a woman or girl.
6. Q11.41: How old were you when you had your first intercourse?
7. Q11.42: The first time you had sex, did you or your partner do anything to prevent pregnancy?
8. Q11.44: The first time you had sex, was it under the influence of alcohol or drugs?
9. Q11.45: Have you had sex with your partner?
10. Q11.46: Do you feel that having sex with your partner allows you to have a closer relationship between the two?
11. Q11.47: Some people worry about sexually transmitted infections. How concerned were\are you about getting a sexually transmitted disease by your partner?
12. Q11.48: Some people are worried about pregnancy. How concerned were\are you about getting pregnant or your couple getting pregnant?
13. Q11.49: Are you or your partner doing something to avoid pregnancy?
14. Q11.53: Do you think you are going to have sex with someone in the next year?

APPENDIX B. EXPLANATORY QUESTIONS

1. Q11.14: How old is your partner?
2. Q11.25: Have you ever spent time alone with someone you love in a private space without adults nearby?
3. Q11.26: How old were you the first time you were alone with someone you were in love in a private area without adults nearby?
4. Q11.35: How old were you the first time you sent a sexual photo?
5. Q11.37: Have you ever touched the private parts of another boy or girl, or your private parts have been touched by someone? With touching private parts of people we mean genitals, breasts, or other parts of the body in a sexual way.
6. Q11.38: How old were you the first time your private parts (genitals) were touched in a sexual way or were touched by someone?
7. Q11.50: What are you doing to avoid pregnancy?
8. Q11.51: Young people have different reasons for not using contraception (birth control).State the reasons that best describe why you and your partner are not currently using contraception. Select all that apply. (Females)
9. Q11.52: Young people have different reasons for not using contraception (birth control). State the reasons that best describe why you and your partner are not currently using contraception. Select all that apply. (Males)
10. Q14.3: How old were you when you drank alcohol for the first time?

11. Q14.4: In your life, how often have you been drunk?
12. Q14.6: . How many cigarettes do you smoke usually during a regular week?