

A study on the current situation and causes of smoking among college students

Guang Deng^{1, *, #}, HuiBing Yan^{1, #}, Esther Chen², Bowen Huang², HengZhang²

¹School of Management, Xiamen University Tan Kah Kee College, Zhangzhou, China

²School of Accounting and Finance, Xiamen University Tan Kah Kee College, Zhangzhou, China

*Corresponding author:750149063@qq.com

#These authors contributed equally

Abstract. The total number of smokers worldwide is now more than one billion, accounting for about a quarter of the world's population, and tobacco use causes more than five million deaths worldwide each year. While smoking brings people pleasure, the toxic substances produced by the ignition of cigarettes can cause a variety of diseases. The data was collected from each school through a questionnaire and distributed online. For the questionnaire data, this paper uses C5.0 decision tree, chi-square test, logistic regression, and structural equation modeling to investigate the basic situation, causes, and influencing factors of smoking among college students, and to provide useful reference for improving the smoking situation among college students. In this paper, Fujian Medical University, Sunshine College, Fujian Information Vocational and Technical College, Fujian Arts Vocational College, Fujian Early Childhood Teacher's College, Jimei University, Chengyi College of Jimei University, and Jiageng College of Xiamen University were selected as the overall sample.

Keywords: College students' smoking; Influencing factors; C5.0 decision tree; Logistic regression.

1. Introduction

Combined with the relatively tobacco-controlled environment and the fact that smoking is hazardous to health, there is public controversy about smoking among college students out of concern for their physical and mental health development [1]. However, in their specific lives, most college students have a tacit and respectful attitude toward the smoking behavior of those around them and even of themselves [2]. Therefore, the real opinion of college students about smoking has aroused curiosity and concern among the public and universities, and whether there is a paradox between it and public opinion needs to be investigated and studied [3]. In order to promote the positive and healthy development of college students and to help the college smoking group to control smoking effectively, it is necessary to analyze the source in order to take targeted measures to help them quit smoking [4]. Therefore, the investigation of college students' smoking status and perceptions of smoking, and the investigation of factors affecting smoking are of great importance to society and universities.

As mentioned above, smoking is a common phenomenon among college students, and their behaviors may be related to their mood states and attitudes toward tobacco; in addition, there are regional differences in smoking rates, and socio-demographic, social environment, and personal characteristics are the main factors influencing their smoking, while parents and family and friends may also have indirect effects on smoking. Since the range of college students nationwide is too broad, a representative sample of college students in Fujian Province was selected as the survey population so as to understand the current status of smoking and their perceptions of smoking among college students, and to further investigate their reasons for smoking and the perceived influencing factors. Conclusions were found from the survey results as a way to provide effective help for college students' tobacco control [5-7].

2. Smoking population distribution characteristics

2.1. Analysis of smoking status data based on Baidu index and Jitterbug index

Baidu Index is a data sharing platform based on the analysis of Baidu's massive Internet users' behavior data. The project team aims to study the keyword focus trends, gain insight into the changing needs of Internet users, and locate the characteristics of digital consumers through the data analysis of Baidu index [8].

1. Analysis of the smoking phenomenon based on the "smoking" search index

The "smoking" search index refers to the number of times smoking is used in the neighborhood, which generally represents the user's attention, but not the actual number of searches for "smoking". The search was conducted with "smoking" as the keyword.

The data from February 16 to 25, 2015 were selected, and the search index for "smoking" was distributed according to the search index in Figure 3-1-1: the search index for "smoking" before July 2016 was at a high level. The "smoking" search index is at a high level; the "smoking" search index is at a low level from July 2016 to August 2019; the "smoking" search index is at a high level from August 2019 to February 2022. The high level of the "smoking" search index between August 2019 and February 2022 indicates that the smoking phenomenon in the past two to three years has been a major problem. This indicates that smoking has rebounded in the last two to three years compared to 2016-2019, as shown in Figure 1.

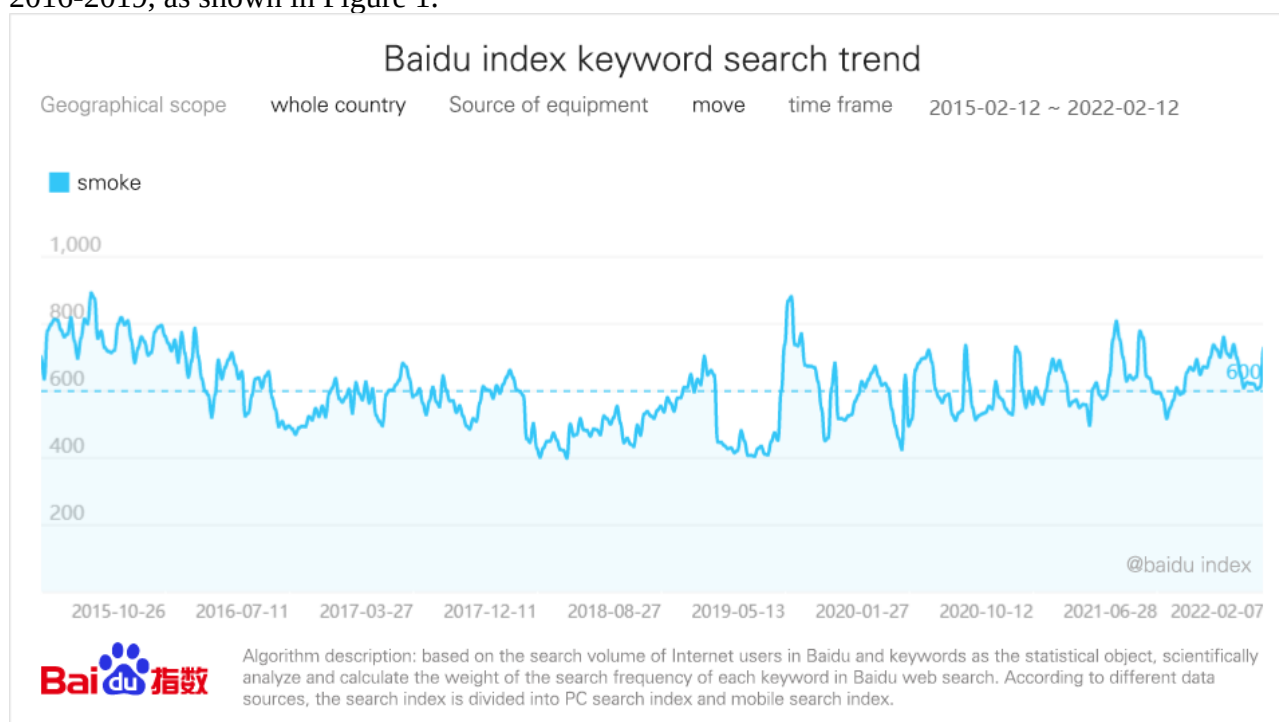


Figure 1 Baidu search index

In the face of the fact that the smoking rate is increasing, out of consideration for people's lives and health and the awareness of civilized social management, coupled with the effectiveness of the original tobacco control is not significant, in order to call on the whole society to pay attention to health, support tobacco control, and advocate the public to develop a healthy and civilized "smoke-free lifestyle." on June 1, 2015, the state issued a local regulations, the Beijing Smoking Control Regulations[9], which became the strictest tobacco control order in history, followed by comprehensive tobacco control in indoor public places in Shenzhen, Guangzhou, and Shanghai, and local smoke-free environment laws and regulations in eighteen cities by the end of 2016. In addition, the new advertising law will be disguised public service announcements by tobacco companies to stop, the country from the overall general environment of the public smoking behavior control, in order to get effective results. After the policy was implemented, the International Tobacco Control

Policy Evaluation Program China Survey Round 1 to Round 5 reported that the quit rate of urban smokers in China increased from 6% to 9.2%. It is due to the active implementation of the national tobacco control policy that the search index for "smoking" was low from July 2016 to August 2019.



Figure 2 Distribution of smoking index in Fujian Province

This survey is mainly for Fujian Province, so the data from February 16, 2015 to February 25, 2022 were selected, and the top four cities are Fuzhou, Quanzhou, Xiamen and Zhangzhou as shown in Figure 2.

2.2. Internet smoking review data collection

Through a comprehensive analysis of the number of downloads and users of domestic APPs (excluding Jitterbug and Baidu), three APPs with more downloads and users were selected, namely, Today's Headline, Zhihu and Weibo. The text data used in this survey are the comments of popular articles on Today's headlines, the answers and comments on Zhihu and the comments on Weibo. There is currently no established database online for this survey, so the required content had to be crawled from the specified websites one by one. The crawler is written in Python, the Requests library is called to request the web pages, and the Pandas library is used to analyze the data.

2.3. Data Processing

Considering the problem that the form of online comments is not uniform, the text is disorganized and interspersed with emoticons and various online terms, and the relevance to the topic is low, if the data is directly text-mined, it will lead to the lack of credibility, validity, and relevance of the results, which in turn will have a negative impact on the conclusion. Based on the above reasons, the text data are pre-processed. Statistics obtained 2190 valid comments and 521 invalid comments, and the effective rate of text data was 76.21%.

3. Analysis of factors influencing college students' smoking behavior

3.1. College students' smoking behavior-Linked analysis

Table 1 Cross tabulation of grade level and whether or not to smoke

			Smoking or not		Total
			Yes	No	
Grade	Sophomore	Count	107	363	470
		Percentage of grade	22.8%	77.2%	100.0%
		Percentage of whether or not they smoke	33.5%	28.5%	29.5%
	Junior	Percentage of total	6.7%	22.8%	29.5%
		Count	141	466	607
		Percentage of grade	23.2%	76.8%	100.0%
	Senior year	Percentage of whether or not they smoked	44.2%	36.6%	38.1%
		Percentage of total	8.9%	29.3%	38.1%
		Count	56	249	305
	Freshman year	Percentage of grade	18.4%	81.6%	100.0%
		Percentage of whether or not they smoke	17.6%	19.5%	19.1%
		Percentage of total	3.5%	15.6%	19.1%
	Master and above	Count	9	127	136
		Percentage of grade	6.6%	93.4%	100.0%
		Percentage of whether or not they smoked	2.8%	10.0%	8.5%
Total		Percentage of total	0.6%	8.0%	8.5%
		Count	6	69	75
		Percentage of grade	8.0%	92.0%	100.0%
		Percentage of whether or not they smoked	1.9%	5.4%	4.7%
		Percentage of total	0.4%	4.3%	4.7%
		Count	319	1274	1593
		Percentage of grade	20.0%	80.0%	100.0%
		Percentage of whether or not they smoke	100.0%	100.0%	100.0%
		Percentage of total	20.0%	80.0%	100.0%

A cross-tabulation of grade and whether or not to smoke was conducted using SPSS 23.0 to produce Table 1 Grade and whether or not to smoke cross-tabulation.

The top three smoking rates were juniors, sophomores and seniors, whose smoking rates were 8.9%, 6.7% and 3.5% of the total, which were much higher than the smoking rates of freshmen and those with degrees at the graduate level and above; the distribution of sophomores and juniors was more concentrated in the grade dimension, with their percentages of 23.2% and 22.8%, respectively. The overall smoking rate was at 20%. It is presumed that freshmen are not familiar with the college environment and will cover up their bad environment, or will not smoke yet. However, as they get older, they will gradually choose to smoke because they know more about the university environment and have more frequent interpersonal interactions, or they will be influenced by their studies. And graduate students have a new level of awareness of their own, the hazards of smoking, etc., smoking is also reduced.

3.2. Smokers' characteristics based mining-C5.0 decision tree model

Based on the above column analysis of the basic profile of the surveyed population and whether they smoked, it was concluded that gender, age and monthly living expenses were relatively independent of whether they smoked or not. Next, the information on the smoking population will be segmented to further explore and filter the factors that have a significant effect on smoking to explain the phenomenon using the C5.0 decision tree model [10].

Decision tree algorithms are guided learning, i.e., the original data must contain both predictor and target variables. Decision trees are so named because the results of their analysis are presented in the form of an inverted tree. The decision tree is composed of root nodes, internal nodes and leaf nodes in order from top to bottom. A node corresponds to a field in the data, i.e., a field - i.e., Question - that divides the data once. Decision trees are divided into categorical decision trees (where the target variable is a subtype value) and regression decision trees (where the target variable is a continuous

type variable). The core of the algorithm is the use of information entropy reduction rate to select the feature variables on each branch [9]. The information entropy is calculated as follows.

$$H(U) = \sum_i P(u_i) \log_2 \frac{1}{P(u_i)} \quad (1)$$

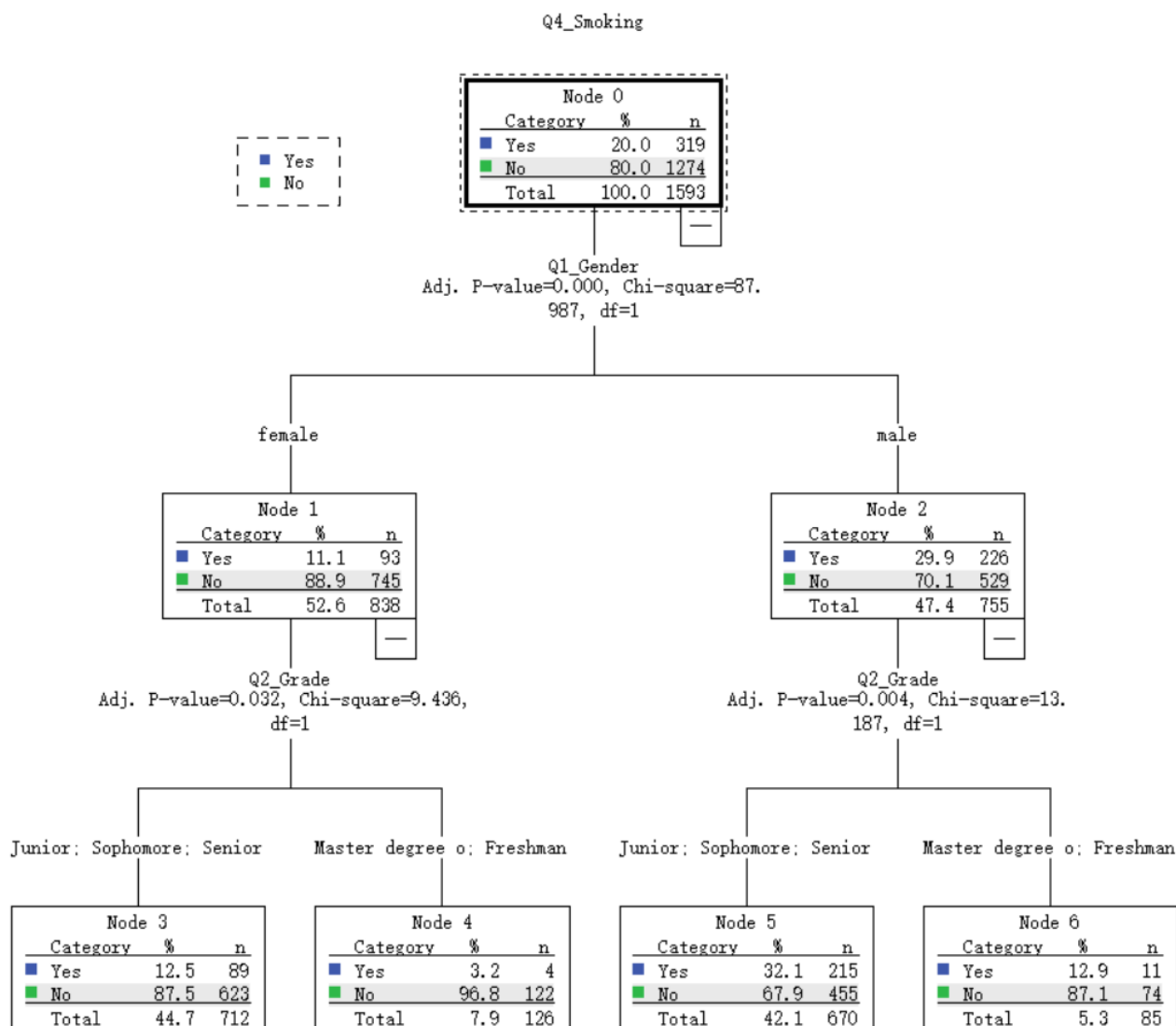


Figure 3 C5.0 Decision Tree Model Results Chart

Through the above Figure 3, it can be found that the p-values of 2 nodes are less than 0.05, indicating that they are highly significant. It indicates that the college students' smoking behavior is more influenced by gender and grade, while the cost of living has a weak influence on their choice of whether to smoke or not, and therefore does not appear in the dendrogram.

In terms of gender, it is clearly observed that the smoking rate of male students is as high as 29.9%, while the smoking rate of female students is on average at 11.1%; on the other hand, grade is also a relevant factor influencing the smoking rate, and in the higher correlation college students' grades (sophomore, junior, and senior), the smoking rate of male students increases slightly to 32.1%, and the smoking rate of female students, although still lower than male students, also increases compared to the general smoking rate. The smoking rate of female students, although still lower than that of male students, increased to 12.5% compared to the general smoking rate; while in the lower grades (freshman) or higher grades (master's and above), both male and female students, their smoking rates decreased significantly, with the male smoking rate at about 12.9% and the female smoking rate dropping to 3.2%.

In response to these results, it can be understood that gender and grade level are the main factors affecting smoking rates. Boys are more likely to choose to smoke compared to girls. The reasons for this may be personal perceptions, inculcation of the surrounding environment, and stress relief. Some boys consider smoking as a handsome behavior to attract the attention of others; some boys are influenced by the people around them and start smoking in order to fit in and follow the footsteps of their friends; some are forced to smoke as their spiritual support to relieve their anxiety, irritability and other bad emotions due to work, study and emotional pressure. Girls, on the other hand, think more about their physical health and the dangers of smoking, so their smoking rate is relatively low.

With regard to grade level, freshmen and master's and doctoral students smoke less compared to sophomores, juniors and seniors. Freshmen are likely to continue their high school habits, focus on their studies, and smoke less; in addition, freshmen face relatively less pressure. Graduate students with advanced degrees have less time to smoke because their lives are focused on research and study. Students in their sophomore and senior years are under pressure from studies, emotions, and work, and are more likely to choose to smoke.

3.3. Analysis of individual differences in the factors influencing smoking - One-way ANOVA

1. Differences in influencing factors by gender

In this study, the independent samples t-test was used to analyze the gender differences, as shown in Table 4-7-1. The results show that the mean scores differed significantly between male and female respondents by gender, and the top three influencing factors in the boys' score ranking were the surrounding environment (4.03), academic and emotional stress (3.99), and poor social environment (3.99); the top three influencing factors in the girls' score ranking were conducting seminars on tobacco control topics (3.97), parents' literacy (3.95), and banning the sale of cigarettes (3.93). There were also significant differences ($p < 0.05$) between male and female students for these factors in terms of general social phenomenon, being influenced by public figures, pressure of club work, bad club atmosphere, personal interests, academic and emotional pressure, and surroundings.

2. Differences in influencing factors by grade level

In this study, the one-way ANOVA test was used to analyze the grade differences in the influencing factors. The results showed that there were significant grade differences ($p < 0.05$) in social prevalence, influence by public figures, control of students' smoking, club work pressure, good club atmosphere, bad club atmosphere, club interpersonal interaction, frequency of club activities, smoking environment, personality can be an influence and academic and emotional stress factors.

To further elucidate the differences in grade level on the above 11 dimensions, the LSD (Least Significant Difference) method was used to conduct a test of differences in means, as the table is too large, see Appendix 3.8: LSD Test of Differences in Means of Grade Levels for details. The results of the multiple comparisons showed that.

1. In the public figure factor the grade is sophomore score will be significantly higher than the score of freshman, senior, graduate and above; the grade is junior in the public figure factor score is significantly higher than the score of freshman, senior, graduate and above.

2. The scores of junior year on the factor of controlling student smoking are significantly higher than the scores of freshman and graduate students and above; the scores of senior year on the factor of controlling student smoking are significantly higher than the scores of graduate students and above.

3. Graduate students and above scored significantly lower than sophomores, juniors, and seniors on the stress factor of club work.

4. Graduate students and above will score significantly lower than sophomores and juniors on the good club atmosphere factor, bad club atmosphere factor, study and emotion factor, and general social phenomenon factor; seniors will score significantly lower than juniors on the bad club factor.

5. The scores of graduate students and above in the interpersonal factor of association are significantly lower than those of students in other grades.

6. On the factors of smoking environment and frequency of club activities, the scores of freshmen were significantly lower than the scores of sophomores.

7. The scores of first-year students were significantly lower than those of sophomores, juniors, and seniors in the factors of study and emotion, frequency of club activities, personality, stress in club work, and social phenomenon.

3. Differences in monthly living costs of the influencing factors

This study used one-way ANOVA test to analyze the differences in monthly living expenses of the influencing factors. The results showed that there was a significant difference in monthly living cost on the factors of appropriate national smoking policy, social prevalence, influence by public figures, no smoking area, club work pressure, good club atmosphere, bad club atmosphere, club interpersonal interaction, frequency of club activities, smoking environment, upbringing, personal interests, personality can be an influence, study, emotional stress and surrounding environment ($p < 0.05$).

To further elucidate the differences in grade levels on the above 11 dimensions, the LSD (Least Significant Difference) method was used to test for differences in means, and because the table is too large, the specific table is shown in Appendix 3.10: LSD Test for Differences in Means of Monthly Living Costs. The results of the multiple comparisons showed that.

1. The scores on the bad association atmosphere factor, smoking environment factor, personal interest factor, public figure influence factor, and national smoking policy factor for monthly living expenses below \$1200 are significantly lower than those for monthly living expenses above \$1200.

2. The scores of those whose monthly living expenses are less than 1200 RMB are significantly lower than those whose monthly living expenses are 1200-2499 RMB for the factors of surrounding environment, study and emotional pressure, frequency of club activities, social phenomenon and social work pressure.

3. The score of those whose monthly living expenses are 2500 RMB or more is significantly lower than the score of those whose living expenses are 1200-2499 RMB.

4. For the factor of no smoking area, the scores of those with monthly living expenses below \$1200 are significantly lower than those with monthly living expenses between \$1200 and \$1799; those with monthly living expenses above \$2500 are significantly lower than those with monthly living expenses between \$1200 and \$2499.

5. The score of the good community atmosphere factor is significantly lower than the score of \$1200-\$2499 for those who live on more than \$2500 per month; the score of those who live on less than \$1200 per month is significantly lower than the score of \$1800-\$2499.

4. Differences in influencing factors on whether or not to smoke population

This study used independent samples t-test to analyze the gender differences. The results show that the mean scores of the respondents differed significantly between whether they smoked or not, and the top three factors influencing the choice of smoking were personality (4.06), academic and emotional stress (4.03), and poor social environment (4.03); the top three factors influencing the non-smoking score were conducting seminars on tobacco control (3.99), surrounding environment (3.97), and parental culture (3.97). There were significant differences ($p < 0.05$) between those who smoked and those who did not smoke in terms of appropriate national smoking policy, social prevalence, influence by public figures, prohibition of cigarette sales, pressure from club work, frequency of club activities, conducting tobacco control activities, and academic and emotional pressure.

Overall, male students chose to smoke with a slightly smaller overall score (75.51) than female students (76.13) for the influencing factors.

3.4. Regression analysis of factors influencing college students' smoking behavior-logistic regression

The logistic regression model constructed is

$$\text{logit } P = \ln \left(\frac{P}{1-P} \right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n \quad (2)$$

Model solving. For the logistic regression analysis, in general, a coefficient of determination of 50%-80% of the regression coefficient is a high level. The model was tested and the results were obtained in Table 2.

Table 2 logistic model test results table

Steps	-2 log-likelihood	Cox-Snell R-squared	Nagorco R-squared
1	767.593a	.405	.641

The regression coefficients in the model in the table reached 40.5% versus 64.1%, with significant regression coefficients. Thus, it can be considered that the model has a good prediction and the data fit the model well. Thus, the next step of calculation is carried out. According to the output of the regression coefficients, there are some regression coefficients that are not significant enough. The three factors significant Sig<0.05, thus the three factors of social prevalence (X2), banning the sale of cigarettes (X4) and conducting tobacco control type activities (X5) are considered to be significant in the equation. However, since the three factors were not sufficient, the above three independent variables were removed and the logistic model was run again with the following equation.

$$\text{Logistic} = -.777 - 0.318X_2 + 0.387X_4 + 2.225X_{14} \quad (3)$$

Get the results

The results of the logistic regression model were observed and analyzed. Based on the significant differences in the regression coefficients, it was found that social prevalence, banning the sale of cigarettes, conducting tobacco control activities, appropriate national smoking policies, being influenced by public figures, and academic and emotional stress had a significant effect on whether college students smoked.

A negative relationship was found between the implementation of tobacco control activities, banning the sale of cigarettes and the issuance of appropriate smoking policies and the probability of smoking among college students, and the more frequent these activities were, the more likely they were to reduce the probability of smoking among college students. At the same time, smoking became a common phenomenon in contemporary society and was positively correlated with the probability of smoking among college students, and the more college students were influenced by the smoking of public figures, the more they tended to smoke. Finally, it was found that smoking among college students was also positively related to academic and emotional stress, and the higher the stress, the higher the probability of smoking among college students.

Exp(B) reflects the change in the explanatory variables per unit change in the explanatory variables under the same circumstances, so the factors that influence the degree of change in the renewal rate according to the magnitude of Exp(B) are ranked as follows: activities to control smoking (X13) > ban on the sale of cigarettes (X4) > general social phenomenon (X2) > appropriate national smoking policy (X1) > academic and emotional stress (X19) > influenced by public figures (X3).

(1) Appropriate policies and the implementation of tobacco control type activities are more likely to reduce the probability of smoking among college students.

The results showed that conducting tobacco control type activities, banning the sale of cigarettes and issuing appropriate smoking policies were effective in controlling college students' smoking. The implementation of tobacco control activities is beneficial to college students' tobacco control. Meanwhile, banning the sale of cigarettes and issuing appropriate smoking policies were the second and fourth influencing factors, indicating that whether or not contemporary college students smoke is influenced by policies and publicity. The ban on the sale of cigarettes and the publication of smoking policies can, to a certain extent, make college students quit smoking passively, especially those who are less addicted to smoking and have tried smoking for the first time. As a result, the smoking rate of college students will be gradually reduced by a combination of policies and publicity.

(2) There is a positive correlation between the prevalence of smoking in contemporary society (public figure bias) and the probability of smoking among college students.

The results show that the more popular smoking is in contemporary society or the more public figures are inclined to smoke, the higher the smoking rate of college students. Some college students are aware of the harmful effects of smoking, but they continue to do so because they believe that smoking is a socially acceptable behavior and is a "classy" and "elegant" thing to do, and this situation is worsened when their idols also smoke. This situation continues to worsen when their idols also smoke. Thus, social tendencies and the idol effect are also an important factor.

(3) The higher the academic and emotional stress, the higher the probability of smoking among college students.

The results show that the more stressful the academic life is, the more likely college students are to smoke. Frustration and the need for psychological support are some of the reasons for smoking, but in the survey, it was found that some of the students started smoking when they lost their love or encountered sadness, and after the first time they smoked, they unconsciously picked up cigarettes whenever they encountered frustration. Therefore, the level of stress and the smoking rate of college students are also closely related.

4. Conclusion

In this paper, we started from both online survey and field research, collected data through questionnaire distribution, interviews, and web crawlers, and organized and tested the obtained data, and obtained the following conclusions by using cross-tabulation analysis, C5.0 decision tree, logistic model, and structural equation modeling.

Based on the C5.0 decision tree model, it was learned that the smoking rate of college students is more influenced by gender and grade level and less influenced by cost of living. Regardless of grade level, male students will have a significantly higher smoking rate than female students; while freshmen and students with graduate degrees or higher will smoke significantly less than sophomores, juniors and seniors.

The prevalence of smoking among college students was positively correlated with the influence of public figures, the prevalence of smoking in contemporary society, and academic and emotional stress, while the prevalence of smoking among college students was negatively correlated with the implementation of tobacco control activities, banning the sale of cigarettes, and issuing appropriate smoking policies.

Cultivating good personal interests among college students can help them to divert and alleviate their academic and emotional stresses, thus blocking the source of smoking or reducing the frequency of smoking.

Family environment and parents' behavior can have a great influence on students' personality, personal interests and behavior, thus indirectly influencing whether college students smoke or not. When college students are assigned intense work tasks in a bad club atmosphere, they may develop negative and angry emotions and try to relieve them by smoking.

The influence of public figures in the Internet era is even greater, and if public figures do not play a good leading role in tobacco control and smoking control, they may make college students who are fans follow or even imitate their behavior.

Banning the sale of cigarettes in schools, setting up smoking areas, regulating students' smoking behavior, and conducting lectures on tobacco control are measures that will have a subtle effect on students' tobacco control and tobacco control from the general environment of the school.

References

- [1] Liu Yixuan, Yin Decai, Li Chenglong, et al. A survey on smoking status and related cognitive attitudes of students in Yan'an University [J]. Journal of Yan'an University (Medical Science Edition).2021,19(02):16-21

- [2] Lu Yanhua, Wang Hongxia, Zhou Chang, et al. Analysis of the current situation of smoking and factors influencing current smoking among college students in Wuhan [J]. South China Preventive Medicine, 2021, 47(02): 150-153.
- [3] Lv Binghui, Deng Qinglong, Xia Juan, et al. A study on the current situation of smoking among some college students and family members in a university in Shanghai[J]. China Health Education, 2020, 36(09):801-806.
- [4] Dai Lifan, Pu Rui, Li Xi, et al. Analysis of the current situation and influencing factors of smoking among college students in Guizhou Province[J]. China School Health, 2019, 40(12):1828-1830+1834.
- [5] Huang Xuanyang, Gao Yongrong, Yuan Jiang. Analysis of the current situation of college students' smoking and its intervention countermeasures--a case study of civil aviation-related students[J]. Green Technology, 2021, 23(03):209-210+213.
- [6] Lv YM, Liu W, Wei ZS. Analysis of smoking situation and influencing factors among undergraduate students in a university[J]. China Health Standards Management, 2021, 12(02):17-20.
- [7] Zheng Tong. Study on the correlation between the current situation of smoking and drinking and psychological factors among college students[D]. Wannan Medical College, 2020.
- [8] Tong Tong. Rationalizing the meaning behind the tightening of tobacco control policies[N]. China Economic Times, 2015-05-06(002)
- [9] Ma Ting, Liu Sihan, Guo Ruixia. Research on larch plantation extraction based on C5.0 decision tree algorithm[J]. Gansu Science and Technology Vertical, 2020, 49(04):7-12.
- [10] Guo Zhengjun, Song Jinggui, Wang Jing, et al. Exploration of C5.0 decision tree model in predicting medication adherence in patients with severe mental disorders[J]. Modern preventive medicine, 2021, 48(01):110-113+189.