

Prediction of Oral Cancer Incidence in Hunan Using a Time-Series Prediction Model Based on Logistic Regression Modeling

Linfeng Peng *

School of Information Science and Technology, Beijing Forestry University, Beijing, China

* Corresponding author: plfoffice@bjfu.edu.cn

Abstract. Scientific and reliable prediction of the incidence of oral cancer is very meaningful, which can assist the medical industry to carry out effective prevention and treatment measures, and also serve as a reference and basis for the formulation of related health policies. Therefore, this paper mainly explores the prediction method based on the Logistic Regression model to predict and analyze the number of oral cancer incidences. The study in this paper found that age, gender, and betel nut consumption were associated with the incidence of oral cancer. Among them, males and middle-aged and older age groups were more susceptible. The time-series prediction indicates that the model accurately predicts oral cancer incidence in multifactorial samples, which provides effective preventive and curative measures for the healthcare industry and serves as a basis for health policy. Future studies can improve the accuracy through comprehensive data collection and advanced model optimization. The results of the study provide a reference for early prevention and intervention of oral cancer, which can help formulate health policies to serve public health and individual health.

Keywords: Oral cancer, incidence, logistic, predictive analysis.

1. Introduction

Cancer is an unfortunate event that has a small chance of occurring for an individual. However, with the changes in the living habits of modern people, such as dependence on tobacco and alcohol products, the incidence of cancer is increasing year by year, among which oral cancer is very representative. According to [1], the incidence and mortality rates of oral cancer in the world are 5.9/100,000 and 3.2/100,000, respectively, and in less developed regions, they are 5.4/100,000 and 3.4/100,000, respectively, and in China, they are 1.6/100,000 and 0.9/100,000, respectively. A study showed that the incidence and mortality rates of oral cancer in China were 2.17/10 million and 0.89/10 million, respectively. The results of the study showed that the incidence and mortality of oral cancer were mainly concentrated in people over 40 years of age, and the incidence and mortality of elderly people were higher in men than in women. In China, men are the main group of smokers, and in addition to bad habits such as drinking and chewing betel nuts, the stress of work and life has led to a decline in the body's immune system.

In Hunan, China, because of its special natural climate environment and historical and cultural factors, the local betel nut market is very prosperous, however, this has also led to Hunan as the province with the highest incidence of oral cancer in China. According to [2], the incidence rates of oral cancer in Hunan from 2009 to 2012 were as follows: Hunan: 332/100,000 males and 227/100,000 females; urban Hunan: 566/100,000 males and 160/100,000 females; rural Hunan: 266/100,000 males and 118/100,000 females. It can be seen that the incidence rate of oral cancer in Hunan is significantly higher than that of the whole country, which has caused many hazards and misfortunes to local people and their families. Therefore, it is very meaningful to make a scientific and reliable prediction of the incidence rate of oral cancer in society as a whole as well as in a specific area. Scientific prediction can not only assist the medical industry to carry out effective preventive and treatment measures but also serve as a reference and basis for the formulation of relevant health policies.

There are many problems and complex factors affecting the predictive analysis of oral cancer, and obtaining the weights of various influencing factors and the probability of occurrence is one of the

problems to be solved in the predictive analysis [3]. Logistic Regression model essentially fits the data into a Logistic Regression model [4], to predict the likelihood of occurrence of the event, which can be used to analyze the relationship between the occurrence of the disease and the risk factors. quantitative relationship between the occurrence of disease and each risk factor.

In this paper, we analyze the characteristics of oral cancer incidence in Hunan and the Logistic Regression model and propose to use the Logistic Regression model to predict the incidence of oral cancer in Hunan and implement the code.

2. Methodology and Data

2.1. Characterization of Oral Cancer

Oral cancer is a general term for malignant tumors occurring in the oral cavity, most of which belong to squamous epithelial cell carcinoma, which is the so-called mucosal variant. In clinical practice, oral cancer includes gingival cancer, tongue cancer, hard and soft palate cancer, jawbone cancer, floor of mouth cancer, oropharyngeal cancer, salivary adenocarcinoma, lip cancer, and maxillary sinus cancer, as well as cancers occurring in facial skin and mucosa [5]. Oral cancer is one of the more common malignant tumors of the head and neck. Causes of oral cancer: 1. addiction to tobacco and alcohol, 2. poor oral hygiene, 3. long-term stimulation by foreign bodies, 4. malnutrition, 5. leukoplakia and erythema of the mucosa, and 6. long-term heavy use of betel nut. These etiologies can be predicted as input features of Logistic Regression model [6].

However, oral cancer, as a kind of cancer, has many influencing factors and is inherently more contingent, which requires precise selection and standardization of its possible influencing factors. Large-scale population prediction has high requirements for model performance and accuracy, which requires rigorous evaluation of the model. Currently, the focus of oral cancer prediction at home and abroad is mostly on the influence of a certain factor or a certain behavior, such as surgery, on the development of oral cancer in individuals, and there is a lack of prediction analysis of individuals and the regional population as a whole. In this paper, we hope to realize the prediction of oral cancer in regional groups and individuals through prediction models.

2.2. Data Sources

The data source for this study was population health data from the Hunan region, collected through the use of questionnaires. The main concerns of this dataset include gender, individual oral cancer disease incidence, betel nut and tobacco use habits, and the presence of prognostic symptoms of oral cancer. To ensure the reliability and representativeness of the data, a rigorous questionnaire design and implementation process was used to ensure that high-quality data were collected. The questionnaire covered several aspects of oral cancer, including disease prevalence, betel nut and tobacco use, and possible early symptoms. The survey was anonymized to protect the privacy of the participants and was collected through an online platform to improve the efficiency of data acquisition. To facilitate subsequent data processing and analysis, this paper chose to store the data in CSV file format. This tabular approach to data storage facilitates flexible and efficient processing in various statistical and analytical tools

2.3. Data Sources

Logistic regression is a statistical model used to deal with classification problems, especially binary classification problems [7, 8]. In oral cancer research, this paper uses Logistic Regression to explore the relationship between different factors and the prevalence of oral cancer, such as betel nut and tobacco use, and prognostic symptoms of oral cancer. The following is a brief description of Logistic regression:

1) Model expression:

The model expression for Logistic regression is as follows:

$$P(Y = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n)}} \quad (1)$$

Where $P(Y = 1)$ is the probability of the event Y occurring. e is the base of the natural logarithm. β_i is a parameter of the model that represents the effect of the variable on Y . X_i is a characteristic variable of the model.

The right-hand side of this expression is a logistic function, also known as a sigmoid function, which converts a linear combination of inputs into a probability value between 0 and 1. The model fits the training data by learning the parameters.

2) Model Application:

Logistic regression is commonly used to solve dichotomous classification problems, such as the prediction of having oral cancer ($Y=1$) versus not having oral cancer ($Y=0$). The model is able to classify new samples based on the input features and provide the probability of an event occurring by learning patterns in the training data.

3) Model Evaluation:

The performance of Logistic regression is usually evaluated by metrics such as confusion matrix, accuracy, precision, recall, and F1 score. These metrics can help in assessing the predictive ability and generalization performance of the model.

The simplicity and interpretability of the logistic regression model make it one of the preferred models for many classification problems. In oral cancer research, by analyzing the model parameters, this paper can understand the relative contribution of each factor to the risk of developing oral cancer, thus providing valuable information for prevention and intervention.

2.4. Algorithm Implementation and Analysis

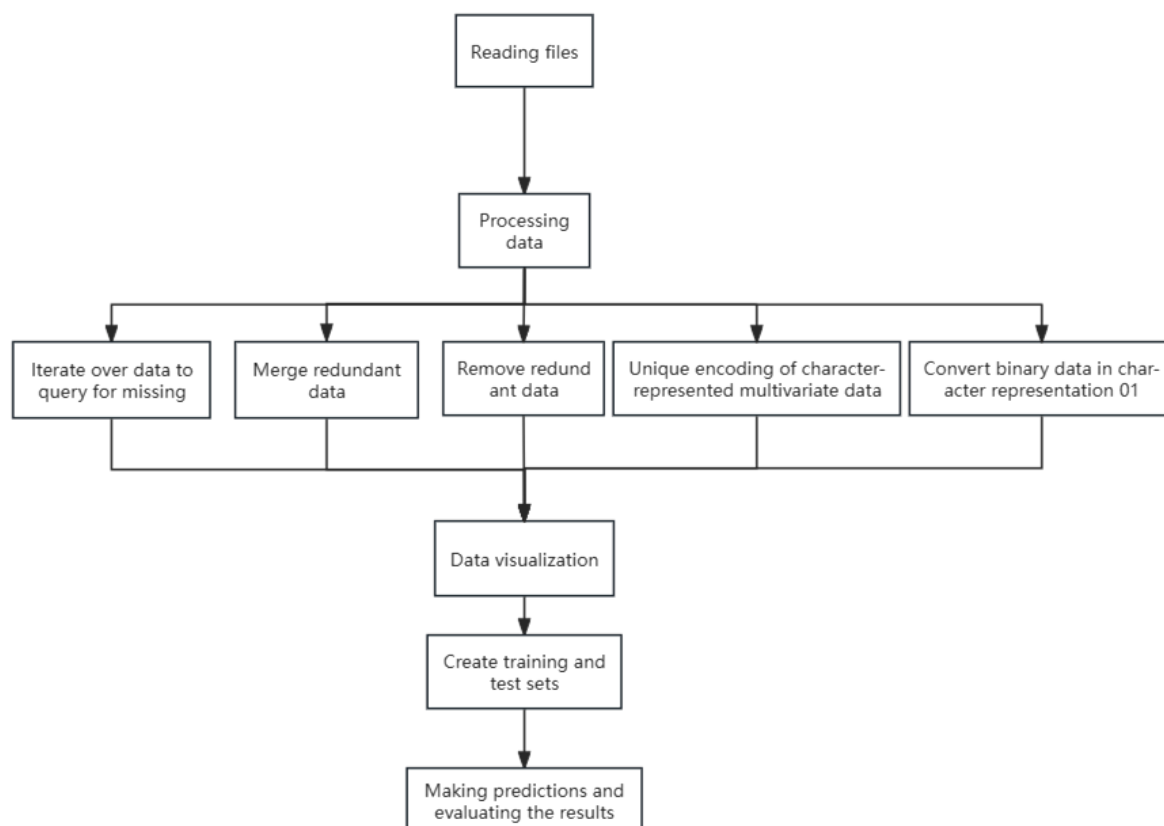


Figure 1. Program Flow Chart

The implementation of the algorithms in this paper focuses on how to use the Logistic Regression model for time-series prediction of data and how to process the data into computer-friendly types. The flowchart is shown in Fig.1.

Firstly, the documents are read and stored in data frames for easy subsequent processing. Secondly, the data is normalized. In actual data collection, the results obtained are usually diverse and data manipulation and organization are necessary to facilitate subsequent investigation and computer processing. For example, in some data items, the results may be expressed as "YES" or "NO", but to adapt to the processing of computer language, it is necessary to convert them to binary representation, usually using 0 and 1 instead. In the process of organizing the data, it is necessary to deal with possible missing or omitted data. Missing data may adversely affect subsequent analyses, so appropriate measures must be taken to deal with them. One common approach is to fill in the missing data, optionally using the mean, median, or other statistical indicators to maintain the integrity of the dataset. Another strategy is to remove samples that contain missing data, provided they are relatively small and do not significantly affect the overall findings. At the same time, care should be taken to deal with outliers in the data, which may be due to recording errors or other reasons. In the process of data organization, outliers can be identified and handled by setting thresholds or using statistical methods to ensure the reliability of the data. Finally, the processed data is imported and the model is trained using the training set function and test set function and a randomly generated test set is input to the model for prediction.

2.5. Result

Fig.2 illustrates the ROC curve of the model. The receiver operating characteristic curve (ROC curve), is also known as the sensitivity curve. the ROC curve is a common method for testing the efficiency and accuracy of binary classification models, and the closer it is to the upper left corner the more that the model predicts more accurately. From the Figure 2, it can be concluded that the model used in this paper is highly accurate and reliable.

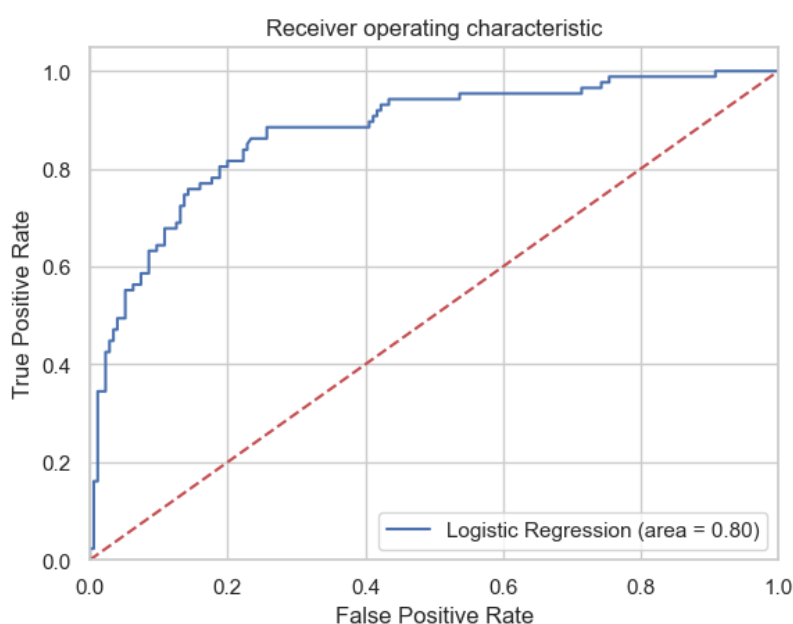


Figure 2. ROC curve for testing the accuracy of the model, the closer the curve is to the upper left corner the more accurate the model prediction is.

3. Data Visualization and Analysis

Fig.3 illustrates the age distribution between male and female positive samples. The age distribution of the positive cases can be seen. Its distribution is approximately the same as the age distribution status of the respective positive cases of men and women. From the Fig.3, it can be seen that the age of positive samples of both sexes is mostly concentrated between 50 and 80 years old, with the highest value around 60 years old. Between the ages of 50 and 60, there is a clear upward trend in the number of positive cases, while the positive cases for females remain stable between the

ages of 60 and 70, while for males there are ups and downs between the ages of 60 and 70, but there is a significant decrease between the ages of 70 and 80 compared to the previous period. The trend of the graph indicates that the majority of patients with oral cancer are middle-aged and elderly. The decrease in the distribution of samples after the age of 60 may be due to the impact of oral cancer on health status.

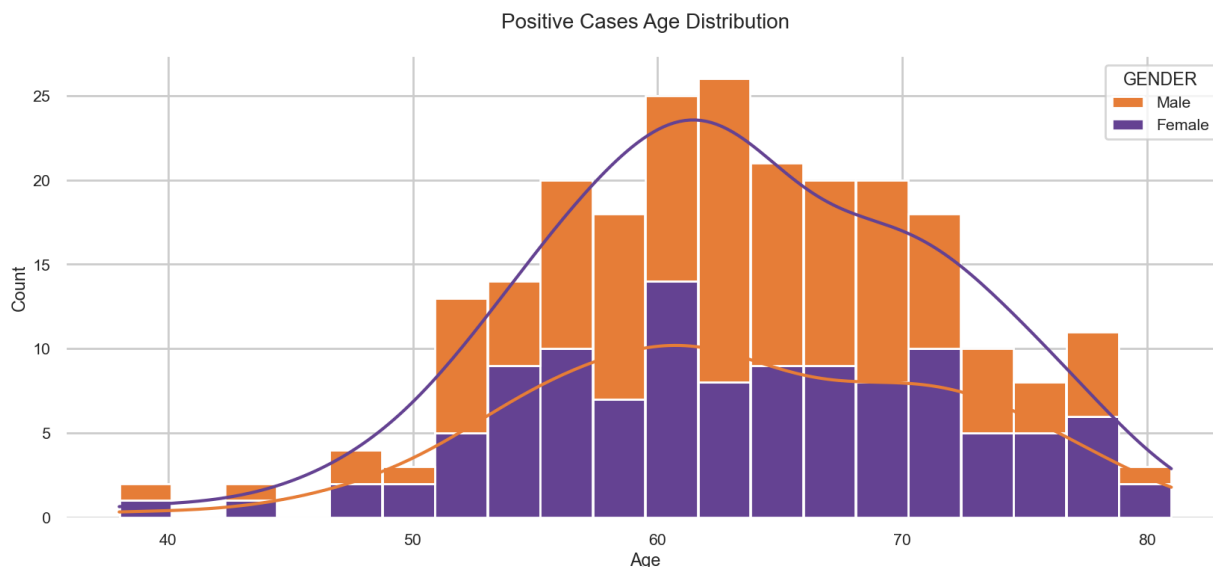


Figure 3. Male and female age curves and histograms of positive cases.

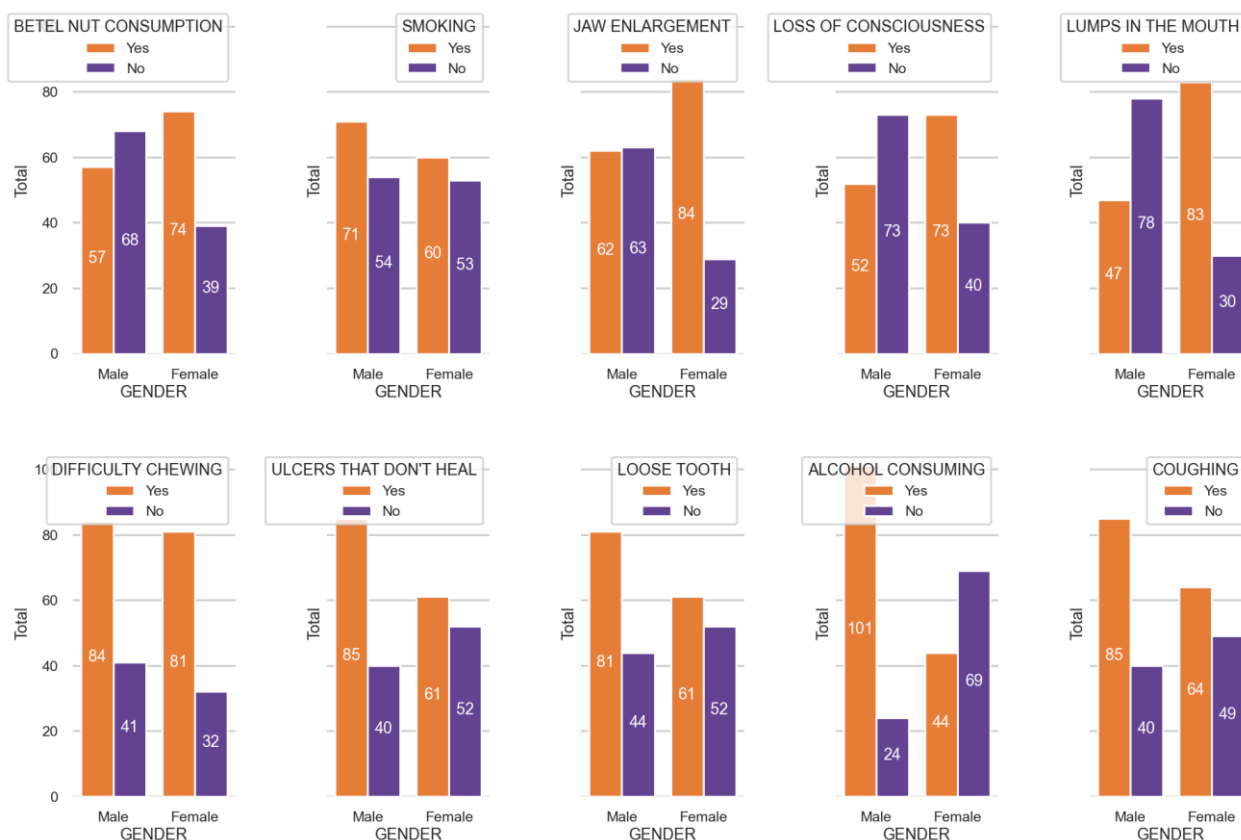


Figure 4. Number of positives and non-positives by gender in each data set.

Fig. 4 shows the comparison of the number of positive and non-positive cases by gender in each data, from which the impact of each symptom on the risk of oral cancer by gender can be simply analyzed. For example, in the data of "shortness of breath", the number of cases with this symptom

among men is 80, which is much more than the number of cases without this symptom is 45; the situation for women is similar, the number of cases with this symptom is 73, and the number of cases without this symptom is 40, which indicates that the symptom of "shortness of breath" has a significant influence on the risk of oral cancer in different genders. This suggests that the symptom "shortness of breath" affects the risk of oral cancer to about the same degree in both sexes. On the other hand, the symptom "pain" was present in 94 cases in men and only 48 cases in women. This suggests that the symptom "pain" has a significant impact on the risk of oral cancer in men, while it has a relatively small impact on the risk of oral cancer in women.

Fig. 5 shows a heat map of the data, with the horizontal and vertical columns corresponding to each data type, e.g., "age" in the first row and column. The cell in a row and column represents the link between the two data in that row and column. For example, the first column corresponds to the connection between "age" and various data, for example, the first row and first column is the relationship between "age" and its own influence, 1.0; the second row and first column corresponds to the relationship between "gender" and "age". The first column of the second row corresponds to the relationship between "gender" and "age". Here we can see that in the ORAL CANCER column, ULCERS THAT DON'T HEAL is the data with the largest positive correlation, excluding itself. This suggests that allergic reactions have the greatest correlation with developing oral cancer and that allergic reactions are likely to be an early symptom of oral cancer.

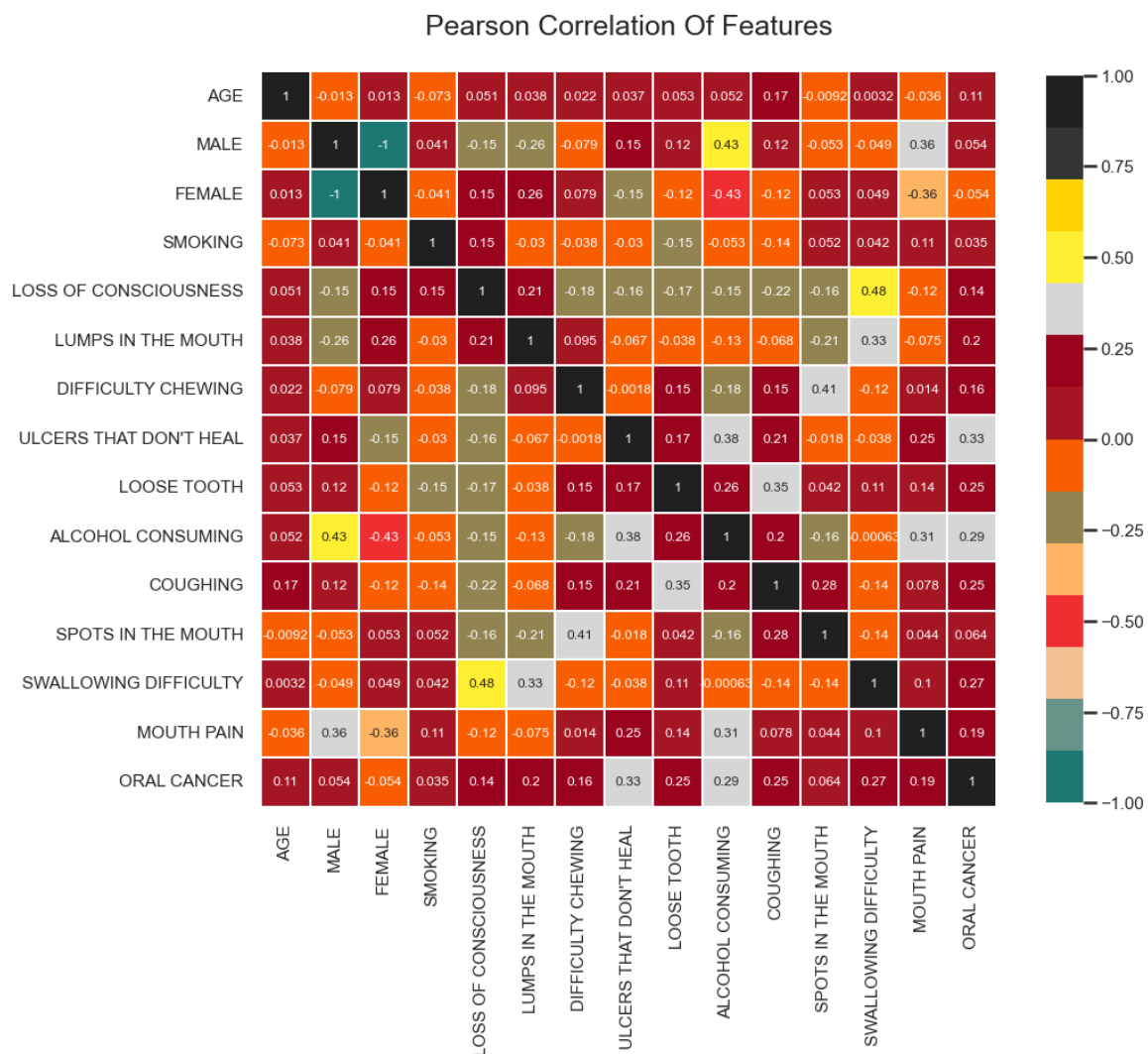


Figure 5. Heatmaps of individual data, reflecting the correlation between the individual data.

4. Conclusion

Through time-series prediction based on Logistic Regression modeling, this paper provides an in-depth discussion of the trends and factors associated with the incidence of oral cancer in Hunan. It was found that factors such as age, gender, and betel nut consumption habits were associated with the incidence of oral cancer. In the data analysis, the incidence rate of oral cancer is higher among men, and the middle-aged and elderly population is more likely to be affected. In the time series prediction, the time series prediction based on the Logistic Regression model realized in this paper can give more accurate prediction results, has a good prediction effect in multifactorial samples with large populations, and can predict the incidence rate according to the health status of the samples, which can be used to assist the medical industry to carry out effective preventive and therapeutic measures and become a reference and basis for the formulation of relevant health policies. The results can be used as a reference for health policy formulation.

The results of the study are summarized in the following table. Future research can further improve the accuracy of prediction through more comprehensive data collection and more advanced model optimization methods. In addition, deeper factors such as biology and genetics can be considered to fully understand the pathogenesis of oral cancer. The results of this study provide a reference for early prevention and intervention of oral cancer, which can help formulate relevant health policies and clinical practices. By continuously deepening the research on the pathogenesis of oral cancer, public health, and individual health can be better served.

References

- [1] Latest world cancer statistics - Globocan 2012: Estimated Cancer Incidence, Mortality and Prevalence Worldwide in 2012.
- [2] Cancer Statistics in China, Wanqing Chen, PhD, MD1; Rongshou Zheng, MPH2; Peter D. Baade, PhD3; Siwei Zhang, BMedSc4; Hongmei Zeng, PhD, MD5. Freddie Bray, PhD6; Ahmedin Jemal, DVM, PhD7; Xue Qin Yu, PhD, MPH8, 9; Jie He, MD10, 2015.
- [3] Lin Jing, Cao Yujie, He Baochang et al. Study on the association between oral hygiene status and the risk of oral cancer [C]//Professional Committee of Geriatric Dentistry, Chinese Dental Association.2019 Compendium of Papers from the 14th National Academic Annual Meeting of Geriatric Dentistry, Professional Committee of Geriatric Dentistry, Chinese Dental Association. [Publisher unknown], 2019: 2. DOI: 10.26914/c.cnkihy. 2019. 024388.
- [4] Baoheng Xue. Application of logistic regression analysis based on logistic regression analysis in studying the purchase of energy-saving electrical appliances by residents in a place in Tianshui [J]. Modern Business Industry, 2023, 44 (21): 253 - 255. DOI: 10.19311/j.cnki.1672 - 3198. 2023. 21. 085.
- [5] Xiong Xinbiao, Xie Xionggang, Yang Peijun et al. Research on the danger of mine floor sudden water based on PCA-Logistic regression model [J]. Coal Mine Safety, 2023, 54 (10): 176 - 181. DOI: 10.13347/j.cnki.mkaq.2023. 10. 023.
- [6] Zang Yang, Zhao Huarong, Liu Pan et al. Predictive value of multifactorial logistic regression model and XGBoost model for the occurrence of oral infection during radiotherapy in tongue cancer patients [J]. Chinese Journal of Modern Medicine, 2023, 33 (19): 66 - 73.
- [7] Journal | [J] The American Statistician. volume 72, Issue 1. 2018. pp 37 - 45 Forecasting at Scale Sean J. Taylor; Benjamin Letham
- [8] Zhu Meihong, Zou Zhandong, Lin Tainan. Binary Logistic Regression Analysis of Factors Influencing College Students' Dry Eye [J]. Journal of Huaqiao University (Natural Science Edition), 2023, 44 (05): 622 - 627.