RESEARCH ARTICLE

Study on the prediction model of main chemical components and ecological environmental factors of Yunnan tobacco leaves

Quan Gao[1], Lijia Wang[1], Yun He[1], Zaosong Li[1], Hongjun Luo[1], Lumin Zhang[2], Kun Luo[3], Jianfeng Chen[3], Chengren Ouyang[4, *]

[1]Big Data College, Yunnan Agricultural University, Kunming, Yunan, China. [2]Technology Centre of Yunnan Tobacco Company, Honghe, Yunan, China. [3]Lijiang Tobacco Company of Yunnan Province, Lijiang, Yunan, China. [4]College of Tobacco Science, Yunnan Agricultural University, Kunming, Yunan, China.

Received: August 16, 2025; **accepted:** December 14, 2025.

In the tobacco sector, the chemical composition of tobacco leaves significantly influences both the leaf quality and tobacco cultivation strategies. This study focused on the Honghe region of Yunnan, China and established prediction models for the main chemical components and ecological environmental factors of tobacco leaves. By performing data cleaning, feature engineering, and standardization on multi-variety tobacco leaf sample data, combined with meteorological, soil, and other ecological environmental data, a high-quality dataset was constructed. Through correlation analysis, prediction models including linear regression, decision trees, and random forests were subsequently constructed. The results showed that the 3 different models of multiple linear regression, decision tree, and random forest algorithm exhibited distinct advantages in predicting the main chemical components of different tobacco varieties. By analyzing the relationship between chemical components and ecological factors and implementing regulation with the aid of predictive models, the quality coordination and safety of tobacco leaves could be effectively enhanced. This research provided strong support for tobacco leaf quality evaluation and enhanced the scientific basis for tobacco quality regulation and planting regionalized planting.

Keywords: Tobacco leaf quality; Chemical components; Correlation analysis; Prediction model.

[*]**Corresponding author:** Chengren Ouyang, College of Tobacco Science, Yunnan Agricultural University, Kunming, Yunan 650201, China. Email: gaoq@ynau.edu.cn.

Introduction

As a crucial economic crop in China, tobacco plays an important role in agricultural development and regional economies. Its quality formation is influenced by multiple factors such as variety, climate, soil, and cultivation management. In recent years, research has made significant progress in constructing prediction models for the main chemical components of tobacco leaves. Scholars worldwide have conducted extensive research in the field of tobacco leaf-related prediction models, covering main chemical components, physical characteristics, as well as other aspects such as diseases and yield.

Ke *et al*. used stepwise regression to build prediction models for 71 tobacco-planting townships, achieving prediction accuracies of 89.04%, 68.90%, and 86.6% for total sugar, chlorine, and potassium content, respectively [1].

Zhang *et al*. used regression analysis and back propagation (BP) neural networks to study the correlation between soil nutrient factors and tobacco leaf chemical components and constructed prediction models [2]. Furthermore, researchers used random forest, gradient boosting decision tree, and extreme gradient boosting methods to predict soil organic matter and total nitrogen content. The gradient boosting decision tree model showed the best prediction accuracy [3]. Liu *et al*. applied multiple methods including multiple linear regression, stepwise linear regression, and BP neural networks to construct a prediction model for nicotine content in flue-cured tobacco. The BP and GA-BP neural network models showed a strong correlation between predicted and actual values with small root mean square errors, indicating effective nicotine content prediction [4]. Xia *et al*. used four algorithms including BP neural network, random forest, linear regression, and stepwise regression to study the relationship between meteorological factors and the single leaf weight of flue-cured tobacco and construct models. The study found that the stepwise regression algorithm model based on multiple factors had the best simulation effect, capable of better simulating the annual peak and trough values of single leaf weight at different leaf positions followed by the random forest algorithm [5]. Chen *et al*. used a multiple linear statistical prediction model and a BP neural network prediction model to construct a prediction model for the physical characteristic indicators of high-quality light-aroma tobacco leaves. The results showed that the neural network prediction model for tobacco leaf physical characteristics was significantly lower than the multiple linear prediction model in terms of root mean square error and normalized root mean square error, indicating a relatively ideal simulation effect [6]. Huang *et al*. applied partial least squares to establish a near-infrared spectroscopy prediction model for heavy metals in sun-cured red tobacco and confirmed that this model performed excellently in predicting heavy metal content in raw tobacco leaves from different origins [7]. Zhang *et al*. used local binary pattern and gray level co-occurrence matrix (LBP-GLCM) algorithm combined with a BP neural network to create a relationship model between the micro-texture and oil content of flue-cured tobacco. The correct recognition rate for predicting flue-cured tobacco oil content grade was 93.33%, and the model correlation coefficient was 0.91486, showing certain advantages in predicting flue-cured tobacco oil content grade [8]. Chen *et al*. used linear regression and quadratic regression models to predict tobacco's response to nitrogen fertilizer and found that the linear model better described the response of total sugar (TS), total nitrogen (TN), and nicotine content to nitrogen application rates, while the quadratic model had advantages in describing the yield response to nitrogen fertilizer application, especially when the nitrogen application rate was 60 - 120 kg N/ha. However, the maximum output value estimated by both models was lower than the actual output value [9]. Zhou *et al*. constructed a sensory quality prediction model based on key chemical indicators affecting the quality, providing technical support for the objective evaluation of tobacco leaf quality [10]. Li *et al*. used four algorithms including BP neural network, four-element random forest, multi-element random forest, and stepwise regression to establish a single leaf weight model for flue-cured tobacco based on annual-scale four meteorological elements and ten-day-scale multiple meteorological elements and found that there was no significant difference among the four algorithms for middle leaves, while the random forest had the highest average accuracy for upper leaves [11]. Han *et al*. comprehensively applied Mann-Kendall trend analysis, Pearson correlation analysis, and stepwise multiple regression methods to deeply analyze the meteorological data of the region and construct prediction models for tobacco black shank and wildfire diseases and identified nine meteorological indicators as strongly correlated with the average incidence rate of black shank. The resulting early-warning models provided strong predictive support for the prevention and control of tobacco diseases [12]. Qi *et al*. used multivariate statistical analysis methods to

construct a tobacco leaf yield prediction model based on climatic factors and found that various climatic factors such as average temperature, accumulated temperature, and precipitation had a significant impact on tobacco leaf yield [13]. Further, He *et al*. applied stepwise regression analysis to construct a flue-cured tobacco yield prediction model based on key meteorological factors with high prediction accuracy [14].

Various statistical analysis methods and machine learning algorithms have been employed to construct different types of prediction models, achieving relatively ideal prediction results. However, most models only target a single tobacco variety or a single chemical component, and their comprehensive prediction ability for multiple varieties and multiple components is weak. This research focused on the development of predictive models for the main chemical components of tobacco leaves in Mile and Luxi counties within the Honghe tobacco cultivation zone in Yunan, China and provided theoretical support for the sustainable development of the local tobacco industry.

Materials and methods

Sample sources and collection

A total of 163 tobacco leaf samples were collected from Mile (87 samples) and Luxi (76 samples) counties within the Honghe tobacco-growing region of Yunnan, China, which included Hongda, K326, and Yunyan 87 three cultivated varieties with 34 , 32, and 10 samples collected in Luxi county and 33, 30, and 24 samples collected in Mile county, respectively. Based on the ecological heterogeneity of tobacco fields (hillside and paddy field areas), combined with the main types of soils including red soil, yellow soil, purple soil and the main cultivated varieties, 164 sampling points were established using stratified sampling methods. Three types of data including tobacco leaf samples, soil samples, and ecological environmental data were collected to obtain multi-dimensional indicators. Middle leaves of tobacco plants (C3F, middle orange grade 3) were selected for tobacco leaf sample collection, while 10 intrinsic chemical composition indicators including total sugar, reducing sugar, total nitrogen, nicotine, potassium (K), chlorine (Cl), petroleum ether extract, starch, nitrogen-to-alkaloid ratio, sugar difference were measured based on commonly used indicators in the tobacco industry [15]. For soil sample collection, representative samples were taken at each sampling point, and key indicators were analyzed. Data of 8 climatic factors and 3 geographical environmental factors were also collected as ecological environmental data. Meteorological sensors were deployed at sampling points in the planting areas to ensure the accuracy and completeness of the data.

Data processing

The collected raw data went through preprocessing including data cleaning, outlier handling, and encoding to enhance data quality and usability for subsequent correlation analysis and prediction model development. The Pandas library (https://pandas.pydata.org/) was used to efficiently complete the data cleaning by filling the existing null values and removing duplicated values. Given the relatively small number of missing values in this research, to maintain data integrity and accuracy, these missing values were directly deleted, thereby further optimizing the dataset and laying a data foundation for subsequent modeling work. Box plots visualization tool was employed for displaying the characteristics of data distribution and played an important role in outlier handling. Through box plots, the central tendency, dispersion, and presence of outliers in the data could be intuitively observed. By applying quartiles and interquartile range to set upper and lower bounds, the outliers could be effectively identified and marked. In this study, soil type, topography, and tobacco type were categorical data with their values presented in character form, which could not be directly applied to modeling analysis. The data needed to be encoded, converting character-based data into numerical data to meet modeling requirements. LabelEncoder (https://scikit-learn.org/stable/

modules/generated/sklearn.preprocessing.LabelEncoder.html) was used to encode categorical data. In this way, character-based data that could not originally be directly involved in modeling could be transformed into numerical form, facilitating subsequent modeling operations and more in-depth data analysis. All data was organized using Microsoft Excel 2016 (https://www.microsoft.com/). Spearman correlation analysis was conducted to explore the associations between the key chemical constituents in tobacco leaves, soil factors, and ecological variables. Multiple linear regression (MLR), decision-trees models (DT), and random forests (RF) were adopted to develop prediction models for tobacco leaf chemical constituents.

Spearman correlation analysis

Spearman correlation coefficient was calculated using rank correlation coefficients as follows [16].

$$\rho = 1 - (6 * \Sigma d\text{^}2) / (n * (n\text{^}2 - 1)) \quad (1)$$

where ρ was the Spearman correlation coefficient. d was the rank difference between X and Y. n was the sample size. The tobacco leaf chemical components, soil data such as soil pH, specific trace element content, *etc*., and ecological environment data such as rainfall, sunshine duration, *etc*. in this research did not all conform to a normal distribution. The Spearman correlation coefficient was calculated based on the ranks of the data, therefore, could effectively measure the correlation between variables for non-normally distributed data.

Multiple linear regression

Multiple linear regression was used to explore the linear connections between several independent variables and a singular dependent variable to leverage data from multiple independent variables to forecast and elucidate the behavior of the dependent variable. Its basic form was shown below [17].

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \varepsilon \quad (2)$$

where Y was the dependent variable. x_1, x_2 ... x_p were the independent variables. β_0 was the intercept, which indicated the value of the dependent variable when all independent variables were set to zero. The coefficients β_1, β_2 ... β_p were the degree of impact each independent variable had on the dependent variable. Additionally, ε was the random error term being presumed to adhere to a normal distribution with a mean of zero. To conduct multiple linear regression, certain assumptions must be met including the presence of a linear relationship, independence of observations, homoscedasticity, normal distribution of residuals, and absence of multicollinearity. The model's fit to the data was assessed through various metrics, one of which was the coefficient of determination (R^2). A value of R^2 approaching 1 indicated a superior fit, implying that a greater proportion of the variance in the dependent variable was accounted for by the independent variables. In this analysis, the Spearman correlation coefficient was utilized to identify essential independent variables, and multiple linear regression was applied to estimate the chemical component concentrations in the tobacco leaves.

Decision tree model

By learning and generalizing from sample data, a decision tree was constructed, which was capable of classifying or predicting unknown data. Feature selection was a key step in building a decision tree to select the most discriminative features from numerous features. Common feature selection metrics included information gain, information gain ratio, Gini index, *etc*. with information gain being used in this study. Information gain represented the extent to which the uncertainty of class Y was reduced by knowing the information of feature X. Let the dataset be D, the number of classes be K, the feature X had n different values, and dividing D into n subsets D_i, the formula for calculating information gain was as follows [18].

$$g(D,X) = H(D) - \sum_{i=1}^{n} \frac{|Di|}{|D|} H(D_i) \quad (3)$$

$$H(D) = -\sum_{K=1}^{K} \frac{|C_k|}{|D|} \log_2 \frac{|C_k|}{|D|} \quad (4)$$

where $H(D)$ was the information entropy of dataset D. $|C_K|$ was the number of samples in D belonging to the k-th class. |D| was the total number of samples. $H(D_i)$ was the information entropy of subset D_i. The larger the information gains, the greater the contribution of that feature to classification. However, decision trees may suffer from overfitting during training, i.e., they fit the training data too precisely, leading to poor performance on test data. Pruning served as a technique to mitigate overfitting by eliminating certain branches, thereby streamlining the decision tree, which was categorized into two types including pre-pruning and post-pruning. Pre-pruning entailed assessing the need for node splitting prior to the actual division during the creation of the decision tree. If the split failed to enhance generalization performance, then it was halted. In contrast, post-pruning involved assessing non-leaf nodes in a bottom-up manner after the decision tree had been constructed. If substituting the subtree linked to the node with a leaf node leads to better generalization performance, the substitution was carried out.

Random forest

Random forest made final decisions by constructing multiple decision trees and integrating their prediction results. Based on the idea of bootstrap aggregating, multiple different bootstrap sample sets were generated by random sampling with replacement from the original training dataset. Each bootstrap sample set was used to train a decision tree, and finally, the results of these decision trees were combined for prediction. Since each sample was random, the training data for each decision tree differed, resulting in different trained decision trees. By combining these different decision trees, the model's variance could be reduced, and its generalization ability could be improved. For regression problems, random forest used the averaging method to obtain the final predicted value as shown below [19].

$$y = \frac{1}{m} \sum_{i=1}^{m} h_i(x) \quad (5)$$

where the final predicted value of the random forest model for the input sample *x* was denoted as *y*. *m* indicated the total number of decision trees within the random forest. Typically, an increase in the number of decision trees (*m*) could lead to improved predictive performance of the model, however, which also resulted in heightened computational expenses and longer training durations. The predicted value for the input sample *x*, as determined by the *i*-th decision tree, was denoted as $h_i(x)$. Random forest could leverage the ensemble advantages of multiple decision trees to effectively predict numerical values of samples in regression tasks.

Results and discussion

Correlation analysis of chemical components of tobacco variety with soil conditions and ecological environment in Luxi county

(1) "Hongda" variety

The heatmap of association analysis between the main chemical components of the "Hongda" variety and ecological environmental factors showed that there were multiple correlations between the chemical elements of the "Hongda" variety and ecological environmental factors. The correlation analysis of tobacco leaves' main chemical components demonstrated that total sugar and starch were correlated with altitude, while starch was correlated with temperature during the tobacco field period, and reducing sugar was correlated with sunshine duration. Total sugar and nitrogen-to-alkaloid ratio were correlated with soil organic matter. Potassium element was correlated with soil hydrolyzable nitrogen. Further, sugar difference was correlated with the available phosphorus of soil,

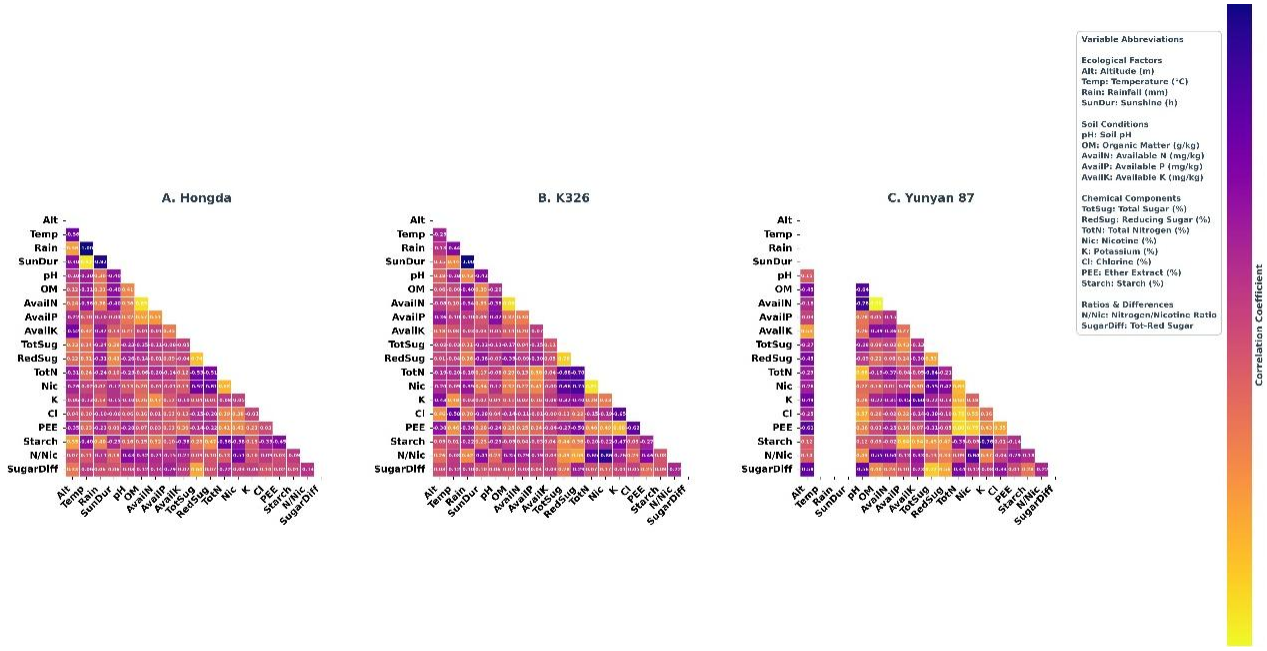

Figure 1. Heatmap of correlations between different tobacco varieties and ecological environmental factors as well as soil conditions in Luxi county.

and starch was correlated with the available potassium of soil (Figure 1A).

(2) "K326" variety

The chemical elements of the "K326" variety exhibited multi-dimensional correlations with ecological environmental factors. Among tobacco leaves' main chemical components, the contents of total nitrogen, nicotine, potassium, chlorine, and petroleum ether extract were correlated with altitude. Chlorine element and starch content were correlated with temperature during the tobacco field period, while nicotine, chlorine, petroleum ether extract, and nitrogen-to-alkaloid ratio were correlated with rainfall. The starch content was correlated with sunshine duration. Potassium element and nitrogen-to-alkaloid ratio were correlated with soil pH. Further, reducing sugar, total nitrogen, nicotine, and nitrogen-to-alkaloid ratio were correlated with soil organic matter, while total nitrogen and nicotine were correlated with soil available phosphorus (Figure 1B).

(3) "Yunyan 87" variety

The results showed that the temperature, rainfall, and sunshine duration of the "Yunyan 87" variety during the field period were constant and did not show a correlation with the main chemical components of "Yunyan 87" variety tobacco leaves. The other factors demonstrated significant correlations. The chemical elements of the "Yunyan 87" variety demonstrated multiple correlations with ecological environmental factors. The main chemical components of tobacco leaves showed that the contents of potassium element, petroleum ether extract, and sugar difference were correlated with altitude. The total nitrogen, chlorine element, petroleum ether extract, nitrogen-to-alkaloid ratio, and sugar difference were correlated with soil pH. Reducing sugar, potassium element, nitrogen-to-alkaloid ratio, sugar difference, and soil organic matter content were correlated. Potassium element and nitrogen-to-alkaloid ratio were correlated with soil hydrolyzable nitrogen and available phosphorus content. Both total and reducing sugars, nicotine, potassium, chlorine, the nitrogen-to-alkaloid ratio, and sugar difference all exhibited associations with the available potassium content in soil (Figure 1C).

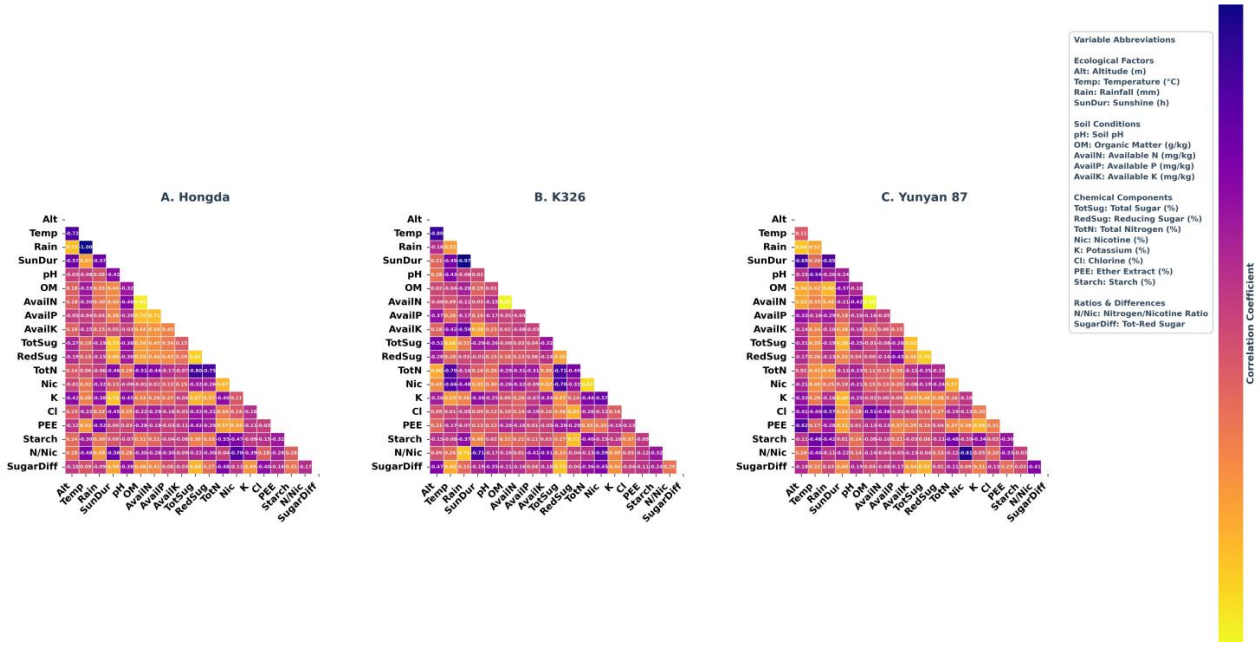

Figure 2. Heatmap of correlations between different tobacco varieties and ecological environmental factors as well as soil conditions in Mile county.

The correlation patterns in Luxi County demonstrated that different tobacco varieties exhibited distinct responses to ecological factors. "Hongda" showed moderate correlations with both soil and meteorological factors, reflecting its relatively balanced environmental adaptability. "K326" presented more extensive correlations, likely due to its genetic sensitivity to environmental changes, which aligned with previous studies on its phenotypic plasticity [20]. "Yunyan 87" relied heavily on soil factors for chemical component regulation, possibly because the stable microclimate masked meteorological factor influences, which emphasized the importance of soil management such as potassium-rich fertilization and pH adjustment for this variety in the region. These findings also highlighted the need for varietal-specific cultivation strategies such as targeted soil nutrient supplementation for "Yunyan 87" and climate-responsive irrigation for "K326".

Correlation analysis of chemical components of tobacco variety with soil conditions and ecological environment in Mile county

(1) "Hongda" variety

The results showed that there was a significant correlation between the chemical elements of the "Hongda" variety and ecological environmental factors. The potassium content was correlated with altitude. Temperature during the field period affected the content of various elements such as total sugar and reducing sugar. Rainfall and sunshine duration were both associated with changes in the content of total sugar and reducing sugar. Soil pH was correlated with the content of total sugar, while soil organic matter content was associated with the content of multiple elements. Hydrolyzable nitrogen affected the content of total sugar, while available phosphorus was related to the content of reducing sugar, and reducing sugar was also correlated with available potassium (Figure 2A).

(2) "K326" variety

The chemical elements of the "K326" variety exhibited multi-faceted correlations with ecological environmental factors. Total sugar, total nitrogen, nicotine, potassium, and sugar difference were correlated with altitude and temperature during the field period. Starch and nitrogen-to-alkaloid ratio were correlated with

rainfall. Total sugar, total nitrogen, and various other elements were correlated with sunshine duration. Sugar difference was correlated with soil pH. Total nitrogen and nicotine were correlated with hydrolyzable nitrogen, while total nitrogen and starch were correlated with available phosphorus, and nicotine and nitrogen-to-alkaloid ratio were correlated with available potassium (Figure 2B).

(3) "Yunyan 87" variety

There was a wide range of correlations between the chemical elements of the "Yunyan 87" variety and ecological environmental factors. Total sugar, total nitrogen, nicotine, potassium, and sugar difference were correlated with altitude and temperature during the field period. Starch and nitrogen-to-alkaloid ratio were correlated with rainfall. Total sugar and various other elements were correlated with sunshine duration. Sugar difference was correlated with soil pH. Total nitrogen and nicotine were correlated with hydrolyzable nitrogen, while total nitrogen and starch were correlated with available phosphorus, and nicotine and nitrogen-to-alkaloid ratio were correlated with available potassium (Figure 2C).

The correlation patterns of tobacco leave in Mile county differed notably from those in Luxi county, particularly for "Yunyan 87", which showed significant responses to meteorological factors. This regional variation suggested that "Yunyan 87" had strong environmental adaptability, adjusting its chemical component regulation mechanisms based on local climate variability. For "Hongda" and "K326", the combined effects of altitude, temperature, and soil nutrients of nitrogen and phosphorus on chemical components highlighted the need for integrated management strategies such as altitude-specific planting plans and balanced nutrient supply to optimize tobacco quality. The enhanced correlation between "K326" and altitude in Mile County also implied that this variety might be more suitable for high-altitude areas in the region, where its nicotine accumulation potential could be maximized. These results underscored the importance of considering both regional and varietal differences in ecological factor interactions when formulating tobacco cultivation practices.

Prediction model for chemical components of tobacco leaves in Luxi county

(1) Prediction model for "Hongda" variety

The results demonstrated that different models performed differently in predicting various chemical elements. The decision tree model showed significant advantages in predicting starch content with a test set R^2 value of 0.829, while random forest was 0.781, and multiple linear regression was 0.703. The prediction effect for nicotine content was poor with the R^2 values of the test set for all three models being negative. For petroleum ether extract prediction, multiple linear regression had a relative advantage, but the R^2 values of the test set was only 0.527. In predicting the nitrogen-to-alkaloid ratio, the multiple linear regression and decision tree models had similar performance with test set R^2 values of 0.505 for both (Figure 3A).

(2) Prediction model for "K326" variety

The prediction results for total nitrogen content were ideal with the random forest test set R^2 at 0.887, multiple linear regression at 0.881, and decision tree at 0.839. For nicotine content prediction, random forest performed excellently with a test set R^2 of 0.93, while multiple linear regression also performed well at 0.784. For starch content prediction, the decision tree performed well with a test set R^2 of 0.829, and random forest at 0.702. For petroleum ether extract prediction, the decision tree test set R^2 was 0.72. For chlorine element prediction, the random forest test set R^2 was 0.712 (Figure 3B).

(3) Prediction model for "Yunyan 87" variety

For sugar difference prediction, the decision tree test set R^2 was 0.972, random forest was 0.952, and multiple linear regression was 0.835 with all three models showing good fit. For potassium element prediction, the multiple linear regression model test set R^2 value was 0.949. For reducing sugar prediction, the random forest

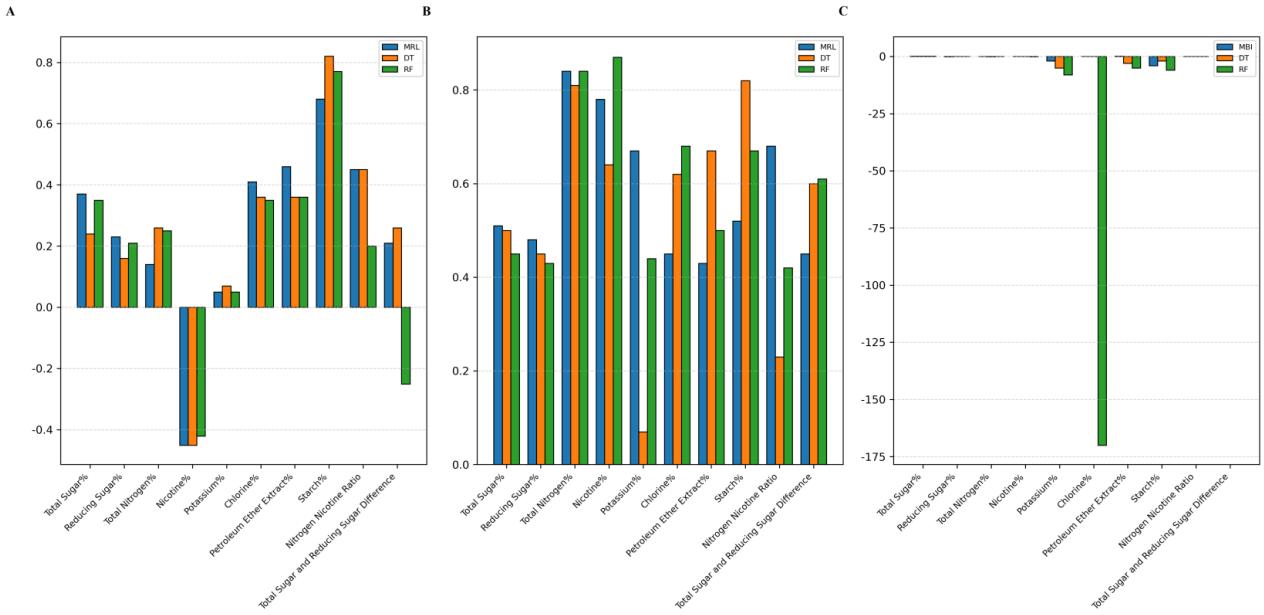

Figure 3. Performance comparison (R^2) of MLR, DT, and RF models for different tobacco varieties in Luxi County. **A.** "Hongda". **B.** "K326". **C.** "Yunyan 87".

model performed excellently with a test set R^2 value reaching 0.999. However, the chlorine element prediction was abnormal with the random forest model test set R^2 value as low as -150 (Figure 3C).

In Luxi county, the prediction model performance highlighted clear varietal and component-specific patterns. "Hongda" benefited from decision tree models for starch prediction, likely due to the model's ability to capture non-linear relationships between starch and ecological factors, while its poor nicotine prediction suggested unaccounted factors such as cultivation practices. "K326" achieved high accuracy with random forest for nicotine and total nitrogen, reflecting the model's strength in handling complex, multi-factor interactions. "Yunyan 87" showed exceptional performance for sugar difference, potassium, and reducing sugar, indicating strong linear or non-linear correlations with selected ecological factors, though abnormal chlorine prediction warranted further data preprocessing. These results suggested that model selection should be tailored to both variety and chemical components with random forest as a versatile choice for most components and multiple linear regression for those with strong linear trends.

Prediction model for chemical components of tobacco leaves in Mile county

(1) Prediction model for "Hongda" variety

The results demonstrated that different models showed varying advantages and disadvantages in predicting various chemical elements. The prediction effect for potassium element was significant with the decision tree model test set R^2 reaching 0.889, random forest at 0.852, and multiple linear regression exceeding 0.5. In total sugar prediction, the decision tree and random forest models performed acceptably with R^2 values of 0.77 and 0.763, respectively, while the two models performed similarly with R^2 values around 0.668 for sugar difference prediction. However, the decision tree had negative R^2 values for reducing sugar, chlorine, and starch prediction. For total nitrogen prediction, the R^2 values for all three models were close to 0. For nicotine prediction, the decision tree and random forest were superior to multiple linear regression. Random forest performed outstandingly for petroleum ether extract prediction, while multiple linear regression had

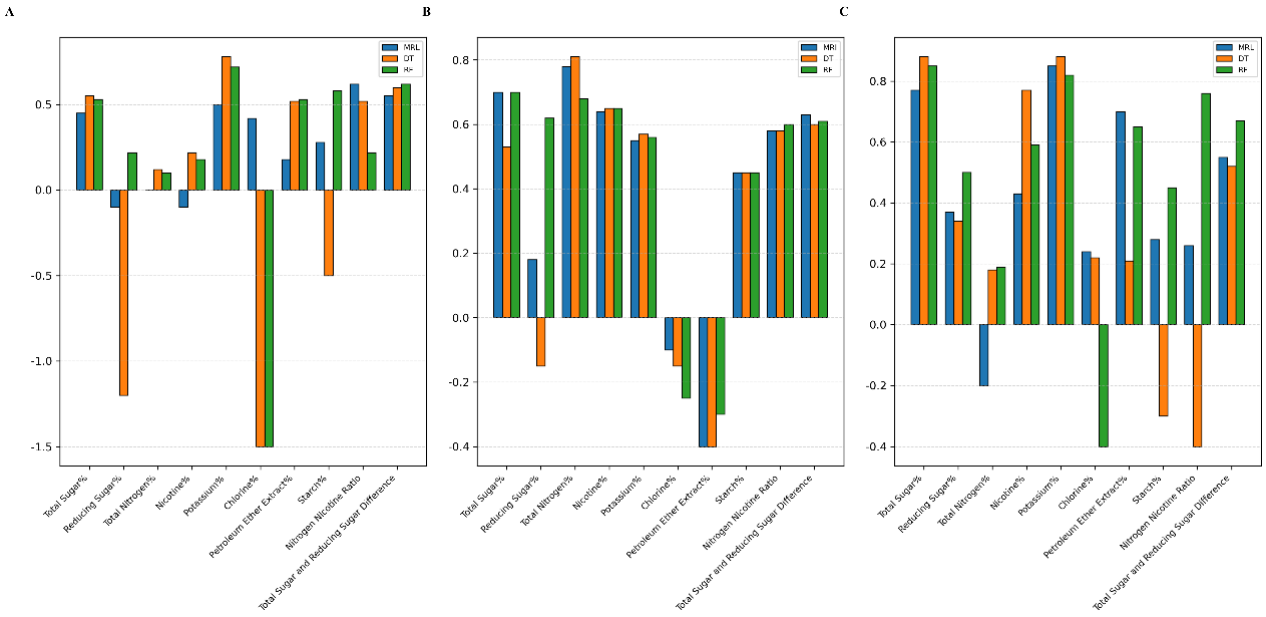

Figure 4. Performance comparison (R^2) of MLR, DT, and RF models for different tobacco varieties in Mile county. **A.** "Hongda". **B.** "K326". **C.** "Yunyan 87".

an R^2 of 0.701 for nitrogen-to-alkaloid ratio prediction, showing excellent performance (Figure 4A).

(2) Prediction model for "K326" variety

The prediction effect for total nitrogen was good with the decision tree model test set R^2 at 0.805, multiple linear regression at 0.778, and random forest at 0.707. In total sugar prediction, the random forest and decision tree models had test set R^2 values of 0.708. For the prediction of reducing sugar, nicotine, potassium element, nitrogen-to-alkaloid ratio, and sugar difference, some models had test set R^2 values greater than 0.6. However, the prediction effects for chlorine element and petroleum ether extract were poor with negative test set R^2 values (Figure 4B).

(3) Prediction model for "Yunyan 87" variety

The prediction effects for total sugar and potassium element were relatively good. In total sugar prediction, the decision tree model test set R^2 was 0.931, random forest was 0.888, and multiple linear regression was 0.756. For potassium element prediction, the test set R^2 values for the three models were 0.879 for multiple linear regression, 0.9 for decision tree, and 0.866 for random forest. For nicotine prediction, the decision tree model had a relative advantage with a test set R^2 value of 0.78. For petroleum ether extract prediction, multiple linear regression performed well with a test set R^2 value of 0.747 followed by random forest, while the decision tree performed poorly with a test set R^2 value close to 0.2. For nitrogen-to-alkaloid ratio prediction, the random forest test set R^2 value was 0.739, while the decision tree test set R^2 value was below -0.4 (Figure 4C).

In Mile county, the prediction model performance exhibited distinct regional characteristics compared to Luxi county. "Hongda" benefited from decision tree models for potassium prediction, aligning with its strong correlation with ecological factors in the region, while its poor performance for reducing sugar and starch suggested higher ecological complexity. "K326" maintained high accuracy for total nitrogen across models, confirming the robustness of ecological factor explanations, though poor chlorine and petroleum ether extract prediction might stem from unmodeled

external factors [21]. "Yunyan 87" showed exceptional total sugar and potassium prediction with decision tree and multiple linear regression, respectively, reflecting its adaptability to Mile's environmental conditions. These findings reinforced the need for regional calibration of prediction models and highlighted random forest as a reliable choice for most components, while decision trees exceled in capturing non-linear relationships for specific varieties such as "Yunyan 87" in total sugar prediction.

The chemical elements of different tobacco varieties in Luxi County and Mile County exhibit extensive correlations with ecological environmental factors [22]. Commonalities included the influence of altitude and key soil attributes such as pH, organic carbon, and nutrients including hydrolyzable nitrogen, available phosphorus, and potassium on multiple chemical elements [23, 24]. However, differences existed between the two locations. For "Yunyan 87" in Luxi county, due to data characteristics, temperature during the field period, rainfall, and sunshine duration showed no correlation with tobacco leaves' main chemical components, whereas this variety in Mile county did show the correlations. Different varieties had unique correlation patterns. "Hongda" in Luxi county demonstrated that starch was correlated with temperature during the field period, while in Mile county, temperature during the field period affected total sugar and other elements. "K326" also showed different correlations in the two locations. Ecological environmental factors intertwined to influence tobacco chemical elements with soil conditions and meteorological factors acting synergistically, affecting tobacco growth and element accumulation. Based on correlation analysis, this study employed three modeling approaches to construct prediction models for tobacco leaf chemical elements [25]. Different models showed significant advantages in predicting specific elements in both counties. Overall, random forest and multiple linear regression performed better for major categories of elements such as sugars and potassium and data with high linear correlation [26]. The comprehensive analysis of the models' performance on different elements, regions, and varieties provided important insights for model selection and optimization. These research findings provided an important scientific basis for tobacco cultivation. By understanding the correlation between chemical components and the ecological environment, planting environments could be optimized in a targeted manner to improve tobacco quality. Meanwhile, the comparison results of the models helped in selecting appropriate prediction models, providing more precise decision support for tobacco production management. However, this study still had certain limitations. Some models performed poorly in predicting certain elements, which might be related to data quality, model selection, and parameter settings. In addition, the study only covered specific varieties in Luxi county and Mile county, and the universality of the research results needed further verification. Future research could expand the scope of the study, collect data from more regions and varieties, optimize data processing and model construction methods, and explore more deeply the relationship between tobacco chemical components and the ecological environment, providing a more solid theoretical foundation for the sustainable development of the tobacco industry. For elements and scenarios with good prediction results of this research, further research should investigate the reasons for the model's success, summarize experiences, and promote their application. For cases with poor prediction results, future in-depth analysis research should address data quality issues such as outliers or missing data by optimizing data preprocessing methods, trying new models, or improving existing models to enhance prediction accuracy and reliability.

Acknowledgements

This study was supported by the Scientific Research Foundation Project of the Yunnan Tobacco Company Science and Technology Plan Project (Grant No. 20235300000241020) and

Lijiang City Company Science and Technology Plan Project (Grant No. 2024530700242004).

References

1. Ke L, Zhao X, Hu Y, Lan L, Zhang X, Dong G, *et al*. 2023. Relationship between soil factors and intrinsic quality of tobacco leaves in Longyan tobacco-growing area and construction of its evaluation model. Jiangxi Agricultural Journal. 35(8):36-41.
2. Zhang M, Liu L, Zhao X, Wu L, Zhang Y, Lin Z, *et al*. 2020. Construction of prediction model for chemical components of tobacco leaves based on BP neural network. Guizhou Agricultural Sciences. 48(2):136-139.
3. Zhang X, Yang C, Liu H, Wu W. 2022. Prediction of soil organic matter and total nitrogen content in tobacco-planting areas based on machine learning. Tobacco Science & Technology. 55(8):20-27.
4. Liu J, Chen Z, Liu Y, Li J, Den P. 2023. Study on prediction model of nicotine content in flue-cured tobacco in "Qingjiangyuan" tobacco area based on meteorological factors. Meteorology and Environmental Sciences. 46(6):32-38.
5. Xia X, Li X, Liu T, Zeng L, Xu J, Wang J, *et al*. 2024. Study on single leaf weight model of flue-cured tobacco in western Guizhou province based on multiple meteorological elements. Southwest China Journal of Agricultural Sciences. 37(8):1850-1861.
6. Chen F, Jing Y, Xie X, Yang J. 2022. Prediction and analysis of physical characteristic indicators of high-quality light flavor high-quality tobacco. Science Technology and Engineering. 22(32):14159-14166
7. Huang Y, Du G, Ma Y, Zhou J. 2021. Predicting heavy metals in dark sun-cured tobacco by near-infrared spectroscopy modeling based on the optimized variable selections. Ind Crops Prod. 172:114003.
8. Zhang F, Ye L, Li D, Wu X. 2023. Prediction model of oil content grade in flue-cured tobacco based on LBP-GLCM feature fusion. Computer and Digital Engineering. 51(7):1524-1528.
9. Chen Y, Ren K, He X, Chen Y, Hu B, Hu X, *et al*. 2020. The response of flue-cured tobacco cultivar K326 to nitrogen fertilizer rate in China. J Agri Sci. 158(5):371-382.
10. Zhou X, Li X, Liu Z, Zhou X, Qiu C, Zhuang Z, *et al*. 2024. Relationship between sensory quality and chemical components of Shandong flue-cured tobacco and establishment of quality prediction model. Chinese Agricultural Science Bulletin. 40(1):128-134.
11. Li X, Xia X, Liu Y, Liu T, Zeng L, Chen L, *et al*. 2024. Comparison of single leaf weight models for flue-cured tobacco in central and eastern Guizhou based on meteorological elements. Chinese Journal of Agrometeorology. 45(9):1012-1026.
12. Han Y, Liu C, Hou Q, He J, Shan S, Zhang L, *et al*. 2024. Meteorological factor analysis and model construction for the occurrence of tobacco black shank and wildfire diseases in Honghe tobacco area. J Biol Disaster Sci. 47(1):101-110.
13. Qi X, Wen M, Huang C, Wang X, Gao J, Cui S, *et al*. 2022. Construction of tobacco leaf yield prediction model based on climatic factors in Zunyi tobacco area by multivariate statistical analysis. Southern Agriculture. 16(9):27-30.
14. He N, Fan Y, Yuan X, Zhang M, Zhang S. 2023. Construction of yield forecast model for flue-cured tobacco in Xiangxi based on key meteorological factors. Chinese Agricultural Science Bulletin. 39(24):96-102.
15. Wang C, Ma S, Wang Z, Liu Q, Li S, Zhao L, *et al*. 2025. Construction of discriminative models for flue-cured tobacco from eight ecological areas based on fisher discriminant analysis and tobacco chemical components. Tobacco Science & Technology. 58(2):1-10
16. Teng L, Jiang G, Ding Z, Wang Y, Liang T, Dai H, *et al*. 2024. Evaluation of tobacco-planting soil quality using multiple distinct scoring methods and soil quality indices. J Clean Prod. 441:140883.
17. Li J, Zhang Q, Li M, Yang X, Ding J, Huang J, *et al*. 2023. Multi-factor correlation analysis of the effect of root-promoting practices on tobacco rhizosphere microecology in growth stages. Microbiol Res. 270:127349.
18. Wang B, Cai J, Zhao L, Shang S, Zhang S, Li Y, *et al*. 2025. Volatile profiles of different tobacco cultivars and their correlation with sensory quality of heated tobacco. Curr Anal Chem. 21(4):345-355.
19. Zhang J, Wang Y, Zhu Y, Su M, Xie Y, Zheng Y, *et al*. 2024. Pyrolysis characteristics of different types of tobacco and its correlation with the released aroma components under heating condition. J Anal Appl Pyrolysis. 181:106646.
20. Sun T, Xue C, Chen Y, Zhao L, Qiao C, Huang A, *et al*.2023. Cost-effective identification of the field maturity of tobacco leaves based on deep semi-supervised active learning and smartphone photograph. Comput Electron Agric. 215:108373.
21. Sinha K, Ghosh N, Sil PC. 2025. Harnessing machine learning in contemporary tobacco research. Toxicol Rep. 14:101877.
22. Fekhar M, Daghbouche Y, Bouzidi N, El Hattab M. 2024. Rapid assessment of smokeless tobacco quality parameters using ATR-FT-MIR spectroscopy: Comparison of analytical/mathematical and machine learning approaches. Microchem J. 201:110670.
23. Nghiem DT, Vu HTT, Nguyen NV, Le CTT. 2024. Growth, yield and quality variability of flue-cured tobacco in response to soil and climatic factors in Northern Vietnam. Ital J Agron. 19(3):100016.
24. Zohar I, Ganem HE, DiSegni DM, Levi AJ. 2024. The impact of alternative recycled and synthetic phosphorus sources on plant growth and responses, soil interactions and sustainable agriculture - lettuce (*Lactuca sativa*) as a case model. Sci Total Environ. 948:174719.
25. Arredondo MG, Fang Y, Jones M, Yabusaki S, Cardon Z, Keiluweit M. 2023. Resolving dynamic mineral-organic interactions in the rhizosphere by combining in-situ microsensors with plant-soil reactive transport modeling. Soil Biol Biochem. 184:109097.
26. Ikram M, Xiao J, Li R, Xia Y, Zhao W, Yuan Q, *et al*. 2022. Identification of superior haplotypes and candidate genes for yield-related traits in tobacco (*Nicotiana tabacum* L.) using association mapping. Ind Crops Prod. 189:115886.