



International Journal of Computational Science and Engineering

ISSN online: 1742-7193 - ISSN print: 1742-7185

<https://www.inderscience.com/ijcse>

A novel fertility intention prediction scheme based on Naive Bayes

Meijiao Zhang, Lan Yang, Weiping Jiang, Gejing Xu, Guoliang Hu

DOI: [10.1504/IJCSE.2025.10068153](https://doi.org/10.1504/IJCSE.2025.10068153)

Article History:

Received:	13 July 2022
Last revised:	09 August 2022
Accepted:	20 September 2022
Published online:	21 December 2024

A novel fertility intention prediction scheme based on Naive Bayes

Meijiao Zhang and Lan Yang*

School of Software,
Quanzhou University of Information Engineering,
Quanzhou, Fujian, China
Email: 80022046@qq.com
Email: yanglan@qziedu.cn
*Corresponding author

Weiping Jiang

School of Computer Science and Engineering,
Hunan University of Science and Technology,
XiangTan, Hunan, China
Email: 2593386261@qq.com

Gejing Xu

School of Software,
Quanzhou University of Information Engineering,
Quanzhou, Fujian, China
Email: 21004775@qq.com

Guoliang Hu

Hunan Branch of National Computer Network Emergency Response,
Technical Team/Coordination Center of China,
Changsha, Hunan, China
Email: hnuhgl@hnu.edu.cn

Abstract: In today's society, the problem of ageing has developed into a global problem. The ageing of the population is related to the decrease in fertility intention. Therefore, it is meaningful to predict the fertility intention of people. In this paper, we propose a fertility intention prediction scheme based on a polynomial Naive Bayesian model, called the FPB scheme. In the proposed scheme, we first extract various features from the data we collected, and we divide the entire dataset into three labels according to the level of fertility intention. Then, we use these data to construct a classifier. Next, we use this classifier to design a perfect prediction algorithm to predict fertility intention. Finally, we conduct extensive experiments to evaluate the performance of the proposed scheme. The experimental results show that the proposed FPB scheme has high accuracy and can help families make accurate fertility decisions.

Keywords: fertility intention; polynomial Naive Bayes; prediction algorithm.

Reference to this paper should be made as follows: Zhang, M., Yang, L., Jiang, W., Xu, G. and Hu, G. (2025) 'A novel fertility intention prediction scheme based on Naive Bayes', *Int. J. Computational Science and Engineering*, Vol. 28, No. 1, pp.1–9.

Biographical notes: Meijiao Zhang received her BS from the Fujian Normal University, China, in 2006, and MSc from the Huaqiao University, China, in 2015. She is currently a Lecturer at the Quanzhou University of Information Engineering, China. Her research interests include machine learning, artificial intelligence, etc.

Lan Yang received her BS from the Huaqiao University, China, in 2004, and MSc from the Huaqiao University, China, in 2013. She is currently a Lecturer at the Quanzhou University of Information Engineering, China. Her research interests include data classification, data mining, etc.

Weiping Jiang is currently pursuing his Bachelor's at the School of Computer Science and Engineering, Hunan University of Science and Technology, China. His current research interests include data mining and big data analytics.

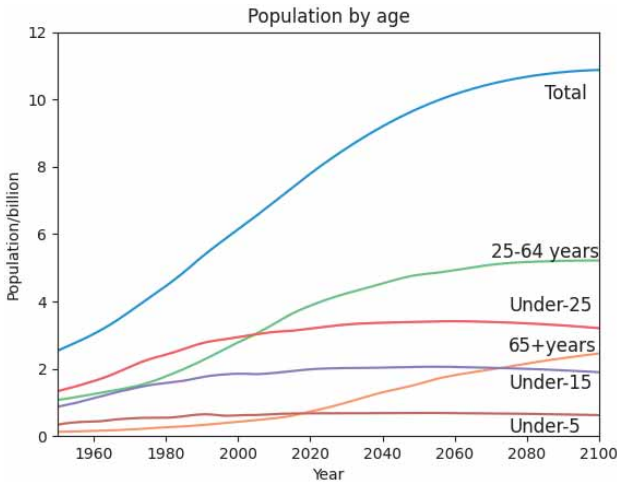
Gejing Xu received her BS in 2004 and Master's in 2014 from the Huaqiao University. She is currently a Lecturer at the Quanzhou University of Information Engineering. Her current research interests include theories of software security, big data analytics and blockchain.

Guoliang Hu received his BS from the Yancheng Institute of Technology, China, in 2010, and MSc from the Hunan University, China, in 2013. He is currently a Senior Engineer at the Hunan Branch of National Computer Network Emergency Response, Technical Team/Coordination Center of China, China. Her research interests include artificial intelligence and network security.

1 Introduction

With the continuous improvement of living standards and the advancement of medical technology, the mortality rate in the world continues to decrease. People's lives become longer. However, according to the statistics from relevant agencies, the global birth rate is declining (World Birth Rate 1950–2022, <https://www.macrotrends.net/countries/WLD/world/birth-rate>), which indicates the proportion of the elderly population is rising and the population is ageing.

Figure 1 Population by Age Bracket with UN Projections, World (see online version for colours)



The number of people in each age group is shown in Figure 1 (Population by Age Bracket with UN Projections, World, <https://ourworldindata.org/grapher/historic-and-un-pop-projections-by-age>). As we can see from Figure 1, the number of teenagers is growing slowly, while the number of seniors over the age of 65 is growing rapidly. Based on Figure 1, we can find that the number of young people may continue to decline in the future, and by around 2070, the number of elderly people will exceed the number of under 15-year-old teenagers. The ageing of the world population at that time will be very serious. We all know that the elderly are mostly incapable of working and make limited contributions to society. Teenagers are the future of the world and they can contribute to the development of the whole world. The ageing of the world's population will slow down the development of the entire world. Therefore, it is urgent to solve the problems of the ageing population in the world.

In order to promote the long-term balanced development of the population, many countries around the world are working hard to propose various policies to stimulate the birth rate. For example, if you have a child in the USA, you can get a tax refund. Childbirth expenses are tax-deductible. The expenses incurred by children going to school can be used for a tax deduction. Some families can even get three or four thousand dollars in tax rebates a year (Wang and Sun, 2017).

Despite this, the ageing of the world's population is still accelerating, and the ageing process has not been alleviated. Through consulting relevant information, we can find that the problem of ageing is related to many factors. A very important factor is that people's fertility intention has changed. People nowadays are under a lot of pressure and are willing to pursue freedom and do not want to be bound by children. As a result, many people choose not to have children. This has led to a decline in the overall birth rate across society and a decline in the average annual growth rate. People's fertility intentions are low. Fertility intentions are complex, diverse, and related to many factors (Huang et al., 2021). Therefore, how to predict the fertility intention accurately is still challenging.

To address the challenge, many researchers have proposed different classification prediction algorithm (Benkhelifa et al., 2020; Zhu et al., 2020), such as decision trees classification (Jiang and Wang, 2021), the k -nearest neighbour algorithm (Shihao, 2017), support vector machine for classification (Wang et al., 2021), logistic regression (Mai et al., 2020), etc. (Geler et al., 2021; Wang et al., 2020). However, these algorithms cannot be directly applied to the prediction of fertility intention. Recently, Wang (2017) and Dai and Wang (2020) proposed a prediction of second-child fertility intention algorithm based on data mining. This paper uses decision trees, support vector machines, random forests, and logistic regression model to predict fertility intention and compare the accuracy of these algorithm predictions. But this paper only considered binary classification, i.e., high and low fertility intentions. Subsequently, Zhang et al. (2021) proposed a fertility intention prediction and analysis algorithm based on a random forest. Nevertheless, the random forest algorithm may not produce good classifications for data with few features and is sensitive to the missing data. The random forest algorithm may not be accurate for prediction results with a small number of samples. Thus, it is necessary to devise a perfect fertility

intention prediction scheme to help families make accurate fertility decisions.

In this paper, we propose a fertility intention prediction scheme based on polynomial Naive Bayesian model, called the FPB scheme. Different from the existing schemes, our proposed scheme is the first to use Naive Bayesian theory to predict fertility intentions and takes into account multivariate classification, and is insensitive to the missing data. The polynomial Naive Bayes algorithm is a classification method based on probability theory. Ideas based on probability theory make Naive Bayesian theory very intuitive and easy to realise. We can obtain intuitive probability estimates when performing classification, which makes it more reliable than other algorithms. The polynomial Naive Bayes classification prediction model defaults that each feature is independent of the others, which simplifies the time complexity of the classification prediction model and improves its efficiency. Meanwhile, our proposed scheme can also have a good accuracy for samples with few features. Furthermore, the industrial significance of our proposed scheme is also important. For example, the proposed scheme can be applied to a real-world employment rate problem. Accurate prediction of employment rate can help the government to change employment policy. At the same time, employees can also understand the current employment situation of their own country. The main contributions of this paper are listed as follows:

- In response to the problem of people's low fertility intention, this paper studies various factors that affect people's low fertility intention. These factors include gender, age, average annual family income, and so on.
- Based on these factors, a novel Naive Bayesian classification prediction scheme is constructed. The classifier prediction scheme can make an accurate prediction of fertility intention and this prediction result can help people make auxiliary decisions on fertility issues.
- we conduct extensive experiments to demonstrate the accuracy and efficiency of FPB scheme in actual applications.

The reminder of this paper is organised as follows. In Section 2, we formalise a brief introduction to the basic principles of polynomial Bayesian classifier. We design the prediction algorithm based on polynomial Naive Bayes classification in detail in Section 3. In what follows, we conduct performance analysis and evaluation of the proposed scheme in Section 4. Finally, Section 5 gives the conclusion of this paper and points out our next research work in the future.

2 Basic principles of polynomial Bayesian classifier

In this section, we will introduce the basic principles of the polynomial Bayesian classifier. In order to describe the

steps of creating a classifier conveniently in this paper, we first elucidate a list of notations. Then we will introduce the theory of polynomial Naive Bayes and the polynomial Naive Bayes classifier.

2.1 Notations

- Y : The label variable, where s represents the number of different labels. $Y = [y_1, y_2, \dots, y_s]$.
- X : The dataset used in our experiments, where n represents the number of samples. $X = [x_1, x_2, \dots, x_n]$, $x_i = [a_{i1}, a_{i2}, \dots, a_{im}, y_i]$. y_i represents x_i belongs to label y_i , $y_i \in Y$. a_{ij} represents the value of each feature. m represents the number of features. If y_i is missing, then x_i is a sample to be predicted.
- Y' : The label variable after being transformed, where s represents the number of different labels. $Y' = [y'_1, y'_2, \dots, y'_s]$.
- X' : The dataset used in our experiments after being transformed, where n represents the number of samples. $X' = [x'_1, x'_2, \dots, x'_n]$, $x'_i = [a'_{i1}, a'_{i2}, \dots, a'_{im}, y'_i]$. y'_i represents x'_i belongs to label y'_i , $y'_i \in Y'$. a' represents the value of each feature. m represents the number of features. If y'_i is missing, then x'_i is a sample to be predicted.

2.2 The theory of polynomial Naive Bayes

Bayes's theorem assumes that the features are independent of each other. We can calculate the maximum posterior probability based on the polynomial Naive Bayesian model.

Given a sample dataset X , x_i is a sample to be predicted. The feature attribute set corresponding to the sample data is $x_i = [a_{i1}, a_{i2}, \dots, a_{im}]$ and the label variable is y_j . The prior probability of y_j is $P_{priori} = P(y_j)$. The posterior probability of y_j is $P_{post} = P(y_j|x_i)$. We can use the prior probability $P_{priori} = P(y_j)$, $P(x_i)$, and $P(x_i|y_j)$ to calculate the posterior probability $P(y_j|x_i)$. The formula for calculating the posterior probability is shown in equation (1).

$$P(y_j|x_i) = \frac{P(y_j)P(x_i|y_j)}{P(x_i)} \quad (1)$$

$$P(x_i) = \prod_{k=1}^m P(a_{ik}) \quad (2)$$

$$P(x_i|y_j) = \prod_{k=1}^m P(a_{ik}|y_j) \quad (3)$$

Therefore, the posterior probability of y_j can be calculated as equation (4).

$$P(y_j|a_{i1}, a_{i2}, \dots, a_{im}) = \frac{P(y_j) \prod_{k=1}^m P(a_{ik}|y_j)}{\prod_{k=1}^m P(a_{ik})} \quad (4)$$

Table 1 Part of the data

<i>Gender</i>	<i>Age</i>	<i>Education</i>	<i>Income</i>	<i>Children</i>	<i>Distance</i>	<i>Occupation</i>	<i>Intention</i>
Male	35	University	7	1	Close	Corporate employees	Low
Male	29	Postgraduate	5	1	A bit far	Medical workers	Low
Female	30	University	8	1	Very far	Government institutions	Medium
Female	40	Middle school	4	2	Close	Self-employed	High
Male	35	University	7	2	Close	Teachers	Medium
Male	27	PhD	10	0	Very far	Students	Low
Female	43	Postgraduate	4	1	Close	Teachers	High
Male	44	Elementary school	3	1	Close	Migrant workers	Low
Female	25	University	7	0	A bit far	Others	High

Because the value of $\prod_{k=1}^m P(a_{ik})$ is fixed, when we need to compare the posterior probabilities we only need to compare the molecular of equation (4).

2.3 The polynomial Naive Bayes classifier

The general process of the polynomial Naive Bayes classifier is described as follows.

- Step 1 We collect training datasets and testing datasets.
- Step 2 Most of the type values of data that we collect are strings, so we need to digitise the data for the convenience of fertility intention prediction.
- Step 3 We calculate the conditional probability of each independent feature based on the training datasets.
- Step 4 We use the testing dataset for testing. After the testing, we calculate the accuracy of the testing data.
- Step 5 After we reach a relatively high accuracy, we can use the algorithm to make a prediction.

3 Fertility intention prediction scheme based on the Bayesian model

In this section, we utilise the polynomial Naive Bayes classifier to design our proposed FPB scheme. Firstly, we will give a formal definition of the FPB scheme. And then we will construct the proposed FPB scheme, including LoadData, DataProc, DataTrain, DataTest, and DataPredict.

3.1 The formal definition of FPB scheme

The proposed scheme is composed of five algorithms, namely: LoadData, DataProc, DataTrain, DataTest, and DataPredict.

- $(X_r, X_t) \leftarrow \text{LoadData}(X)$: It takes as input the original datasets X and outputs the initial training datasets X_r and initial testing sets X_t .

- $(X'_r, X'_t) \leftarrow \text{DataProc}(X_r, X_t)$: It takes as input the initial training datasets X_r , initial testing sets X_t and outputs the new training sets X'_r and testing sets X'_t .
- $(Pb) \leftarrow \text{DataTrain}(X'_r)$: It takes as input the new training sets X'_r and it outputs the predicted label Pb .
- $(Ac) \leftarrow \text{DataTest}(X'_t)$: It takes as input the new testing sets X'_t and outputs the accuracy Ac .
- $(Pb) \leftarrow \text{DataPredict}(X'_t)$: It takes as input the set of samples to be predicted X'_t and outputs the predicted label Pb .

3.2 Construction of FPB scheme

3.2.1 LoadData

3.2.1.1 Preparing data

We collected datasets from the whole society through questionnaires. The values of these data are real. They show the real situation of people's factors affecting their fertility intention. Therefore these data can be used for fertility intention prediction. By setting a certain ratio, we will divide the whole raw data into the training datasets and the testing datasets for our subsequent experiments. Part of the data is shown in Table 1.

3.2.1.2 Feature and category extraction

Fertility intention is related to many aspects. For example, in general, the fertility intention of males is lower than the fertility intention of females, and families with higher income may have a higher fertility intention than families with lower income. There are also some other common influencing factors, such as age, education, number of children, etc. With investigation, we have obtained some factors that mainly affect fertility intention. We take these influencing factors as the feature sets of the classifier.

The following feature sets and label sets are used in our proposed scheme. Feature sets include gender, age, education background, average annual family income, number of children, the distance of residence to school,

and occupation. Label sets consist of low fertility intention, medium fertility intention, and strong fertility intention.

Table 2 Part of the processed data

Gender	Age	Education	Income	Children	Distance	Occupation	Intention
0	1	2	2	1	0	1	0
0	0	3	1	1	1	4	0
1	0	2	2	1	2	0	1
1	1	1	1	2	0	3	2
0	1	2	2	2	0	2	1
0	0	4	3	0	2	6	0
1	2	3	1	1	0	2	2
0	2	0	1	1	0	5	0
1	0	2	2	0	1	7	2

3.2.2 DataProc

The type value of sample data exists in the form of strings. Before using the Naive Bayes classifier model, we first need to convert the type value of sample data into integer. The operation of processing the data makes our experiment more convenient, and it does not affect the accuracy of the experiment.

- **Features:** For gender, we can use '0' to represent male and '1' to represent female. For age, we can use '0' to represent age between 20 and 30 years old, '1' to represent between 31 and 40 years old, and '2' to represent more than 40. For education, '0' represents the elementary school, '1' represents the middle school, '2' represents the university, '3' represents postgraduate, and '4' represents PhD with regard to average family income, we can use '0' to represent income less than 30,000, '1' to represent income between 30,000 to 60,000, '2' to represent income between 60,000 to 90,000, '3' to income between 90,000 to 120,000, and '4' to represent income more than 120,000. For the number of children, the feature value can be set to 0, 1, or 2. For the distance of residence to schools, we can use '0' to represent close, '1' to represent a little far and '2' to represent very far. With respect to occupation, '0' represents government institutions, '1' represents corporate employees, '2' represents teachers, '3' represents self-employed, '4' represents medical workers, '5' represents migrant workers, '6' represents students and '7' represents others.
- **Label (fertility intention):** We use '0' to represent low fertility intention, '1' to represent medium fertility intention, and '2' to represent high fertility intention.

After we finished the above process, we convert the string data into integer data. Part of the processed data is shown in Table 2. Thus, after we run the *DataProc* algorithm, we can get the new training sets and testing sets.

3.2.3 DataTrain

This algorithm mainly calculates the conditional probability of each feature at a specific value through the processed

training set. For example, if we find that women have a high fertility intention, the conditional probability of high fertility intention is the highest when the gender is female. Based on the equation (1), $P(x_i)$ is fixed, we only need to compare the numerators. For multiple features, we take the features A , B , and C as an example. Assuming that these features are independent of each other, according to the Naive Bayes, we can get:

$$P(A, B, C|I)P(I) = P(A|I)P(B|I)P(C|I)P(I) \quad (5)$$

where A , B , and C are the sample features values, and I is the specific label.

When we use the prediction Naive Bayesian classifier, we need to calculate the product of multiple probabilities, such as $P(A|I)P(B|I)P(C|I)P(I)$. But if the probability of one of them is 0, the final product is 0. If the product is 0, we cannot make the final comparison prediction. To solve this problem, we use Laplacian smoothing, which is often used to solve zero probability problems. In Laplacian smoothing, we can set the initial value of all features to λ , and the initial value of each classification label to $N\lambda$, where N is an integer, and then the probability of all features will not be 0 anymore. Assuming that the datasets are large, the change in the probability caused by adding λ to the number of each feature and $N\lambda$ to the number of each label is negligible. Therefore, using Laplace smoothing does not have any adverse effect on the final result of the experiment, and it is very convenient and effective to avoid the 0 probability problem.

After Laplace smoothing processing, the calculation method of prior probability and the probability of label y_i is shown in equations (6) and 7, respectively. The formula for calculating the posterior probability is shown in equation (8). We use $\sum S(y_i)$ represents the number of samples whose label is y_i and $\sum S(a_j, y_i)$ represents the number of samples whose label is y_i and the feature is a_j . The posterior probability is calculated for all labels in $Y' = [y'_1, y'_2, \dots, y'_s]$. If the posterior probability of y_i is the highest, y_i is the predicted label.

$$P(a'_j|y'_i) = \frac{\sum_{k=1}^n S(a_j, y_i) + \lambda}{\sum_{k=1}^n S(y_i) + N\lambda} \quad (6)$$

$$P(y_i) = \frac{\sum_{k=1}^n S(y_i) + \lambda}{n + N\lambda} \quad (7)$$

$$P(y'_i|a'_1, a'_2, \dots, a'_k) = \frac{P(y'_i) \prod_{j=1}^k P(a'_j|y'_i)}{\prod_{j=1}^k P(a'_j)} \quad (8)$$

3.2.4 DataTest

In order to verify the accuracy of predict algorithm, we must conduct the test experiment. The label of each sample in the testing data is known, and we get a predicted label after prediction. If the predicted label is the same as the sample label, it means that we predict to succeed. Otherwise, if the predicted label is different from the sample label, it means that we predict to fail. We count the number of times of successful predictions. The number of successful predictions divided by the size of the testing set is the prediction accuracy. We also use recall rate and precision rate to evaluate the performance of the proposed scheme.

The accuracy Ac we use is the proportion of the total that the model predicts correctly. The size of the testing set is $Total$. We regard the value of predicting low as low LP , predicting high as high HP , and predicting medium as medium MP . The number of successful predictions is $LP + MP + HP$. The formula of accuracy Ac is shown in equation (9).

$$Ac = \frac{LP + MP + HP}{Total} \quad (9)$$

The recall rate is used to evaluate the completeness of the results. We regard the value of predicting the actual low as low LP , predicting the actual low as high HP , and predicting the actual low as medium MP . The formula for the recall rate of low fertility intention is shown in equation (10).

$$LP_{recall} = \frac{LP}{LP + HP + MP} \quad (10)$$

The precision rate is often used to evaluate the quality of results. We regard the value of predicting low as low LP , predicting high as low HP_2 , and predicting medium as low MP_2 . The formula for the precision rate of low fertility intention is shown in equation (11).

$$LP_{precision} = \frac{LP}{LP + HP_2 + MP_2} \quad (11)$$

3.2.5 DataPredict

The purpose of this algorithm is to make predictions on the given features. We can input a set of features with unknown labels to predict. We calculate the posterior probability according to equation (8). Then, we get a predicted label Pb whose posterior probability is the highest. The algorithm pseudo-code of prediction of fertility intention is listed as Algorithm 1.

4 Performance analysis and evaluation

In this section, we first present the experimental settings. Then, we conduct the experiments and analyse the results to demonstrate the accuracy of our proposed scheme.

Algorithm 1 Prediction of fertility intention

Input: the sample data and feature set of unknown label x_i (Fs).
Output: the accuracy of prediction (Ac) and prediction label (Pb).

- 1: Obtain the data Y, X and convert the value of data to boolean or integer: Y', X' ;
- 2: $num = 0$; num represents the number of successful predictions.
- 3: $total = n$; $total$ represents the total number of samples predicted
- 4: Calculate the conditional probability $P(y'_j)$ of each label;
- 5: Calculate the conditional probability of each feature if the label is y'_j , $P(a'_i|y'_j)$;
- 6: Find the value of label variance y'_i that maximise the expression $P(y'_j) \prod_{k=1}^m P(a'_{ik}|y'_j)$.
- 7: **if** predict succeed **then**
- 8: $num = num + 1$;
- 9: **end if**
- 10: Calculate the accuracy of prediction (Ac), $num/total$;
- 11: For a samples that need to be predicted x_i , we repeat the above process to calculate the posterior probability to obtain the predicted label (Pb);

4.1 Experimental settings

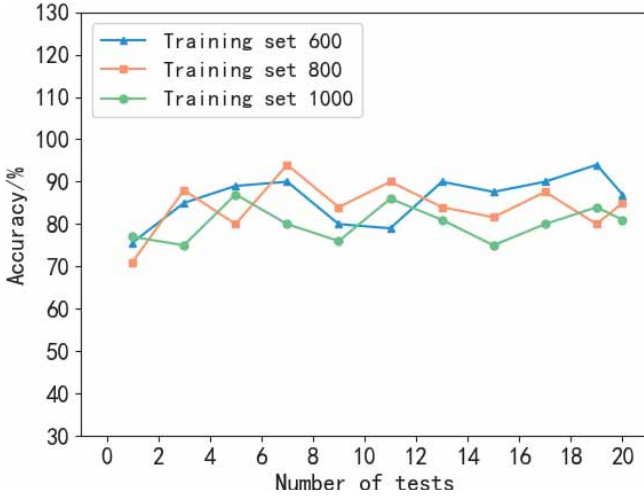
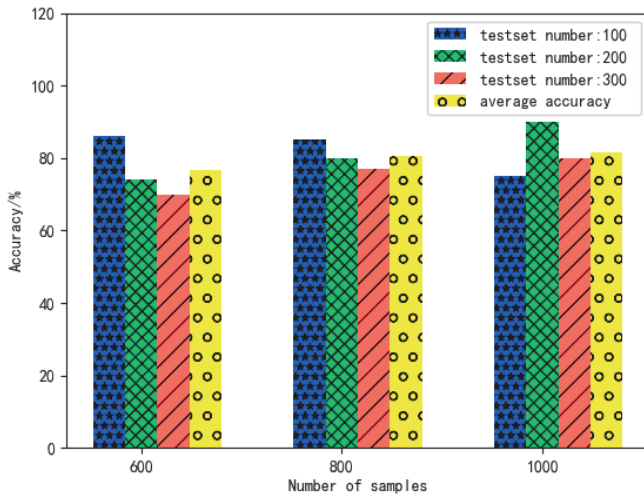
We collected about 1,500 pieces of data, which means about 1,500 people participated in our questionnaires. These data reflect the fertility intention of modern people. We use 80% of the data as training datasets and 20% of the data as testing datasets. Part of the testing datasets data we used is shown in Table 3. Our experiments are coded in python running on a machine with Intel(R)Core(TM) i5-8265U CPU, 8 GB of memory, and a Windows 10 professional operating system.

4.2 Experimental results

To evaluate the effect of λ , we use the same data samples and different λ to perform 20 experiments. Table 4 is the comparison of the accuracy of the proposed scheme under different λ . When $\lambda = 1$, $N = 2$, the accuracy of the proposed scheme is the same as those of $\lambda = 2$, $N = 2$. We can clearly see that the accuracy of the proposed scheme is maintained between 70% and 90%. According to the principle of Laplace smoothing, the effect of the change of the initial value on the conditional probability can be ignored when the data sample is large. Based on theoretical analysis and experiment results, we conclude that the accuracy of the proposed scheme is independent of λ .

Table 3 Part of the test data

Gender	Age	Education	Income	Children	Distance	Occupation	Intention
Male	28	University	6	1	Very far	Self-employed	High
Male	33	Middle school	8	1	A bit far	Corporate employees	Medium
Female	34	Postgraduate	9	2	A bit far	Corporate employees	High
Male	44	PhD	15	2	Close	Medical workers	High
Female	31	University	8	2	Close	Teachers	Medium
Male	27	PhD	10	0	Very far	Students	Low

Figure 2 The accuracy of FPB scheme (see online version for colours)**Figure 3** Average accuracy of FPB scheme (see online version for colours)

We get the predicted results of some samples as shown in Table 5, where $\lambda = 2$. In Table 5, we can find that women may have a higher fertility intention than men. And people who have a child in the family, have a decent job, or have a better family income may have a higher fertility intention. Because they have the ability to raise more children and have the energy and time to take care of them. Conversely, families with low income, living far away from school, or

having multiple children may have lower fertility intention. Besides, there are some young couples who do not want to be bound by their children. They want to live a free life, so these people's fertility intentions will be relatively low.

Table 4 Comparison of accuracy (%) of the proposed scheme ($\lambda = 1$ and $\lambda = 2$)

Run of numbers	1	3	5	7	9	11	15	17	20
$\lambda = 1$	89	85	87	78	86	84	75	85	70
$\lambda = 2$	89	85	87	78	86	84	75	85	70

We subsequently analyse the accuracy of the FPB scheme. The accuracy of the results of the FPB scheme is shown in Figure 2.

In Figure 2, when the number of samples is set to 600, 800, and 1,000, the accuracy of predictions of the FPB scheme are mostly between 70% and 90%. Figure 2 demonstrates that we can keep the accuracy of our proposed FPB scheme stable.

Figure 3 shows the corresponding average accuracy when the testing sets are 100, 200 and 300. As we can see from Figure 3, the average accuracy of the FPB scheme is relatively high. Figure 3 show that our proposed scheme can predict with high accuracy.

We can further analyse the precision rate and recall rate of the FPB scheme. The experimental results of the FPB scheme are shown in Table 6. For example, the number of low actual fertility intention and low predict fertility intention is 18. The total number of low actual fertility intention is 24, and the total number of low predicted fertility intention is 23. Thus the precision rate is $18/23 = 78.26\%$, and the recall rate is $18/24 = 75.00\%$. We can find that both precision rate and recall rate are more than 70%. It shows that our scheme predicts precisely and comprehensively.

After getting a relatively high accuracy, precision rate, and recall rate, we can make the flowing predictions. Given a sample $x_i = (\text{male}, 42, \text{middle school}, 5, 2, \text{very far}, \text{medical})$, the final result based on historical indications is likely to be low fertility intention. After prediction, we can find that the predicted label by our proposed scheme is 'low fertility intention'. It indicates that our proposed scheme can make accurate predictions. This result demonstrates the accuracy of our proposed scheme.

Table 5 Part of the test results

<i>Gender</i>	<i>Age</i>	<i>Education</i>	<i>Income</i>	<i>Children</i>	<i>Distance</i>	<i>Occupation</i>	<i>Predicted intention</i>
Male	28	Middle school	4	2	Close	Corporeate employees	Medium
Female	35	Postgraduate	8	2	Close	Other	High
Male	41	PhD	15	0	A bit far	Other	Low
Female	31	University	14	2	Close	Self-employed	High
Male	45	Elementary school	2	1	Very far	Migrant worker	Low
Female	35	University	9	1	Close	Teachers	High

Table 6 The recall and precision rate of FPB scheme

	<i>Actual fertility intention</i>				<i>Precision rate</i>	
	<i>Low</i>	<i>Medium</i>	<i>High</i>	<i>Total</i>		
Predict fertility intention	Low	18	4	1	23	78.26%
	Medium	2	12	2	16	75.00%
	High	4	1	16	21	76.19%
	Total	24	17	19	60	
Recall rate		75.00%	70.58%	84.00%		

5 Conclusions

This paper proposes a fertility intention prediction scheme based on polynomial Naive Bayesian model, called the FPB scheme. By performing a simple analysis of the data, we can obtain the prediction result through the calculation of conditional probability. After conducting a lot of experiments, we successfully verify that the FPB scheme is suitable for the prediction problems of fertility intention.

Our proposed scheme has been able to make the basic predictions of people's fertility intention, and the accuracy of the predictions is relatively high. However, a shortcoming of our scheme may be that Naive Bayes requires multiple features to be independent of each other. But in real society, complete independence of multiple features is impossible. Thus in future work, we need to solve the problem of independence of multiple features or minimise the error caused by the dependence of multiple features to improve the accuracy rate of the prediction algorithm.

Acknowledgements

This work is supported in part by the National Natural Science Foundation of China (No. 61702180), the Research Foundation of Education Bureau of Hunan Province (No. 21B0493), the Natural Science Foundation of Fujian Province (No. 2022J05106), and the Hunan Province Science and Technology Project Funds (No. 2018TP1036). The authors are grateful to the editor and anonymous reviewers for their invaluable feedback, and any remaining errors are attributed to the authors.

References

- Benkhelifa, R., Biskri, I., Laallam, F.Z. and Aïmeur, E. (2020) 'User content categorisation model, a generic model that combines text mining and semantic models', *International Journal of Computational Science and Engineering*, Vol. 21, No. 4, pp.536–555.
- Chen, C. (2021) 'Improved multiple Naive Bayes classification algorithms and Python implementation', *Journal of Jingdezhen University*, Vol. 36, No. 3, pp.92–95.
- Dai, Y. and Wang, T. (2020) 'Prediction of customer engagement behaviour response to marketing posts based on machine learning', *Connection Science*, Vol. 33, No. 4, pp.891–910.
- Diao, C., Zhang, D., Liang, W., Li, K-C., Hong, Y. and Gaudiot, J-L. (2023) 'A novel spatial-temporal multi-scale alignment graph neural network security model for vehicles prediction', *IEEE Transactions on Intelligent Transportation Systems*, Vol. 24, No. 1, pp.904–914.
- Geler, Z., Savić, M., Bratić, B., Kurbalija, V., Ivanović, M. and Dai, W. (2020) 'Sentiment prediction based on analysis of customers assessments in food serving businesses', *Connection Science*, Vol. 33, No. 3, pp.674–692.
- Huang, J., Chen, Y. and Ying, T. (2021) 'Investigation on the status quo of reproductive intentions of Ningbo residents and research on influencing factors', *Management and Statistics*, Vol. 36, No. 9, pp.78–82.
- Jiang, X. and Wang, J. (2021) 'Prediction of disease classification based on decision tree algorithm', *Information and Computer (Theoretical Edition)*, Vol. 33, No. 11, pp.51–53.
- Li, X., Bo, J. and Zhangrui, Z. (2021) 'Weighted Naive Bayes news text classification algorithm based on TF-IDF', *Network Security Technology and Application*, Vol. 11, No. 11, pp.31–33.
- Mai, J., Chen, J. and Zhao, H. (2020) 'An analysis of the influencing factors of college teachers' evaluation of teaching based on the orderly multi-class logistic regression model', *Journal of Leshan Normal University*, Vol. 35, No. 4, pp.111–116.

- Mustofa, R.L. and Prasetyo, B. (2021) 'Sentiment analysis using lexicon-based method with Naive Bayes classifier algorithm on #newnormal hashtag in Twitter', *Journal of Physics: Conference Series*, Vol. 1918, No. 4, pp.256–264.
- Ning, D., Ming, Y., Ning, S., Shuguang, Q. and Qihong, L. (2010) 'Classification of arrester defects based on Naive Bayesian inference', *IOP Conference Series: Earth and Environmental Science*, Vol. 804, No. 4.
- Peng, L., Peng, M., Liao, B., Huang, G., Liang, W. and Li, K. (2017) 'Improved low-rank matrix recovery method for predicting miRNA-disease association', *Scientific Reports*.
- Population by Age Bracket with UN Projections, World [online] <https://ourworldindata.org/grapher/historic-and-un-pop-projections-by-age> (accessed 23 April 2022).
- Shihao, L. (2017) *Prediction of Short-term Traffic Flow on Urban Roads based on Bayesian Classification and K-Nearest Neighbor Method*, 61pp, Lanzhou Jiaotong University.
- Sun, J. (2021) 'Research on application of Gaussian Naive Bayes algorithm in predicting college students' performance', *Computer Knowledge and Technology*, Vol. 17, No. 20, pp.23–26.
- Wang, Y. (2017) 'Prediction of second-child fertility intention of women of childbearing age based on data mining', *Computer Engineering & Software*, Vol. 38, No. 11, pp.55–59.
- Wang, Y. and Sun, M. (2017) 'Policies to encourage fertility and their effects: international experience, review and prospects', *International Journal of Embedded Systems*, Vol. 47, No. 5, pp.19–29.
- Wang, D., Lu, Y., Chen, B. and Zhao, Y. (2020) 'Wind weather prediction based on multi-output least squares support vector regression optimised by bat algorithm', *International Journal of Embedded Systems*, Vol. 12, No. 2, pp.137–145.
- Wang, S., Xu, F., Chao, H., Liu, Y. and Zuo, D. (2021) 'Research on fault classification of textile hot rolling mill based on support vector machine', *Modern Manufacturing Engineering*, Vol. 6, No. 6, pp.116–121.
- World Birth Rate 1950–2022 [online] <https://www.macrotrends.net/countries/WLD/world/birth-rate> (accessed 20 June 2022).
- Wu, F. and Li, J. (2021) 'Three signals that China is at risk of a birth crisis: low birth rate, low birth intention and birth deficit', *Shanxi Normal University News (Social Science Edition)*, Vol. 49, No. 1, pp.61–68.
- Yuan, Y. (2021) 'Research on sentiment classification of online comment text based on Naive Bayes', *Inner Mongolia Technology and Economy*, Vol. 18, No. 18, pp.91–94.
- Zhang, H., Hou, L., Ma, P. and Qiu, H. (2021) 'Prediction of fertility behavior and influencing factors analysis of 'post-80s' occupational population based on random forest algorithm', *Journal of Ningxia Medical University*, Vol. 43, No. 2, pp.149–155.
- Zheng, K. and Zhu, C. (2021) 'Naive Bayesian text classification algorithm based on context semantics', *Computer and Modernization*, Vol. 6, No. 6, pp.1–6.
- Zhu, G., Wang, Q., Liu, H. and Zhang, S. (2020) 'The mining method of trigger word for food nutrition matching', *International Journal of Computational Science and Engineering*, Vol. 23, No. 3, pp.205–213.