Research On an Optimal Nipt Timing Decision Model Driven by Multidimensional Factors Based on Machine Learning

Yangzhiping Chen *

School of Mathematics and Economics, Sichuan University, Chengdu, China, 610065

* Corresponding Author Email: chenyangzhiping@163.com

Abstract. The accuracy of non-invasive prenatal testing (NIPT) is highly dependent on selecting the optimal testing time, which varies due to individual differences among pregnant women. To address the limitations of the traditional "one-size-fits-all" testing approach, this paper constructs an intelligent model framework that integrates prediction, stratification, and decision-making. First, the framework employs a CatBoost regression model to accurately predict the key indicator of fetal cell-free DNA (cffDNA)—Y chromosome concentration—based on multidimensional physiological indicators of pregnant women, achieving a coefficient of determination (R^2) of 0.717 on the test set. Next, a K-means clustering algorithm is used to achieve data-driven refined stratification of the study population, and principal component analysis (PCA) is applied to process high-dimensional features. Building on this, the study innovatively introduces a Cox proportional hazards model to construct a risk function and establishes an optimization model to calculate personalized optimal NIPT timing for different stratified groups of pregnant women. For example, in the comprehensive multi-factor model, the optimal testing times determined for the low-BMI and high-BMI groups were 11.997 weeks and 10.420 weeks, respectively. In addition, to address the issue of detecting female fetal chromosomal aneuploidy, a decision tree classification model was developed, achieving an average accuracy of 0.886 with five-fold cross-validation. The core contribution of this study lies in the deep integration of machine learning prediction models with operational optimization and decision models, achieving a paradigm shift from "passive prediction" to "active decision-making," and providing a scientific, efficient, and individualized decision support solution for clinical NIPT practice.

Keywords: Machine learning, CatBoost regression, Cox regression, K-means clustering, decision optimization.

1. Introduction

NIPT, as a revolutionary prenatal screening technology, can efficiently screen for chromosomal aneuploidies such as trisomy 21 syndrome by analyzing cffDNA present in maternal peripheral blood [1]. Due to its high sensitivity and specificity, this technology has become a mainstream screening method worldwide. However, the effectiveness of NIPT is closely related to the timing of the test. Testing too early may result in false negatives due to insufficient cffDNA concentration, while testing too late may miss the optimal clinical intervention window [2]. Therefore, determining the best timing for NIPT for pregnant women with different physiological characteristics has become a key challenge in current clinical practice.

At present, the timing of NIPT testing mostly relies on broad clinical guidelines and generally lacks fine-grained consideration of individual differences. Existing research and practice often neglect the complex effects of multiple factors—such as maternal BMI, age, and gestational age—on cffDNA concentration [3]. The non-linear relationships and interactions between these physiological indicators make it difficult for traditional statistical methods to construct accurate predictive models, thus failing to provide optimal testing time recommendations for individuals. This "one-size-fits-all" approach not only affects the accuracy of the test but also does not fully utilize the potential of NIPT technology, highlighting the urgent need for an advanced methodology that can integrate multidimensional information to achieve personalized decision-making [4].

In response to these shortcomings, this paper proposes an innovative research framework that combines machine learning with optimized decision-making. What sets this study apart is that it not only predicts cffDNA concentrations, but, for the first time, organically integrates a high-precision

CatBoost prediction model, data-driven stratification based on K-means clustering, and survival analysis models based on Cox regression, constructing a complete closed-loop process from "prediction" to "decision." This framework can quantify the risk of testing at different time points according to the multidimensional personal characteristics of pregnant women, and determine the optimal gestational week for testing by balancing the risks of "missed detection" and "delayed intervention," thus providing a new paradigm for achieving individualized precision medicine.

The main research work in this paper centers around this model framework. First, we performed detailed preprocessing and feature engineering on the raw data. Next, we built and validated a CatBoost regression model for predicting Y chromosome concentrations in male fetuses, and analyzed its interpretability using methods such as SHAP. Subsequently, pregnant women were stratified using K-means clustering and principal component analysis, and, on this basis, an optimal NIPT timing decision model was established and solved. Furthermore, for the issue of chromosomal abnormality determination in female fetuses, we constructed an efficient decision tree classification model. Finally, the overall performance and potential application value of the model are summarized and discussed.

2. Methods

2.1. Data Preprocessing and Feature Engineering

In this study, we collected over 1,600 pieces of data related to NIPT among pregnant women through manual investigation. A systematic preprocessing of the raw data was first conducted to ensure the quality and reliability of model training. This process included removing invalid samples that did not meet minimum detection standards or had missing values in key indicators (such as Y-chromosome concentration), and applying the Interquartile Range (IQR) method to sequencing quality metrics (such as GC content) to identify and eliminate outliers. To enable effective use of differently formatted data by the model, discrete gestational age records (e.g., "11w+6d") were converted into continuous numerical variables. Additionally, to address the severe class imbalance present in the classification task for fetal chromosomal aneuploidy in female fetuses, we used a combination of oversampling and undersampling techniques to balance the proportions of positive and negative samples—thus avoiding bias toward the majority class and improving the model's ability to identify the minority (abnormal) cases.

2.2. Y-Chromosome Concentration Prediction Model

To accurately capture the complex non-linear relationship between pregnant women's multidimensional features and male fetal Y-chromosome concentration, this study ultimately selected the CatBoost regression model after comparing various machine learning algorithms. CatBoost is an advanced ensemble learning method based on Gradient Boosted Decision Trees (GBDT), and its core advantages lie in its ability to automatically and efficiently handle categorical features, as well as to effectively prevent overfitting through its symmetric tree growth strategy and Ordered Boosting mechanism [5][6]. The reason for selecting this model was that preliminary linear regression analyses revealed only a weak linear relationship (R^2 was just 0.018) between variables and Y-chromosome concentration, indicating the need for a more powerful nonlinear model. Compared with other ensemble models (such as XGBoost and Random Forest), CatBoost demonstrated lower overfitting risk and better generalization performance on our dataset. The training process follows the standard gradient boosting framework, iteratively building a series of decision trees, with each tree aimed at correcting the residuals from the previous iteration, and ultimately integrating all trees' predictions to achieve high-accuracy results [7]. This model serves as a critical prerequisite for subsequent timing optimization.

2.3. Personalized Timing Decision Framework

To shift from prediction to decision-making, this study constructed a multi-stage personalized timing decision framework. The first stage involves automated stratification of the pregnant population. We used the K-means clustering algorithm to group data based on BMI and other key physiological indicators. The optimal number of clusters was determined by the Elbow Method, ensuring statistical soundness in stratification. When handling multi-dimensional input features, PCA was first employed for dimensionality reduction to extract the most representative composite features, thus improving the efficiency and effectiveness of the clustering algorithm [8]. The second stage involves the construction and optimization of a risk function. Innovatively, this study introduced the Cox proportional hazards model, commonly used in the medical field, to characterize the probability of "Y-chromosome concentration reaching threshold" at different time points for different groups of pregnant women. The Cox model does not require a preset event time distribution and can effectively assess the impact of multiple covariates (such as BMI group, age, etc.) on event incidence [9][10]. Based on the fitted risk functions, we defined two risk functions: a threshold risk function for "failure to reach target concentration due to testing too early," and a window risk function for "missing the intervention window due to testing too late." Finally, we established an optimization model whose objective function minimizes the sum of these two risks, thereby deriving the optimal NIPT testing time point for each stratified group by solving this optimization problem. This framework transforms a complex clinical decision-making problem into a structured mathematical optimization problem.

2.4. Intelligent Aneuploidy Classification Model

For the classification task of female fetal trisomy 21, 18, and 13 chromosomes, this study constructed a decision tree classification model. This model can make a series of binary decisions based on multiple input features (such as the Z-scores of each chromosome, GC content, read counts, maternal BMI, etc.), ultimately classifying samples as "normal" or "abnormal." Prior to model development, we conducted strict feature selection based on medical literature and statistical association tests, removing variables that were not independently significantly associated with aneuploidy (such as number of deliveries), so as to enhance model accuracy and clinical practicality [11][12]. Decision trees offer inherent interpretability: each split node and path corresponds to a concrete clinical decision rule, making the model's logic clear and transparent [13]. We optimized the model's hyperparameters (such as maximum depth, minimum samples per split, etc.) using training data and evaluated its generalization ability through five-fold cross-validation, ensuring it remains robust and reliable when classifying unseen data.

3. Results

3.1. Performance and Interpretability Analysis of the Y-Chromosome Concentration Prediction Model

To gain a deeper understanding of the internal working mechanisms of the CatBoost prediction model and its dependence on clinical variables, we conducted both performance and interpretability analyses.

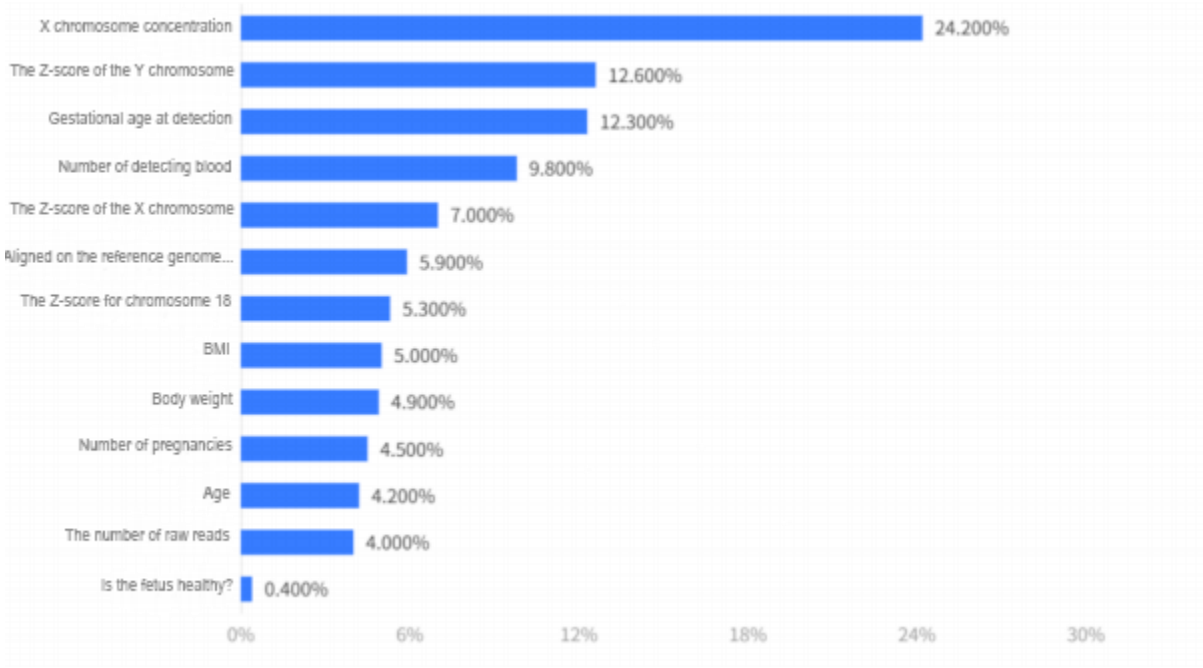

Figure 1. Feature Importance Ranking

As shown in Figure 1, the model's feature importance analysis clearly identifies the key factors influencing the prediction of Y chromosome concentration. X chromosome concentration tops the list with an importance of 24.2%, which is highly consistent with biological mechanisms—namely, the relative proportion of X and Y chromosome fragments in maternal plasma is a direct reflection of cffDNA concentration. Following closely is the Z-value of the Y chromosome (12.6%), gestational week at testing (12.3%), and number of blood draws (9.8%). These respectively represent statistical significance, cumulative effect of time, and possible variations between testing batches. This result validates that the model has successfully captured core biological and clinical time series information, with its decision-making basis grounded in solid scientific principles rather than mere surface-level correlations.

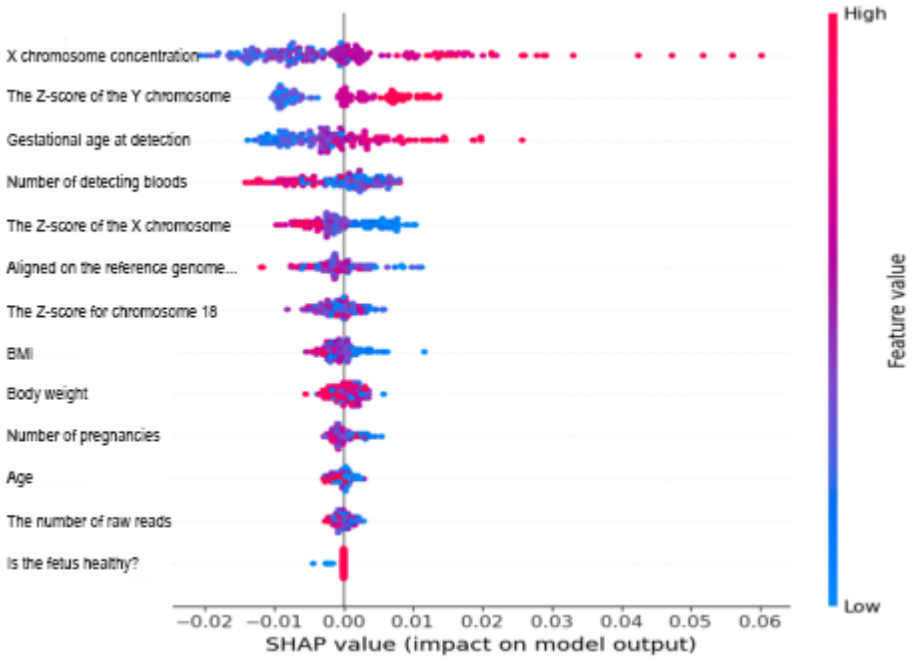

Figure 2. Model SHAP Value Analysis

To further explore the specific direction and magnitude of each feature's effect on the prediction results, we introduced SHAP (Shapley Additive exPlanations) analysis, as shown in Figure 2. This figure not only confirms the importance ranking from Figure 1 but also reveals how changes in feature values drive increases or decreases in the predicted values. For example, when the X chromosome concentration (X_concentration) feature value is high (red dots in the figure), the corresponding SHAP value is negative, indicating that a high X chromosome concentration strongly predicts a lower Y chromosome concentration—fully consistent with clinical logic. Conversely, when the feature value for gestational week (Test_gestational_week) is high (red dots), the SHAP value is generally positive, showing that as gestational week increases, the model's prediction of Y chromosome concentration also increases. This clear interpretability, consistent with domain knowledge, greatly enhances the credibility of the model and demonstrates the rationality of its internal decision logic.

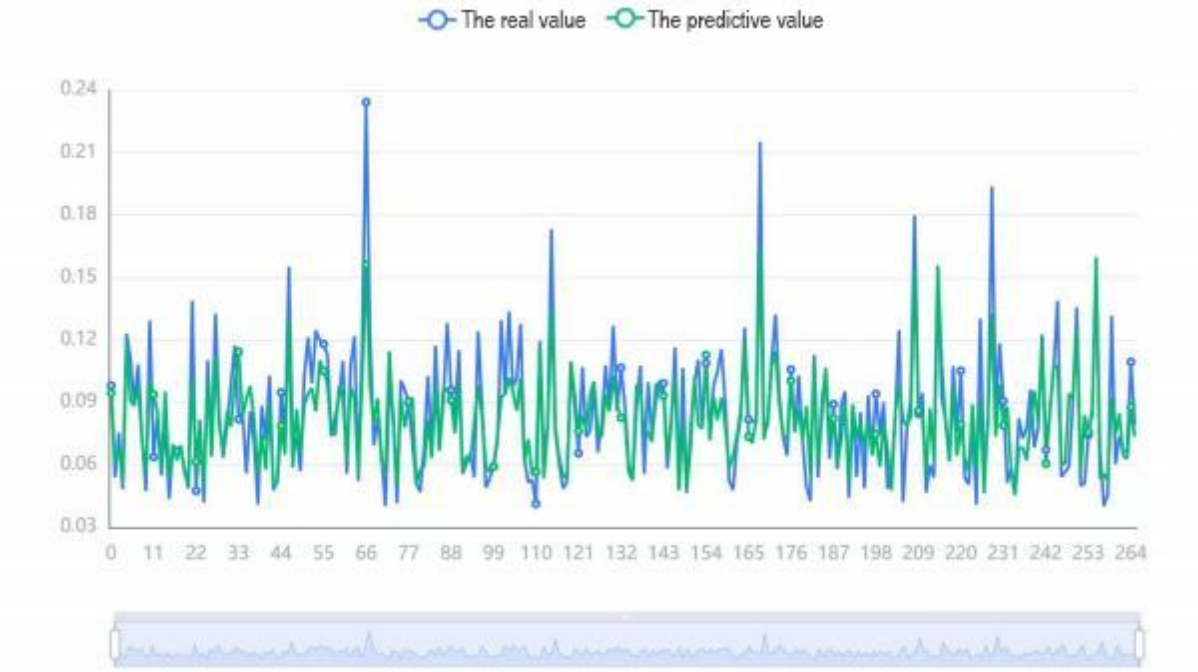

Figure 3. Model Prediction Performance on Test Set

As shown in Figure 3, the model’s predicted values on the test set align closely with the actual values. The points are tightly clustered around the y=x diagonal line, indicating small prediction errors and strong generalization ability. This figure intuitively demonstrates that the CatBoost model can achieve stable and accurate predictions of Y chromosome concentration in new samples, and the reliability of its prediction results provides a robust data foundation for subsequent timepoint optimization decisions. This serves as a fundamental guarantee for the effectiveness of the entire decision-making framework.

3.2. NIPT Timing Optimization Analysis Based on BMI Stratification

After verifying the effectiveness of the prediction model, we proceeded to stratify the population of pregnant women to enable personalized decision-making.

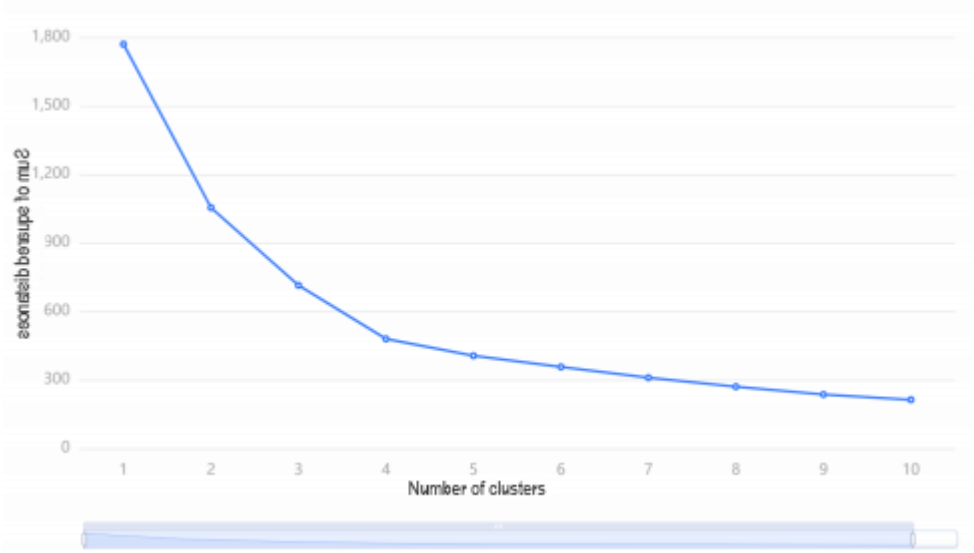

Figure 4. Elbow method analysis chart for K-means clustering

As shown in Figure 4, this chart illustrates the process of determining the optimal number of clusters (K) using the elbow method. The horizontal axis represents the number of clusters K, and the vertical axis shows the sum of squared errors within clusters (SSE). It is clearly observable that as K increases from 1 to 4, the SSE drops sharply; however, when K exceeds 4, the decrease in SSE significantly slows down, forming an "elbow" shape similar to an arm. This inflection point (K=4) marks the threshold where the marginal benefit of increasing the number of clusters begins to diminish, making the choice of 4 clusters the data-driven optimal decision. This step ensures that the subsequent stratification of pregnant women is statistically meaningful and avoids the bias introduced by subjective classification [14].

Figure 5. Scatterplot of BMI clustering results

As shown in Figure 5, this scatterplot visually presents the clustering results of the K-means algorithm (K=4). Each point in the figure represents a pregnant woman, and different colors distinguish the four distinct clusters. The plot demonstrates that the four groups form relatively separate and well-defined clusters in two-dimensional space, with tight cohesion within clusters and clear separation between them. This provides intuitive evidence that the clustering algorithm successfully divided pregnant women into four subgroups with different characteristics based on their BMI features, offering a solid foundation for developing differentiated NIPT testing strategies tailored to each group.

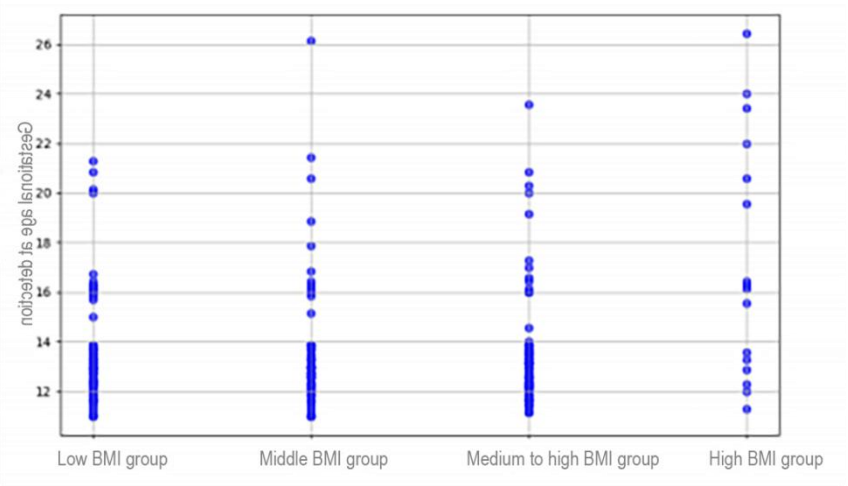

Figure 6. Distribution of gestational age at testing across different BMI groups

As shown in Figure 6, this box plot compares the distribution of gestational age at testing within the four BMI groups. Through this figure, we can observe differences not only in the median and range of gestational age at testing across groups, but more importantly, it clearly highlights extreme outliers in each group (points far from the box). These outliers may result from special clinical circumstances or data recording errors and can disproportionately impact the optimization of the model. Therefore, this plot provides an intuitive basis for identifying and removing these outliers, which is a critical quality control step to ensure that the final calculated optimal timing is both robust and representative.

3.3. Validation of the Proportional Hazards Model with Multiple Factors

When considering the influence of multiple factors on the timing of NIPT, we employed the Cox regression model, the use of which requires the proportional hazards assumption to be satisfied [15].

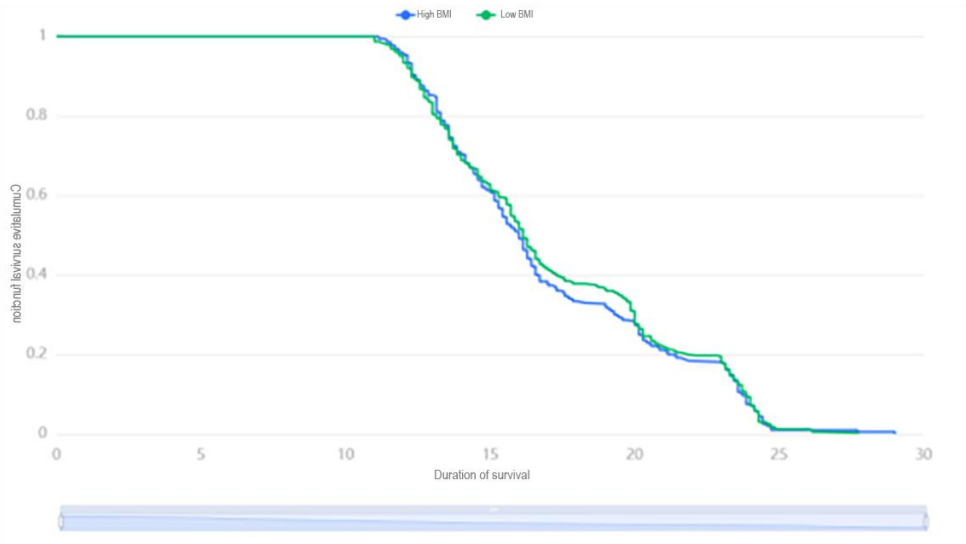

Figure 7. Kaplan-Meier survival curve

As shown in Figure 7, this figure plots the Kaplan-Meier survival curves for two groups divided by BMI (low BMI group and high BMI group) after dimensionality reduction of multiple factors. In this context, the "survival event" is defined as the Y chromosome concentration reaching the detection threshold. The two curves in the figure are roughly parallel, without significant crossover or divergence, which visually provides strong support for the proportional hazard's assumption. Meeting this assumption means that the hazard ratio between the different BMI groups remains essentially constant throughout the observation period, which is a prerequisite for using the Cox model to assess risk. Statistical confirmation of this assumption was further obtained through tests such as the Log-

Rank test ($p>0.05$), thereby ensuring the methodological validity of subsequent risk function construction and timing optimization based on the Cox model [16].

3.4. Feature Analysis of the Aneuploidy Classification Model

Finally, for the task of determining female fetal chromosomal aneuploidy, we analyzed the feature importance of the decision tree model.

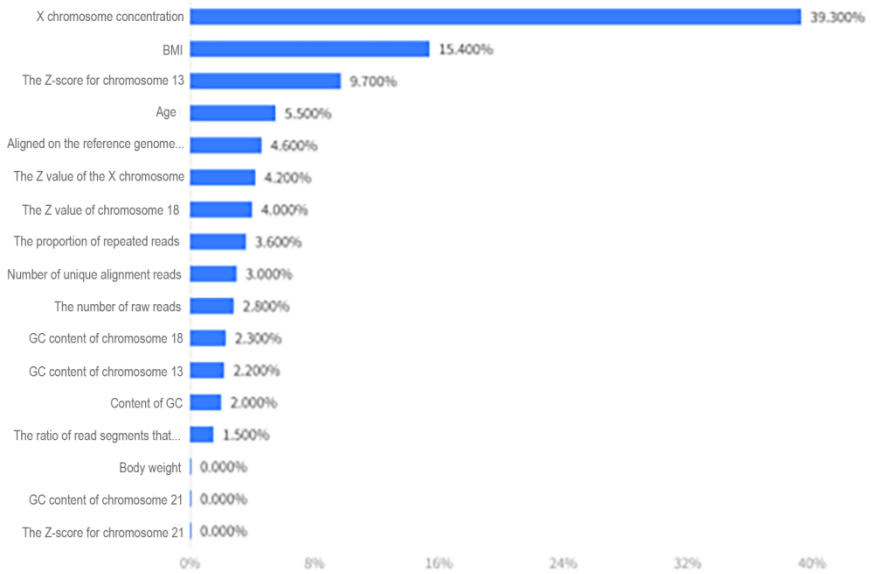

Figure 8. Feature importance chart of the decision tree model

As shown in Figure 8, the chart reveals which decision criteria the model relies on most when identifying chromosomal abnormalities in female fetuses. Unlike the model for predicting Y chromosome concentration, here the mother's weight and BMI are the two most important features, with importance values of 18.8% and 17.8%, respectively. This indicates that the maternal physiological state is a key background factor influencing aneuploidy screening signals. Following closely are the Z value and concentration of chromosome 13, suggesting that quantitative indicators for specific chromosomes serve as direct evidence for discrimination. This finding carries significant clinical implications: it highlights the necessity of combining individual maternal physical parameters with specific chromosomal sequencing data when assessing aneuploidy risk [17]. It also explains why a dedicated model—one with a completely different set of feature weights from the cffDNA concentration prediction model—is required for this particular task.

4. Conclusions

This paper successfully establishes and validates a multidimensional factor-driven intelligent decision-making framework for identifying the optimal timing of NIPT based on machine learning. The core contribution of the study lies in the effective transformation of clinical problems into solvable mathematical models by integrating a high-precision CatBoost predictive model, data-driven K-means clustering stratification, and an optimized decision-making model based on Cox regression. This approach provides personalized and dynamic recommendations for the optimal testing time for pregnant women with different physiological characteristics. In addition, this study offers a reliable decision tree classification scheme for determining fetal chromosomal aneuploidy. The advantages of the model are reflected in its rigorous data processing procedures, flexible and appropriate model selection, as well as strong interpretability and generalizability ensured through methods such as SHAP analysis and cross-validation. However, there are certain limitations to this study. The model's performance is highly dependent on the quality and scale of the training data, and the universality of its conclusions remains to be verified with more diverse population data. At the same time, the setting

of the risk function in the optimization model relies on certain assumptions, and more complex functional forms may be explored in future research. Future directions may focus on two aspects: first, integrating more dimensions of bioinformatics data and dynamic monitoring data to construct sequence prediction models for more precise real-time decision support; second, exploring end-to-end deep learning methods to attempt learning optimal decision strategies directly from raw data, further enhancing the model's level of intelligence.

References

[1] Bianchi D W, Parker R L, Wentworth J, et al. DNA sequencing versus standard prenatal aneuploidy screening[J]. New England journal of medicine, 2014, 370(9): 799-808.

[2] Jayashankar S S, Nasaruddin M L, Hassan M F, et al. Non-invasive prenatal testing (NIPT): reliability, challenges, and future directions[J]. Diagnostics, 2023, 13(15): 2570.

[3] Hsieh V, Sherer D M, Davydovych K, et al. The Art (and Science) of Individualized Selection of Non-Invasive Prenatal Screening (NIPS)[J]. International Journal of Women's Health, 2025: 1271-1283.

[4] Faieta M, Falcone R, Duca S, et al. Test performance and clinical utility of expanded non-invasive prenatal test: Experience on 71,883 unselected routine cases from one single center[J]. Prenatal Diagnosis, 2024, 44(8): 936-945.

[5] Kulkarni C S. Advancing gradient boosting: A comprehensive evaluation of the CatBoost algorithm for predictive modeling[J]. J. Artif. Intell. Mach. Learn. Data Sci, 2022, 1(5): 54-57.

[6] Dorogush A V, Ershov V, Gulin A. CatBoost: gradient boosting with categorical features support[J]. arXiv preprint arXiv:1810.11363, 2018.

[7] Biau G, Cadre B. Optimization by gradient boosting[M]//Advances in Contemporary Statistics and Econometrics: Festschrift in Honor of Christine Thomas-Agnan. Cham: Springer International Publishing, 2021: 23-44.

[8] Zubair M, Iqbal M D A, Shil A, et al. An improved K-means clustering algorithm towards an efficient data-driven modeling[J]. Annals of Data Science, 2024, 11(5): 1525-1544.

[9] Kalbfleisch J D, Schaubel D E. Fifty years of the cox model[J]. Annual Review of Statistics and Its Application, 2023, 10(1): 1-23.

[10] Abd ElHafeez S, D'Arrigo G, Leonardis D, et al. Methods to analyze time-to-event data: the Cox regression analysis[J]. Oxidative medicine and cellular longevity, 2021, 2021(1): 1302811.

[11] Wei L, Zhang J, Shi N, et al. Association of maternal risk factors with fetal aneuploidy and the accuracy of prenatal aneuploidy screening: a correlation analysis based on 12,186 karyotype reports[J]. BMC Pregnancy and Childbirth, 2023, 23(1): 136.

[12] Wei L, Zhang J, Shi N, et al. Effects of Maternal Factors on Fetal Aneuploidy and Reliability of Screening: A Cohort Study Based on 12,186 Karyotype Reports[J]. 2022.

[13] Mienye I D, Jere N. A survey of decision trees: Concepts, algorithms, and applications[J]. IEEE access, 2024, 12: 86716-86727.

[14] Kassambara A. Practical guide to cluster analysis in R: Unsupervised machine learning[M]. Sthda, 2017.

[15] Sjölander A, Dickman P W. Why test for proportional hazards—or any other model assumptions?[J]. American journal of epidemiology, 2024, 193(6): 926-927.

[16] Schober P, Vetter T R. Kaplan-Meier curves, log-rank tests, and cox regression for time-to-event data[J]. Anesthesia & Analgesia, 2021, 132(4): 969-970.

[17] Elmerdahl Frederiksen L, Ølgaard S M, Roos L, et al. Maternal age and the risk of fetal aneuploidy: A nationwide cohort study of more than 500 000 singleton pregnancies in Denmark from 2008 to 2017[J]. Acta obstetricia et gynecologica Scandinavica, 2024, 103(2): 351-359.