# Supplementary File C. T2DM prediction model based on machine learning

### C.1 Population

This study developed a diabetes prediction model based on the Hongguang population, while the Kailuan and Baoan population were utilized as external validation groups to assess the generalizability and robustness of the model. The inclusion and exclusion criteria for all three populations were consistent with those applied in the external validation of the risk score model.

### C.2 Model Construction

The dataset was randomly partitioned into training and testing sets at a ratio of 7:3. Subsequently, several machine learning models were constructed, including Logistic Regression, Support Vector Machines (SVM), Random Forest (RF), eXtreme Gradient Boosting (XGBoost).

The dataset in this study exhibited a class imbalance issue, which was addressed through the application of several advanced resampling techniques, including the Synthetic Minority Oversampling Technique (SMOTE), Adaptive Synthetic Sampling (ADASYN), Borderline-SMOTE (BLSMOTE), and Density-Based SMOTE (DBSMOTE). The SMOTE algorithm generates synthetic data by identifying the k-nearest neighbors for each minority class sample and interpolating between them using a random weight within the range [0, 1]. The ADASYN method enhances this approach by constructing a weighted distribution for minority class samples, thereby synthesizing additional training data for harder-to-learn instances. BLSMOTE and DBSMOTE represent improved variants of SMOTE, refining the generation of synthetic samples. Specifically, BLSMOTE focuses on borderline samples to enhance classifier performance in handling complex decision boundaries, while DBSMOTE employs density estimation to identify dense and sparse regions within the minority class, preserving the intrinsic structure of the data and reducing noise introduced by random sample generation. The encoding results of the variables are presented in S3 Table.

### C.3 Metrics for model evaluation

In this study, 5 metrics were used to examine the performance of the prediction model, including the area under the subject's working characteristic curve (AUC), accuracy (ACC), specificity (SPE), sensitivity (SEN), G-mean, higher values of these metrics indicate better prediction performance of the model. Specificity represents the proportion of true survivors correctly predicted as survivors, reflecting the classifier's ability to identify negative cases. Sensitivity indicates the proportion of true deceased cases correctly predicted as deceased. The G-mean is a composite metric that combines specificity and sensitivity, evaluating the classifier's capability to accurately identify both deceased and surviving samples. Accuracy denotes the proportion of correctly predicted samples, taking into account both positive and negative cases comprehensively. The AUC assesses the overall performance of the classification model by considering the classifier's ability to distinguish between positive and negative classes, providing a robust evaluation even in the presence of imbalanced datasets.

S3 Table The encoding of variables in machine learning model

|  |  |
| --- | --- |
| Variables | Encoding |
| gender | Male = 1，Female = 0 |
| central obesity | waist circumference < 85 cm(female) and waist circumference < 90 cm(male)=0，waist circumference ≥85cm (female) and waist circumference ≥90 cm(male)=1 |
| exercise | No exercise =0，exercise =1 |
| smoking | No smoking=0，smoking=1 |
| hyperhistory | No hypertension=0，hypertension=1 |
| DM\_family\_history | No family history of diabetes = 0, family history of diabetes = 1. |
| Fattyliver | No fatty liver=0，fatty liver=1 |
| FBG | < 5.6=0，≥5.6=1 |
| TG | < 1.7=0，≥1.7=1 |
| TYG | Continuous variable |
| age | Continuous variable |
| BMI | Continuous variable |
| WHtR | Continuous variable |