**数据预处理**

消除高度相关的特征是机器学习中的一个重要数据预处理步骤，尤其是对于高维度的数据集。此步骤有助于解决多重共线性问题，降低过拟合风险，提高计算效率，并增强模型泛化。Spearman相关系数通常用于评估特征之间的成对相关性。大多数特征对表现出弱相关性（Spearman 相关系数 <0.7）。高度相关的特征对（Spearman 秩相关系数> 0.9）为血肌酐和eGFR，与结果相关性较低的特征将从数据集中消除，即消除eGFR。纳入机器学习的剩下18个指标。

图表, 散点图

AI 生成的内容可能不正确。

**机器学习模型开发和比较**

80% 用于训练，20% 用于验证（内部验证），外部数据集用于测试（外部验证）。

总共17个特征用于开发预测模型。连续变量缺失数据采用均值插补法处理，分类变量缺失数据采用中位数插补法处理。

5个ML模型，即随机森林、LightBGM、XGBoost、决策树和Logistic用于预测DKD。表中列出了5个模型的判别性能。

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Model** | **AUC** | **Sensitivity** | **Specificity** | **PPV** | **NPV** | **Accuracy** | **F1-score** |
| RF | 0.883 | 0.800 | 0.875 | 0.823 | 0.858 | 0.843 | 0.811 |
| DT | 0.734 | 0.662 | 0.805 | 0.711 | 0.767 | 0.745 | 0.686 |
| LightGBM | 0.872 | 0.738 | 0.860 | 0.793 | 0.819 | 0.809 | 0.764 |
| XGBoost | 0.874 | 0.752 | 0.870 | 0.807 | 0.829 | 0.820 | 0.779 |
| LR | 0.866 | 0.745 | 0.890 | 0.831 | 0.828 | 0.829 | 0.785 |
| RF: random forest; LightGBM: light gradient boosting machin; LR: logistic regression; XGBoost: Extreme Gradient Boosting; DT: decision tree. | | | | | | | |

递归特征筛选（Recursive Feature Elimination, RFE）是一种常见的特征选择方法，其核心思想是基于模型自带的特征重要性排名，结合设定的步长和最低保留特征数量，对特征集进行逐步递归地筛选。在每次迭代中，根据模型计算的特征贡献度（如基于随机森林的Gini重要性或基于Boosting算法的分裂增益等），剔除一部分排名最低的特征，直到达到设定的最低特征数量或性能平衡点，该方法能够在减少特征数量的同时尽量保持模型性能的稳定性。在根据特征重要等级减少特征的过程中，这5个模型的 AUC 变化如下图，RF 模型几乎保持了这 5 个模型中最好的预测能力。

图表, 条形图

AI 生成的内容可能不正确。

图表, 折线图

AI 生成的内容可能不正确。

随后我们定义 DeLong's test 函数，通过DeLong's test 的相关算法实现了两个相关的 ROC 曲线下面积差异的比较。我们比较了使用前3、9、10个指标的模型与使用全部特征的模型，计算 AUC 差值（ΔAUC）和对应的P值。显示比较结果：9个指标的AUC与全部指标的P值为0.015，存在统计学差异；而纳入前10个指标时的AUC与全部指标的P值为0.176，无统计学差异。

图表, 折线图

AI 生成的内容可能不正确。

因此，我们后续建模及验证采用重要性前10的指标即DR、Duration of DM、HbA1c、Serum creatinine、TC、Urine protein excretion、FBG、BMI、Age、SBP进行建模及内部和外部验证。为了更好地评估模型性能和，避免模型的过拟合和欠拟合，我们在内部验证队列中采用5倍和10倍交叉验证，最终模型在5倍和10倍交叉验证中均实现了0.88± 0.02的平均 AUC。最终的RF模型在5倍交叉验证中预测DKD方面AUC为0.88，灵敏度为0.73，特异性为0.88，PPV为0.81，NPV为0.82，准确性为0.82，F1评分为0.77。

图表, 折线图

AI 生成的内容可能不正确。

图表

AI 生成的内容可能不正确。

5倍交叉验证结果：

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **AUC** | **Sensitivity** | **Specificity** | **PPV** | **NPV** | **Accuracy** | **F1-score** |
| MEAN | 0.88 | 0.73 | 0.88 | 0.81 | 0.82 | 0.82 | 0.77 |
| Std | 0.02 | 0.02 | 0.03 | 0.04 | 0.01 | 0.02 | 0.02 |
| Std: Standard Deviation. | | | | | | | |

如前所述训练的 RF 模型的性能在外部测试集中保持稳定(AUC: 0.83)。图表, 折线图

AI 生成的内容可能不正确。

我们使用SHAP方法对模型进行解释，首先是对模型全局进行解释：变量对结果的影响可以通过 SHAP 值的大小（由颜色变化表示）和变量横轴上的趋势来直观地解释。如 SHAP汇总图所示，使用平均SHAP值评估特征对模型的贡献，并按降序显示。然后评估单个特征如何影响预测模型的输出例如，DR与没有DR的患者相比，DKD发生的概率更大；对于糖尿病病程，时间越长，DKD的概率越大。

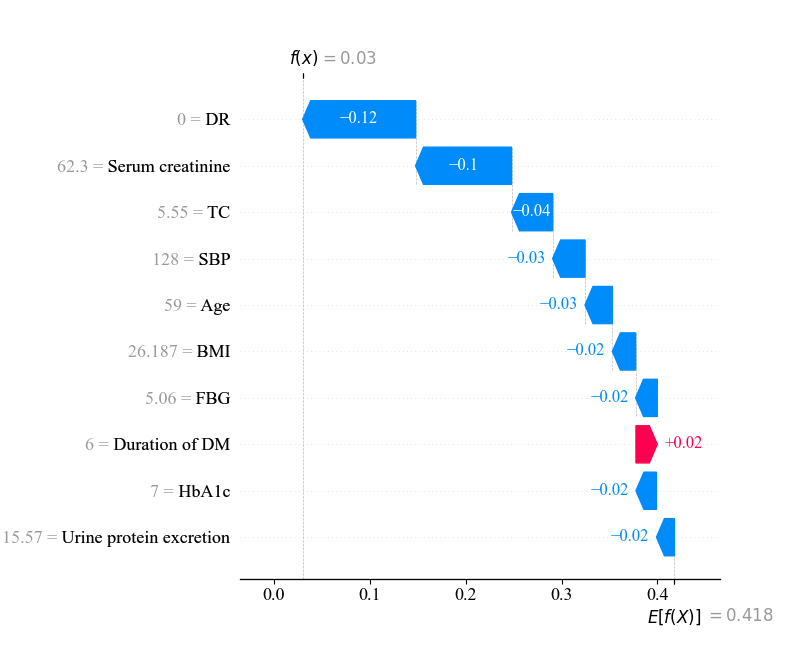
图表, 条形图

AI 生成的内容可能不正确。

图表

AI 生成的内容可能不正确。

对于单样本的解释，我们随机选择了一名患者，患者的基本信息如下：患者59岁，BMI 26.19，SBP128mmHg，患有糖尿病6年，无糖尿病视网膜病变，生化检测显示HbA1c 7，FBG5.06，Serum creatinine 62.3，TC 5.55，Urine protein excretion 15.57，通过该模型计算出患者患有DKD的概率为0.03，该患者的临床诊断为NDKD，与我们的模型相符。



图形用户界面, 文本, 应用程序

AI 生成的内容可能不正确。

DKD患者，患者的基本信息如下：患者55岁，BMI 30.11，SBP173mmHg，患有糖尿病2年，有糖尿病视网膜病变，生化检测显示HbA1c 5.9，FBG5.36，Serum creatinine 135.7，TC 5.12，Urine protein excretion 2.36，通过我们的模型计算出患者患有DKD的概率为076。

