目录

[1. 问题重述 5](#_Toc85194543)

[1.1 问题背景 5](#_Toc85194544)

[1.2 问题要求 5](#_Toc85194545)

[2. 模型假设 7](#_Toc85194546)

[3. 符号说明 8](#_Toc85194547)

[4. 模型准备 8](#_Toc85194548)

[5. 问题一： 10](#_Toc85194549)

[5.1 问题分析 10](#_Toc85194550)

[5.2 模型建立 10](#_Toc85194551)

[5.3 模型求解 10](#_Toc85194552)

[5.4 算法设计 10](#_Toc85194553)

[6. 问题二： 11](#_Toc85194554)

[6.1 问题分析 11](#_Toc85194555)

[6.2 模型建立 11](#_Toc85194556)

[6.3 模型求解 11](#_Toc85194557)

[6.4 算法设计 11](#_Toc85194558)

[7. 问题三： 12](#_Toc85194559)

[7.1 问题分析 12](#_Toc85194560)

[7.2 模型建立 12](#_Toc85194561)

[7.3 模型求解 12](#_Toc85194562)

[7.4 算法设计 12](#_Toc85194563)

[8. 模型检验与评价 13](#_Toc85194564)

[9. 模型的优缺点与改进 14](#_Toc85194565)

[9.1 模型检验 14](#_Toc85194566)

[9.2 模型评价 14](#_Toc85194567)

[9.2.1 模型优点 14](#_Toc85194568)

[9.2.2 模型缺点 14](#_Toc85194569)

[参考文献 15](#_Toc85194570)

[附录 16](#_Toc85194571)

[附录1 代码 16](#_Toc85194572)

[附录2 资料 16](#_Toc85194573)

# 问题重述

## 问题背景

乳腺癌是世界上最常见的癌症之一，也是女性患癌死亡的第二大原因[1]，根据2008 年 GLOBOCAN统计，全球乳腺癌的发病率约38. 9 /10 万。

乳腺癌的发生和发展与雌激素的作用密切相关，雌激素由其受体(estmgen receptor，ER)介导，引起ER构象发生改变，刺激各种与细胞生长有关的基因表达增高，诱导肿瘤细胞生长、增殖增速，并减弱对其基因的抑制作用，引起乳腺癌等恶性肿瘤[2]。ER包括ERα和ERβ两种亚型，其控制细胞增殖和分化，ERα参与细胞以及癌细胞的增殖，ERβ则具有相反的作用，因此可被认为是肿瘤抑制剂。对于ER呈阳性的患者，目前最常见的治疗方式是采用内分泌治疗[3]，通过药物调节雌激素受体的活性靶向抑制雌激素对乳腺癌细胞生长的促进作用。ERα被认为是治疗乳腺癌的重要靶标，能够拮抗ERα活性的化合物可能是治疗乳腺癌的候选药物,比如，临床治疗乳腺癌的经典药物他莫昔芬和雷诺昔芬就是ERα拮抗剂。

内分泌治疗是针对激素受体阳性乳腺癌患者的重要治疗手段之一，具有疗效确切、不良反应小、治疗费用较低等特点，临床应用广泛[6]。因此，筛选出一种能够针对ERα有抑制作用，同时在人体内具备良好的药代动力学性质和安全性的活性化合物显得尤为重要。

## 问题要求

本文根据提供的ERα拮抗剂信息（1974个化合物样本，每个样本都有729个分子描述符变量，1个生物活性数据，5个ADMET性质数据），构建化合物生物活性的定量预测模型和ADMET性质的分类预测模型，从而为同时优化ERα拮抗剂的生物活性和ADMET性质提供预测服务。

现根据以上研究背景和所提供数据，本文需要研究和解决以下问题：

**问题一：寻找建模主要变量**

根据文件“Molecular\_Descriptor.xlsx”和“ERα\_activity.xlsx”提供的数据，针对1974个化合物的729个分子描述符进行变量选择，根据变量对生物活性影响的重要性进行排序，并给出前20个对生物活性最具有显著影响的分子描述符（即变量），并请详细说明分子描述符筛选过程及其合理性。

**问题二：建立ERα生物活性的定量预测模型**

结合问题1，选择不超过20个分子描述符变量，构建化合物对ERα生物活性的定量预测模型，请叙述建模过程。然后使用构建的预测模型，对文件“ERα\_activity.xlsx”的test表中的50个化合物进行IC50值和对应的pIC50值预测，并将结果分别填入“ERα\_activity.xlsx”的test表中的IC50\_nM列及对应的pIC50列。

**问题三：建立化合物的分类预测模型**

利用文件“Molecular\_Descriptor.xlsx”提供的729个分子描述符，针对文件“ADMET.xlsx”中提供的1974个化合物的ADMET数据，分别构建化合物的Caco-2、CYP3A4、hERG、HOB、MN的分类预测模型，并简要叙述建模过程。然后使用所构建的5个分类预测模型，对文件“ADMET.xlsx”的test表中的50个化合物进行相应的预测，并将结果填入“ADMET.xlsx”的test表中对应的Caco-2、CYP3A4、hERG、HOB、MN列。

**问题四：优化模型**

寻找并阐述化合物的哪些分子描述符，以及这些分子描述符在什么取值或者处于什么取值范围时，能够使化合物对抑制ERα具有更好的生物活性，同时具有更好的ADMET性质（给定的五个ADMET性质中，至少三个性质较好）。

# 模型假设

假设1：所有样本的原料性质、操作位点的数据测量正确、数据处理步骤正确；

假设2：所有位点均是可以操作改变的控制变量；

假设3：在优化主要变量时认为所提出的预测模型结果准确。

# 符号说明

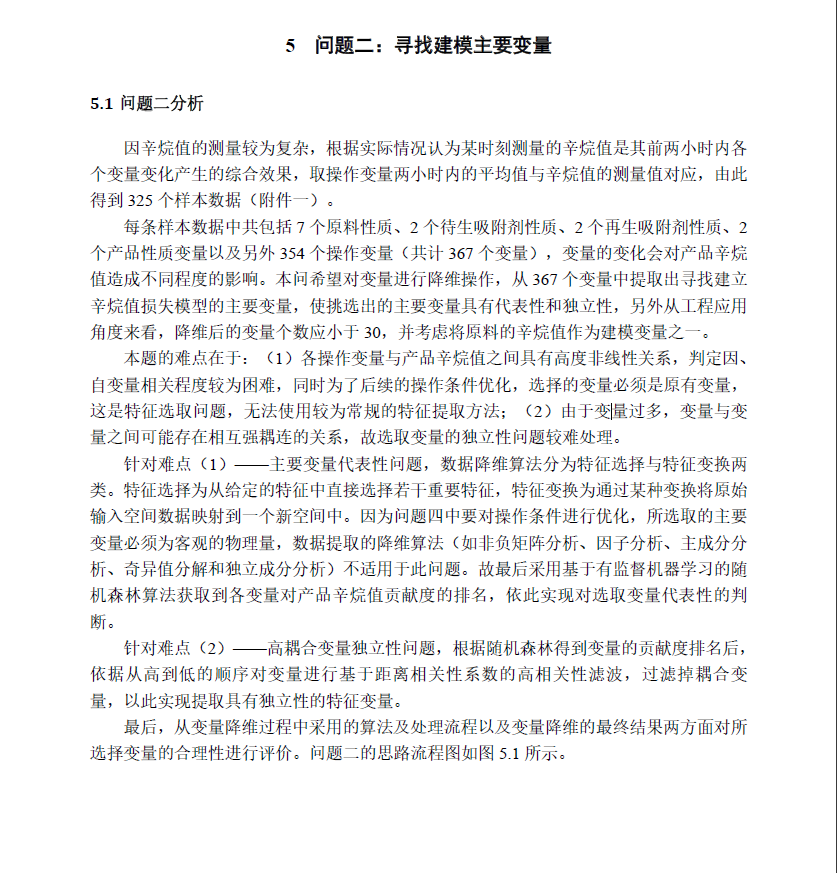
|  |  |  |
| --- | --- | --- |
| 符号 | 含义 | 单位 |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

# 模型准备

# 问题一：

## 问题分析

参考B20102470089.pdf



## 模型建立

随机森林、灰色预测，空间距离相关系数，SPSS Spearman相关性分析

## 模型求解

## 变量选择合理性评价

# 问题二：

## 问题分析

## 模型建立

## 模型求解

## 算法设计

# 问题三：

## 问题分析

## 模型建立

## 模型求解

## 算法设计

# 模型检验与评价

# 模型的优缺点与改进

## 模型检验

## 模型评价

### 模型优点

### 模型缺点

参考文献

附录

1. 代码
2. 资料

(

1‑1

)

图片包含 游戏机, 画

描述已自动生成

图 1.1 图片标题

表 1.1 表格标题

|  |  |  |  |
| --- | --- | --- | --- |
| 表格内容 |  |  |  |
| 1 |  |  |  |
| 2 |  |  |  |
| 3 |  |  |  |