Аналитический отчёт курсовой работы

Цель исследования.

На основании предоставленных данных от химиков необходимо построить прогноз, позволяющий подобрать наиболее эффективное сочетание параметров для создания лекарственных препаратов.

Задачи исследования.

Создать несколько максимально эффективных моделей для решения следующих задач:

•Регрессия для IC50

•Регрессия для CC50

•Регрессия для SI

•Классификация: превышает ли значение IC50 медианное значение выборки

•Классификация: превышает ли значение CC50 медианное значение выборки

•Классификация: превышает ли значение SI медианное значение выборки

•Классификация: превышает ли значение SI значение 8

Сравнить между собой полученные модели и их результаты, выполнить анализ, обосновать выбор наиболее качественных решений.

Описание предмета исследования

Был проанализирован датасет в котором содержатся данные с числовыми характеристиками химических соединений и параметры, характеризующие эффективность, обозначаются как IC50, CC50 и SI, они и являлись нашими целевыми переменными. Датасет содержит 1001 строку и 213 признаков, 3 из которых являются целевыми переменными.

Датасет имеет 110 дискретных признаков 18 из них имеют только нулевые значения. В последствии они были удалены так, как они для модели будут бесполезными. Проанализировав столбчатые диаграммы зависимостей дискретных признаков и целевых переменных IC50, mM, CC50, mM, SI, был сделан вывод о том, что признаки имеют категории которые имеют выраженную связь с целевыми переменными IC50, mM и SI.

Датасет имеет 103 непрерывных признака . Проанализировав гистограммы числовых непрерывных признаков, был сделан вывод о том, что большинство признаков имеют разное распределение, что говорит о хорошей разделяющей способности данных признаков, но есть признаки с похожим распределением.

Проанализировав пропуски, было обнаружено, что 12 признаков имеют 3 пропуска, эти 3 строки были удалены.

Проанализировав корреляционную матрицу было обнаружено большое количество признаков с высокой корреляцией более 90% между ними: fr_phenol - fr_phenol_noOrthoHbond, fr_Al_OH - fr_phenol, fr_Al_OH - fr_phenol_noOrthoHbond,fr_nitro_arom -

fr_nitro_arom_nonortho,fr_Al_OH_noTert - fr_Al_OH,NHOHCount - NumHDonors,VSA_EState3 - NumHDonors,fr_C_O - fr_C_O_noCOO,fr_C_O - VSA_EState2,fr_benzene - NumAromaticCarbocycles,NumAromaticCarbocycles - SMR_VSA7,fr_benzene - SMR_VSA7,SMR_VSA7 - VSA_EState6,SMR_VSA7 - SlogP_VSA6, VSA_EState6 - SlogP_VSA6,fr_Nhpyrrole - fr_Ar_NH,fr_COO - fr_COO2,fr_COO - fr_Al_COO,fr_COO2 - fr_Al_COO,fr_Ar_NH - fr_Nhpyrrole,MolMR - HeavyAtomMolWt,MolMR - Kappa1,MolMR - MolWt, MolMR - ExactMolWt, MolMR - Chi0, MolMR - Chi1, MolMR - HeavyAtomCount,MolMR - NumValenceElectrons,MolMR - Chi1n,MolMR - Chi1v,MolMR - Chi0n,MolMR - LabuteASA,MolMR - Chi0v,MolMR - Kappa2,Kappa2 - Chi0n,Kappa2 - Chi0v,Kappa2 - Kappa1,Kappa2 - Kappa3,HeavyAtomCount - Chi0,HeavyAtomCount - Chi1,HeavyAtomCount -LabuteASA,HeavyAtomCount - NumValenceElectrons, HeavyAtomCount - ExactMolWt, HeavyAtomCount - MolWt, HeavyAtomCount - HeavyAtomMolWt,HeavyAtomCount - Kappa1, HeavyAtomCount - Chi0v,HeavyAtomCount - Chi0n,HeavyAtomCount - Chi1n,HeavyAtomCount - Chi1v,HeavyAtomCount - Kappa2,Chi1 - BertzCT,Chi1 - Chi0,Chi1 - LabuteASA,Chi1 - NumValenceElectrons,Chi1 - ExactMolWt,Chi1 - MolWt, Chi1 - HeavyAtomMolWt,Chi1 - Kappa1,Chi1 - Chi0v,Chi1 - Chi0n,Chi1 - Chi1n,Chi1 - Chi1v,Chi1 - Kappa2,Chi0 - LabuteASA,Chi0 - NumValenceElectrons,Chi0 - ExactMolWt, Chi0 - MolWt, Chi0 - HeavyAtomMolWt,Chi0 - Kappa1,Chi0 - Chi0v,Chi0 - Chi0n,Chi0 - Chi1n,Chi0 - Chi1v,Chi0 - Kappa2,LabuteASA - NumValenceElectrons,LabuteASA - ExactMolWt, LabuteASA - MolWt,LabuteASA - HeavyAtomMolWt,LabuteASA - Kappa1,LabuteASA - Chi0v, LabuteASA - Chi0n, LabuteASA - Chi1n, LabuteASA - Chi1v, LabuteASA - Kappa2, NumValenceElectrons - ExactMolWt, NumValenceElectrons - MolWt, NumValenceElectrons - HeavyAtomMolWt, NumValenceElectrons - Kappa1, NumValenceElectrons - Chi0v, NumValenceElectrons - Chi0n,NumValenceElectrons - Chi1n, NumValenceElectrons - Chi1v,NumValenceElectrons - Kappa2, ExactMolWt - MolWt,ExactMolWt - HeavyAtomMolWt,ExactMolWt - Kappa1,ExactMolWt - Chi0v,ExactMolWt - Chi0n, ExactMolWt - Chi1n, ExactMolWt - Chi1v,ExactMolWt - Kappa2,MolWt - HeavyAtomMolWt,MolWt - Kappa1,MolWt - Chi0v, MolWt - Chi0n,MolWt - Chi1n, MolWt - Chi1v,MolWt - Kappa2,HeavyAtomMolWt - Kappa1,HeavyAtomMolWt - Chi0v, HeavyAtomMolWt - Chi0n, HeavyAtomMolWt - Chi1n, Kappa1 - Chi0v, Kappa1 - Chi0n, Kappa1 - Chi1n, Kappa1 - Chi1v, Chi0v - Chi0n,Chi0v - Chi1n,Chi0v - Chi1v,Chi0n - Chi1n,Chi0n - Chi1v,Chi1n - Chi1v,NumAliphaticCarbocycles - NumSaturatedCarbocycles,NumAliphaticCarbocycles - SMR_VSA4,Kappa1 - Chi0n,Chi0n - Chi0v,NOCount - NumHeteroatoms, NOCount - TPSA,NOCount - NumHAcceptors,SlogP_VSA11 - SMR_VSA9,Chi4v - Chi2n,Chi4v - Chi2v,Chi4v - Chi3n,Chi4v - Chi3v,Chi4v - Chi4n,Chi4n - Chi2n,Chi4n - Chi3n,Chi4n - Chi3v,Chi3v - Chi2n,Chi3v - Chi2v,Chi3v - Chi3n,Chi3n - Chi2n,Chi3n - Chi2v,Chi2v - Chi1v,Chi2v - Chi2n,Chi2v - Chi1n,Chi2n - Chi1v,BertzCT - HeavyAtomMolWt,BertzCT - Chi1,BertzCT - HeavyAtomCount,BertzCT - MolWt,BertzCT - ExactMolWt,FpDensityMorgan1 - FpDensityMorgan2,FpDensityMorgan2 - FpDensityMorgan3,MaxAbsPartialCharge - MaxEStateIndex,MaxEStateIndex - MaxAbsEStateIndex, MaxPartialCharge - MinAbsPartialCharge, Chi0n — Chi2n.

После этого были удалены признаки: BertzCT , Chi0 , Chi0v , Chi1 , Chi1n , Chi1v , Chi2v , Chi3n , Chi3v , Chi4n , Chi4v , ExactMolWt , FpDensityMorgan1 , FpDensityMorgan3 , HeavyAtomCount , HeavyAtomMolWt , Kappa1 , Kappa2 , Kappa3 , LabuteASA , MaxAbsEStateIndex , MaxAbsPartialCharge , MolMR , MolWt , NHOHCount , NumAromaticCarbocycles , NumHAcceptors , NumHeteroatoms , NumSaturatedCarbocycles ,

NumValenceElectrons , SMR_VSA4 , SMR_VSA9 , SlogP_VSA6 , TPSA , VSA_EState2 , VSA_EState3 , VSA_EState6 , fr_Al_OH_noTert , fr_COO , fr_COO2 , fr_C_O_noCOO , fr_Nhpyrrole , fr_benzene , fr_nitro_arom_nonortho , fr_phenol , fr_phenol_noOrthoHbond , MinAbsPartialCharge , Chi2n . Оставим наиболее встречающиеся признаки: fr_nitro_arom, NumHDonors, fr_C_O, fr_Ar_NH, Chi0n, fr_Al_COO, Chi0n, Chi2n, SMR_VSA7, NOCount, fr_Al_OH, MaxEStateIndex, FpDensityMorgan2, NumAliphaticCarbocycles, SlogP_VSA11, MaxPartialCharge.

Проанализировав матрицы корреляций признаков с целевыми переменными можно сделать вывод, что целевые переменные имеют очень слабую или слабую линейную корреляцию с другими признаками. Соответственно применение линейных моделей нецелесообразно.

Проанализировав выбросы целевых переменных в задачах регрессии были сделаны выводы о том, что во всех целевых переменных мы можем наблюдать выбросы, но основное негативное влияние на метрики оно оказывает на целевую переменную SI, поэтому мы удалим строки в которых значение целевой переменной больше 250.

Из анализа обзора распределения целевых переменных в задачах классификации видно, что во всех задачах, кроме задачи «превышает ли значение SI значение 8» классы распределены равномерно.

Методы исследования.

В задачах регрессии применялись 4 модели, одна модель это дерево решений (DecisionTreeRegressor) и три модели градиентного бустинга (CatBoostRegressor, LGBMRegressor, XGBRegressor).

Метрики были выбраны:

Коэффициент детерминации, или r2, измеряет долю дисперсии целевой переменной, которую модель объясняет. Этот коэффициент варьируется от бесконечности до 1.

Корень из среднеквадратичной ошибки (RMSE) — измеряет среднее квадратичное отклонение предсказанных значений от истинных.

В задачах классификации применялись 4 модели, одна модель это дерево решений (DecisionTreeClassifier) и три модели градиентного бустинга (CatBoostClassifier, LGBMClassifier, XGBClassifier).

Метрики были выбраны:

Accuracy — это метрика которая показывает долю правильно классифицированных объектов от общего числа объектов.

F-мера является гармоническим средним между точностью и полнотой. F1-меру используют, когда важно учитывать как ложные положительные, так и ложные отрицательные срабатывания.

В качестве валидации была применена отложенная выборка (hold-out). Данные были разбиты в пропорциях train — 80%, test — 20%.

Анализ и выводы по результатам исследования.

- **Регрессия для IC50**

После предобработки данных, подбора признаков, подбора гиперпараметров были получены следующие результаты:

	R2 - train	RMSE - train	R2 - test	RMSE - test
DecisionTreeRegressor	0.89	136.89	-0.01	355.50
CatBoostRegressor	0.80	183.27	0.55	238.20
LGBMRegressor	0.85	159.33	0.57	206.88
XGBRegressor	0.89	137.00	0.46	259.84

Наилучший результат и наименьшее переобучение показала модель lightgbm с результатом на трейне: r2=0.84 и RMSE=159.33 и тесте: r2=0.57 и RMSE=206.88. Данная модель объясняет 57% дисперсии между тренеровочными и тестовыми данными и корень среднеквадратичной ощибки составляет 206.88 микромоль (мкМ).

- **Регрессия для CC50**

После предобработки данных, подбора признаков, подбора гиперпараметров были получены следующие результаты:

	R2 - train	RMSE - train	R2 - test	RMSE - test
DecisionTreeRegressor	0.90	195.29	0.29	561.51
CatBoostRegressor	0.89	202.64	0.56	440.52
LGBMRegressor	0.83	261.90	0.63	402.31
XGBRegressor	0.89	209.80	0.59	424.86

Наилучший результат и наименьшее переобучение показала модель lightgbm с результатом на трейне: r2=0.83 и RMSE=261.9 и тесте: r2=0.63 и RMSE=402.32. Данная модель объясняет 63% дисперсии между тренеровочными и тестовыми данными и

корень среднеквадратичной ошибки составляет 402.32 микромоль (мкМ). Также хорошую метрику показала модель CatBoostRegressor, но у нее больше переобучение.

- **Регрессия для SI**

После предобработки данных, подбора признаков, подбора гиперпараметров были получены следующие результаты:

	R2 - train	RMSE - train	R2 - test	RMSE - test
DecisionTreeRegressor	0.79	13.20	-0.06	40.10
CatBoostRegressor	0.77	13.68	0.29	32.66
LGBMRegressor	0.70	15.54	0.30	32.38
XGBRegressor	0.78	13.20	0.25	33.75

Наилучший результат и наименьшее переобучение показала модель lightgbm с результатом на трейне: r2= 0.70 и RMSE= 15.54 и тесте: r2=0.31 и RMSE=32.38. Данная модель объясняет 32% дисперсии между тренеровочными и тестовыми данными и корень среднеквадратичной ощибки составляет 32.38 микромоль (мкМ).

- **Классификация: превышает ли значение IC50 медианное значение выборки**

После предобработки данных, подбора признаков, подбора гиперпараметров были получены следующие результаты:

	f1_score - train	accuracy_score - train	f1_score - test	accuracy_score - test
DecisionTreeClassifier	0.95	0.95	0.61	0.64
CatBoostClassifier	0.95	0.95	0.72	0.73
LGBMClassifier	0.95	0.95	0.69	0.70
XGBClassifier	0.95	0.95	0.70	0.70

Наилучший результат показала модель CatBoostClassifier с результатом на трейне: f1_score= 0.95 и accuracy_score= 0.95 и тесте: f1_score=0.72, accuracy_score= 0.725.

- **Классификация: превышает ли значение CC50 медианное значение выборки**

После предобработки данных, подбора признаков, подбора гиперпараметров были получены следующие результаты:

	f1_score - train	**accuracy_score - train**	**f1_score - test**	**accuracy_score - test**
DecisionTreeClassifier	0.97	0.97	0.7	0.7
CatBoostClassifier	0.97	0.97	0.79	0.78
LGBMClassifier	0.97	0.97	0.78	0.78
XGBClassifier	0.97	0.97	0.78	0.78

Наилучший результат показала модель CatBoostClassifier с результатом на трейне: f1_score= 0.97 и accuracy_score= 0.97 и тесте: f1_score=0.7864, accuracy_score=0.78.

- **Классификация: превышает ли значение SI медианное значение выборки**

После предобработки данных, подбора признаков, подбора гиперпараметров были получены следующие результаты:

	f1_score - train	**accuracy_score - train**	**f1_score - test**	**accuracy_score - test**
DecisionTreeClassifier	0.95	0.95	5.9	0.6
CatBoostClassifier	0.95	0.95	0.67	0.68
LGBMClassifier	0.95	0.95	0.68	0.66
XGBClassifier	0.95	0.95	0.6	0.6

Наилучшие результаты показали модели LGBMClassifier с результатом на трейне: f1_score= 0.95 и accuracy_score= 0.95 и тесте: f1_score=0.676, accuracy_score=0.661 и XGBClassifier с результатом на трейне: f1_score= 0.95 и accuracy_score= 0.95 и тесте: f1_score=0.674, accuracy_score=0.677.

- **Классификация: превышает ли значение SI значение 8**

После предобработки данных, подбора признаков, подбора гиперпараметров были получены следующие результаты:

	f1_score - train	accuracy_score - train	f1_score - test	accuracy_score - test
DecisionTreeClassifier	0.91	0.94	0.45	0.67
CatBoostClassifier	0.92	0.94	0.58	0.76
LGBMClassifier	0.89	0.93	0.59	0.76
XGBClassifier	0.91	0.94	0.59	0.76

Наилучшие результаты показали модели LGBMClassifier с результатом на трейне: f1_score= 0.89 и accuracy_score= 0.93 и тесте: f1_score=0.5892, accuracy_score=0.764 и XGBClassifier с результатом на трейне: f1_score= 0.92 и accuracy_score= 0.94 и тесте: f1_score=0.5913, accuracy_score=0.759.

Рекомендации по решению проблем, выявленных в ходе исследования.

В задачах регрессии для улучшения значения метрики и снижения переобучения нужно будет увеличить количество данных. Также можно попробовать логарифмировать целевую переменную.

В задачах классификации для улучшения значения метрики и снижения переобучения нужно будет увеличить количество данных.