Министерство цифрового развития

Федеральное государственное бюджетное образовательное учреждение высшего образования
«Сибирский государственный университет телекоммуникаций и информатики»
(СибГУТИ)

Кафедра прикладной математики и кибернетики

Отчёт

по лабораторной работе № 3 «Классификация методом дерева решений»

Выполнил:

студент группы ИП-312

Дорогин Н.С.

Работу проверил: старший преподаватель кафедры ПМиК Дементьева К.И.

Новосибирск 2025 г.

Введение (задание)

Цель: освоить практическое применение метода решающего дерева для задач классификации. Исследовать влияние гиперпараметров на качество модели и научиться проводить базовый анализ важности признаков.

Основная часть:

1. Подготовка данных

1.1. Выбираем для работы датасет из первой лабораторной работы, а конкретно файл Stress_Dataset. За целевую переменную берём вопрос Have you recently experienced stress in your life?

1.2. Загружаем и выполняем предобработку.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 843 entries, 0 to 842
Data columns (total 26 columns):
 #   Column                                                Non-Null Count
---  ------                                                --------------
 0   Gender                                                843 non-null
 1   Age                                                   843 non-null
 2   Have you recently experienced stress in your life?    843 non-null
 3   Have you noticed a rapid heartbeat or palpitations?   843 non-null
 4   Have you been dealing with anxiety or tension recently?   843 non-null
 5   Do you face any sleep problems or difficulties falling asleep?   843 non-null
 6   Have you been dealing with anxiety or tension recently?.1   843 non-null
 7   Have you been getting headaches more often than usual?   843 non-null
 8   Do you get irritated easily?                          843 non-null
 9   Do you have trouble concentrating on your academic tasks?   843 non-null
 10  Have you been feeling sadness or low mood?            843 non-null
 11  Have you been experiencing any illness or health issues?   843 non-null
 12  Do you often feel lonely or isolated?                 843 non-null
 13  Do you feel overwhelmed with your academic workload?   843 non-null
 14  Are you in competition with your peers, and does it affect you?   843 non-null
 15  Do you find that your relationship often causes you stress?   843 non-null
 16  Are you facing any difficulties with your professors or instructors?   843 non-null
 17  Is your working environment unpleasant or stressful?   843 non-null
 18  Do you struggle to find time for relaxation and leisure activities?   843 non-null
 19  Is your hostel or home environment causing you difficulties?   843 non-null
 20  Do you lack confidence in your academic performance?   843 non-null
 21  Do you lack confidence in your choice of academic subjects?   843 non-null
 22  Academic and extracurricular activities conflicting for you?   843 non-null
 23  Do you attend classes regularly?                      843 non-null
 24  Have you gained/lost weight?                          843 non-null
 25  Which type of stress do you primarily experience?     843 non-null
dtypes: int64(25), object(1)
```

```
70/30

590.0999999999999

252.89999999999998


80/20

674.4000000000001

168.60000000000002
```

```
0    Eustress (Positive Stress) - Stress that motiv...
1    Eustress (Positive Stress) - Stress that motiv...
2    Eustress (Positive Stress) - Stress that motiv...
3    Eustress (Positive Stress) - Stress that motiv...
4    Eustress (Positive Stress) - Stress that motiv...
5    Eustress (Positive Stress) - Stress that motiv...
6    Eustress (Positive Stress) - Stress that motiv...
7    Eustress (Positive Stress) - Stress that motiv...
8    Eustress (Positive Stress) - Stress that motiv...
9    No Stress - Currently experiencing minimal to ...
Name: Which type of stress do you primarily experience?, dtype: object


 Заменим наш единственный категориальный признак на числовой.


0    0
1    0
2    0
3    0
4    0
5    0
6    0
7    0
8    0
9    1
Name: Which type of stress do you primarily experience?, dtype: object
--------------------------------------------------

Поделим 70 на 30

Размер обучающей:  590
Размер тестовой:  253

 Всего после разделения:  843
```

Итого: пропущенных значений нет, единственный категориальный признак в числовой, разделили выборку на обучающую и тестовую в соотношении 70 на 30.

2. Базовое дерево

```
de_tree = DecisionTreeClassifier(random_state=42)

all_parameters = list(stress_dataset.columns)
parameters = all_parameters[:2] + all_parameters[3:]
target = all_parameters[2]

print("Признаки:", parameters)
print("Целевая переменная:", target)

X_learn = learn_part[parameters]
y_learn = learn_part[target]
clf = de_tree.fit(X_learn, y_learn)

X_test = test_part[parameters]
y_test = test_part[target]

predictions = clf.predict(X_test)
acc = accuracy_score(y_test, predictions)

print(f"\nТочность на тестовой выборке: {acc}")
print(f"Размер обучающей выборки: {len(learn_part)}")
print(f"Размер тестовой выборки: {len(test_part)}")

print("\n\n")
print("-" * 50, "\n")
print("Важность признаков:\n\n");

feature_importance = clf.feature_importances_

importance = pd.DataFrame({
    'Признак': parameters,
    'Важность': feature_importance
}).sort_values('Важность', ascending=False)

print(importance)
```

```
Признаки: ['Gender', 'Age', 'Have you noticed a rapid heartbeat or palpitations?', 'Have you been dealing wit
Целевая переменная: Have you recently experienced stress in your life?

Точность на тестовой выборке: 0.5731225296442688
Размер обучающей выборки: 590
Размер тестовой выборки: 253



--------------------------------------------------

Важность признаков:


                                           Признак  Важность
2   Have you noticed a rapid heartbeat or palpitat...  0.097489
23                      Have you gained/lost weight?  0.079576
1                                             Age  0.067064
21  Academic and extracurricular activities confli...  0.066782
14  Do you find that your relationship often cause...  0.064001
22                  Do you attend classes regularly?  0.057889
8   Do you have trouble concentrating on your acad...  0.057133
12  Do you feel overwhelmed with your academic wor...  0.049430
3   Have you been dealing with anxiety or tension ...  0.048110
15  Are you facing any difficulties with your prof...  0.045202
4   Do you face any sleep problems or difficulties...  0.040183
6   Have you been getting headaches more often tha...  0.039955
10  Have you been experiencing any illness or heal...  0.037127
11              Do you often feel lonely or isolated?  0.037118
9          Have you been feeling sadness or low mood?  0.033957
20  Do you lack confidence in your choice of acade...  0.033710
5   Have you been dealing with anxiety or tension ...  0.030431
19  Do you lack confidence in your academic perfor...  0.023445
13  Are you in competition with your peers, and do...  0.022435
7                       Do you get irritated easily?  0.021189
16  Is your working environment unpleasant or stre...  0.016632
24  Which type of stress do you primarily experience?  0.010935
18  Is your hostel or home environment causing you...  0.007699
0                                            Gender  0.007551
17  Do you struggle to find time for relaxation an...  0.004960
```

Итого: точность на тестовой выборке с параметрами по умолчанию примерно 57,3%; три наиболее важных признака: ускоренное сердцебиение, лишний вес и возраст.

3. Подбор гиперпараметров

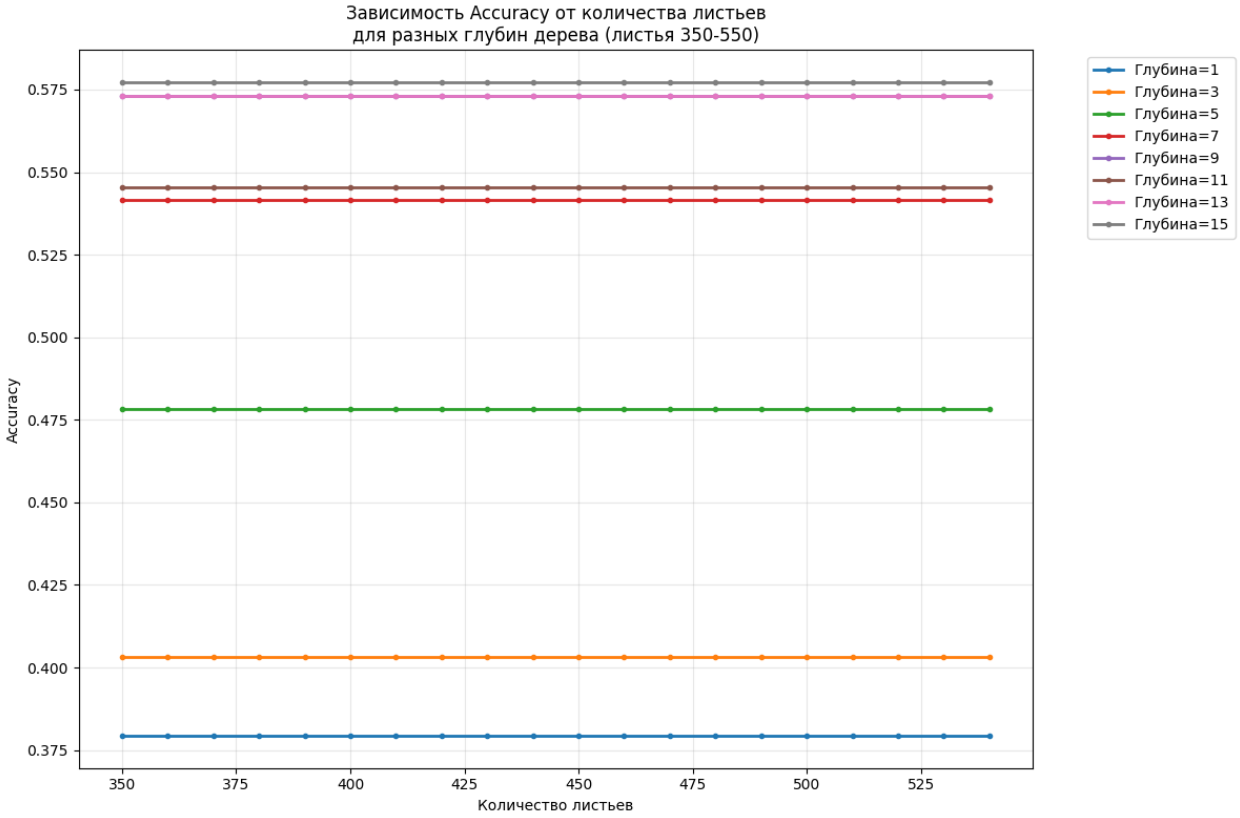
```
all_parameters = list(stress_dataset.columns)
parameters = all_parameters[:2] + all_parameters[3:]
target = all_parameters[2]

print("Признаки:", parameters)
print("Целевая переменная:", target)


print("max_depth      |     max_leaf_nodes      |   accurancy")
print('-' * 100)
for i in range(1, 16):
  for j in range(350, 550):
    de_tree = DecisionTreeClassifier(max_depth = i, max_leaf_nodes=j, random_state=42)
    X_learn = learn_part[parameters]
    y_learn = learn_part[target]
    clf = de_tree.fit(X_learn, y_learn)

    X_test = test_part[parameters]
    y_test = test_part[target]

    predictions = clf.predict(X_test)
    acc = accuracy_score(y_test, predictions)
    print(i,"             |     ",j,"             |   ",acc)
```

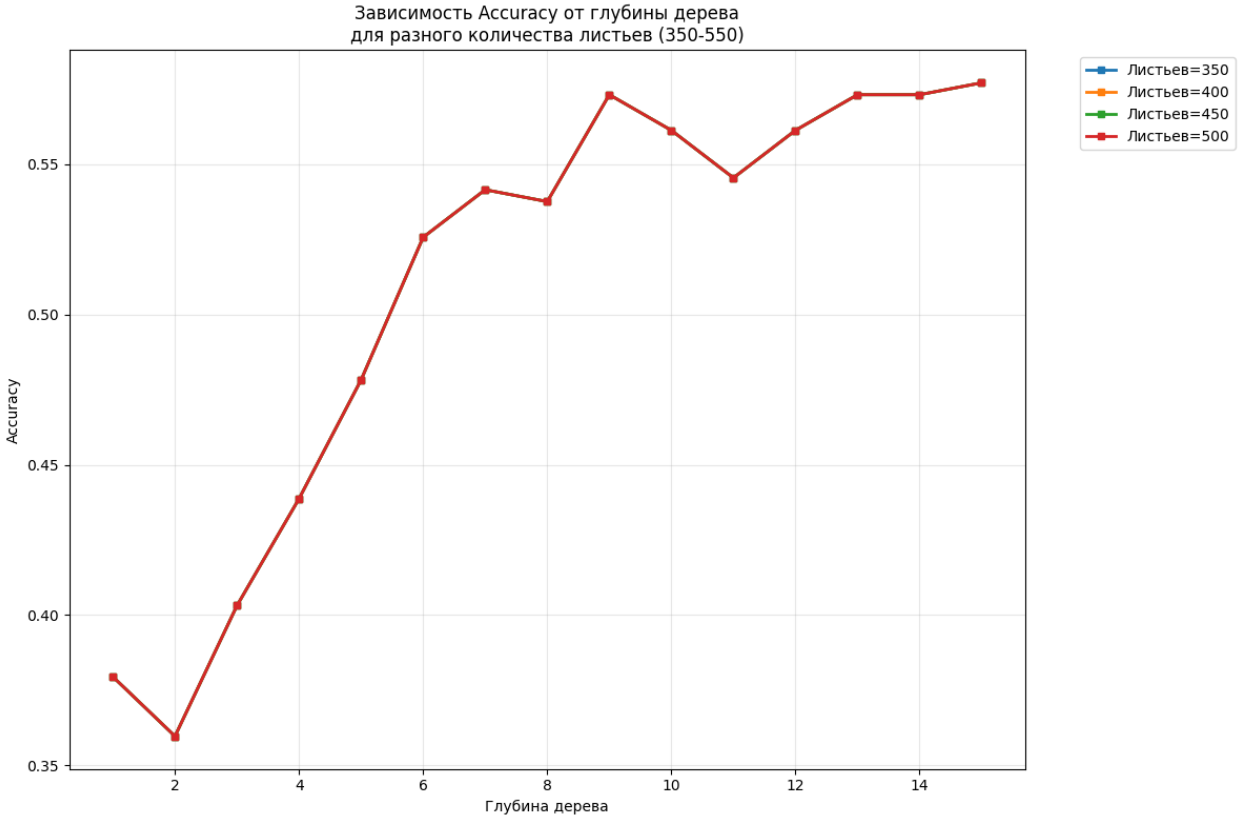
```
max_depth      |     max_leaf_nodes      |   accurancy
----------------------------------------------------------------------------------------------------
1              |     350             |   0.3794466403162055
1              |     351             |   0.3794466403162055
1              |     352             |   0.3794466403162055
1              |     353             |   0.3794466403162055
1              |     354             |   0.3794466403162055
1              |     355             |   0.3794466403162055
1              |     356             |   0.3794466403162055
1              |     357             |   0.3794466403162055
1              |     358             |   0.3794466403162055
1              |     359             |   0.3794466403162055
1              |     360             |   0.3794466403162055
1              |     361             |   0.3794466403162055
1              |     362             |   0.3794466403162055
1              |     363             |   0.3794466403162055
1              |     364             |   0.3794466403162055
1              |     365             |   0.3794466403162055
1              |     366             |   0.3794466403162055
1              |     367             |   0.3794466403162055
1              |     368             |   0.3794466403162055
1              |     369             |   0.3794466403162055
1              |     370             |   0.3794466403162055
1              |     371             |   0.3794466403162055
```

15	524	0.5770750988142292
15	525	0.5770750988142292
15	526	0.5770750988142292
15	527	0.5770750988142292
15	528	0.5770750988142292
15	529	0.5770750988142292
15	530	0.5770750988142292
15	531	0.5770750988142292
15	532	0.5770750988142292
15	533	0.5770750988142292
15	534	0.5770750988142292
15	535	0.5770750988142292
15	536	0.5770750988142292
15	537	0.5770750988142292
15	538	0.5770750988142292
15	539	0.5770750988142292
15	540	0.5770750988142292
15	541	0.5770750988142292
15	542	0.5770750988142292
15	543	0.5770750988142292
15	544	0.5770750988142292
15	545	0.5770750988142292
15	546	0.5770750988142292
15	547	0.5770750988142292
15	548	0.5770750988142292
15	549	0.5770750988142292

4. Анализ результатов

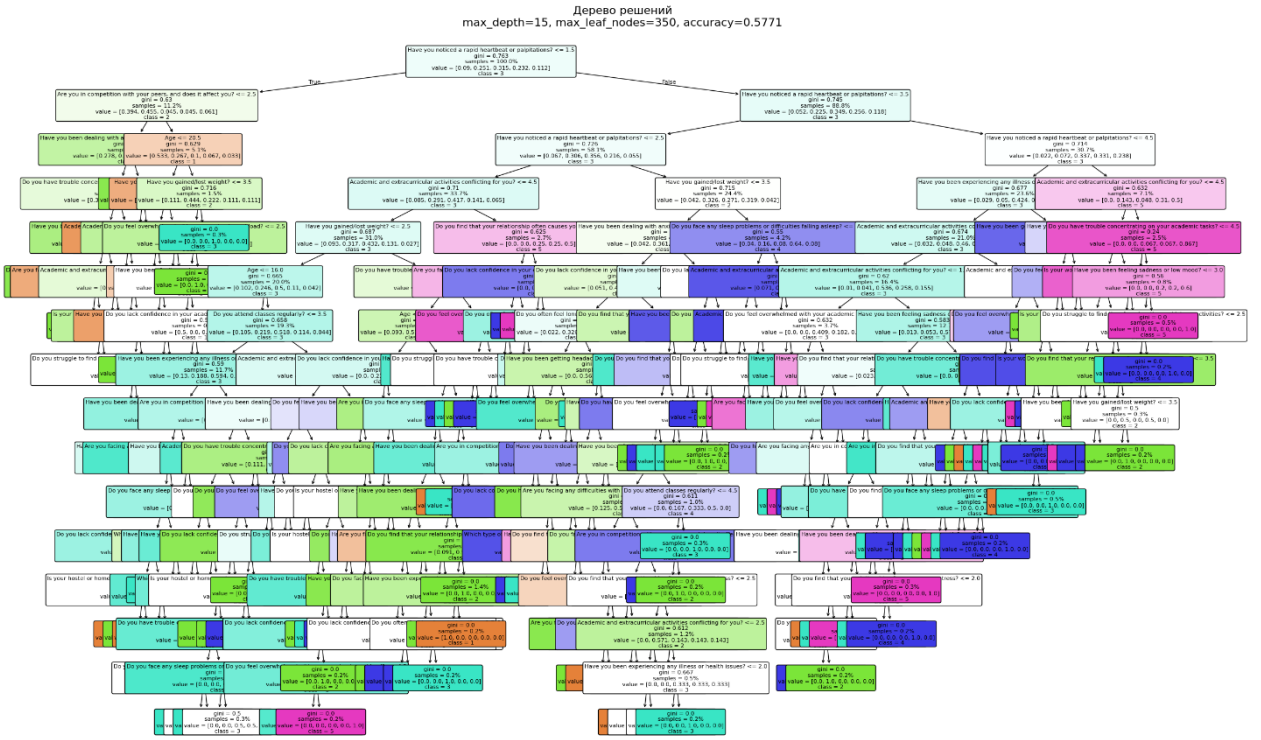
Максимальная точность 0.5770750988142292, то есть примерно 57.7%, на 0.04% больше, чем у базового дерева.





5. Визуализация

Полное дерево



Первые 4 уровня