

# Депрессия у Студентов

**Gender**

**City**

**Profession**

**Sleep Duration**

**Dietary Habits**

**Degree**

**Suicidal Thoughts?**

**Family History of  
Mental Illness**



# **Depression. 1 or 0?**

**Age**

**Academic Pressure**

**Work Pressure**

**GPA**

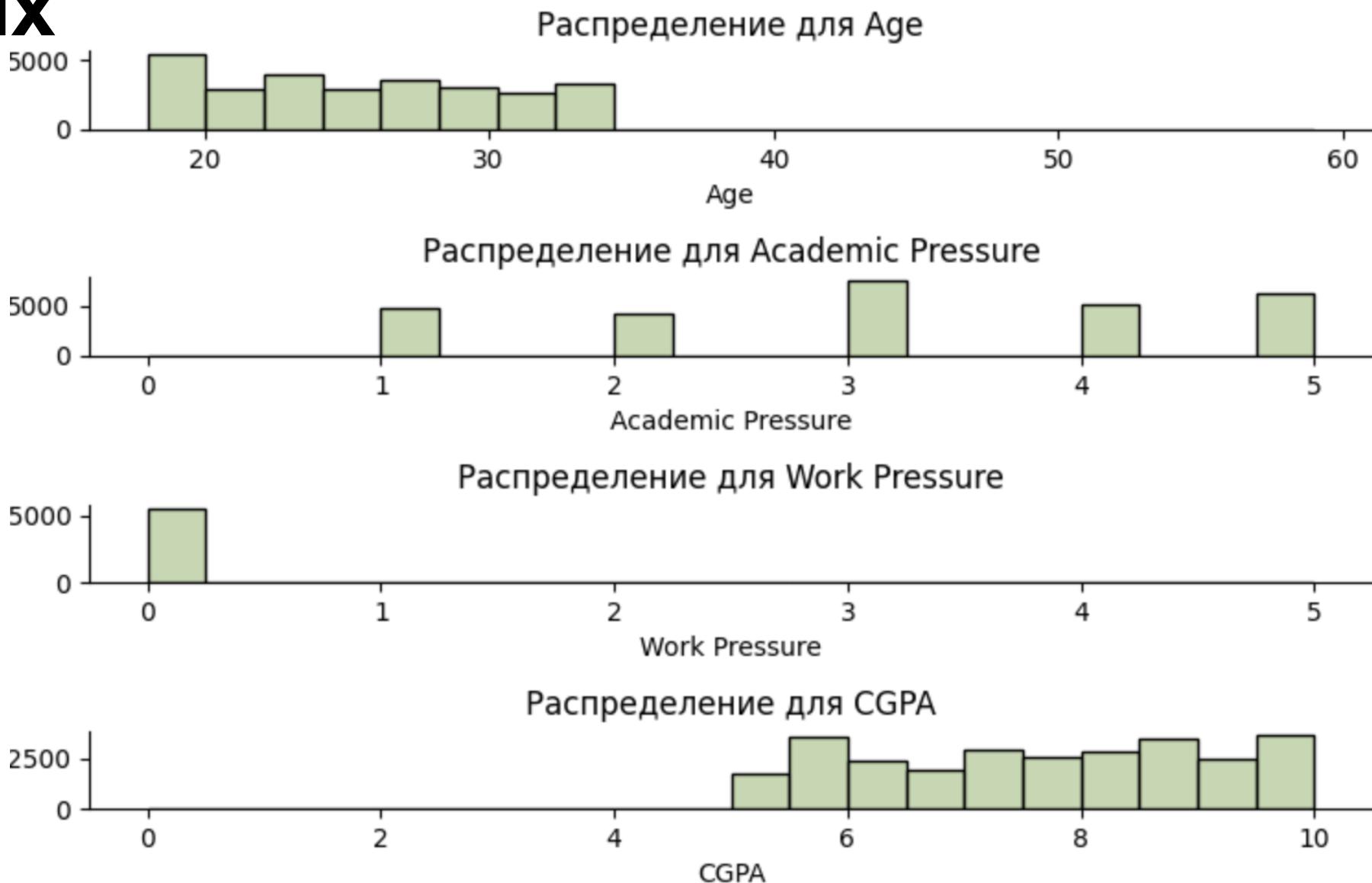
**Study Satisfaction**

**Job Satisfaction**

**Work/Study Hours**

**Financial Stress**

# Распределение Количественных Переменных

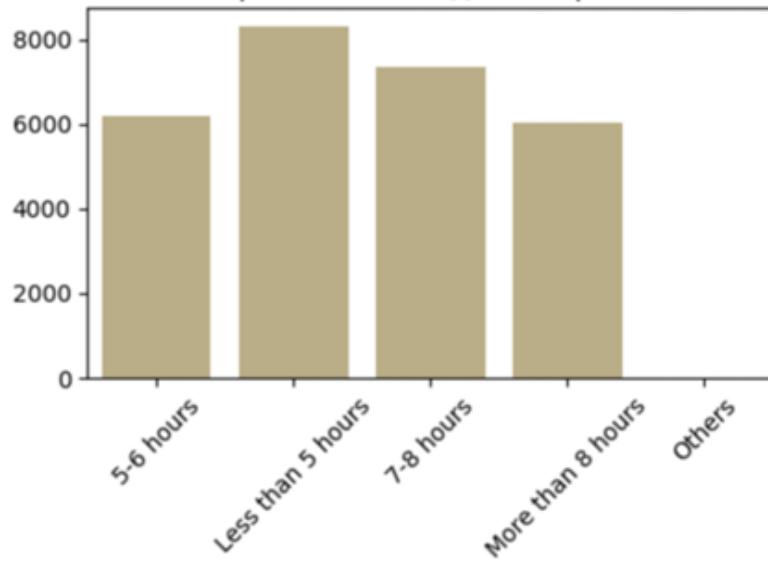


# Распределение Количественных Переменных

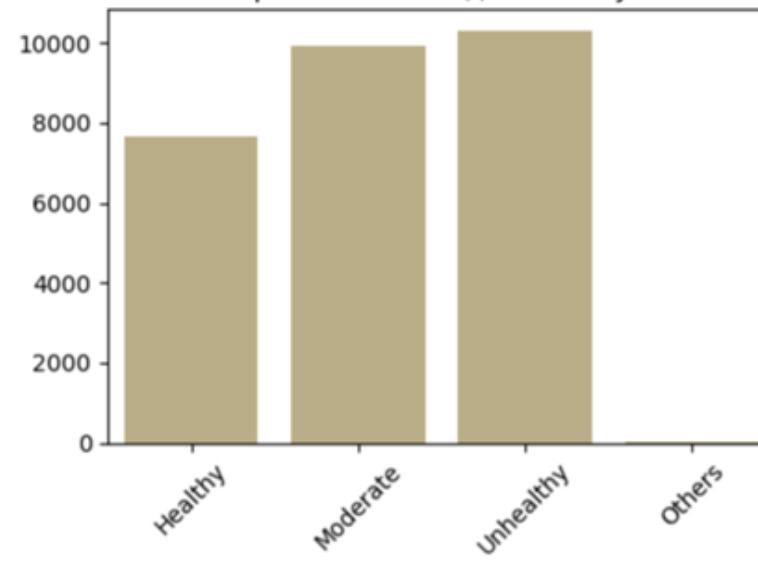


# Распределение категориальных переменных

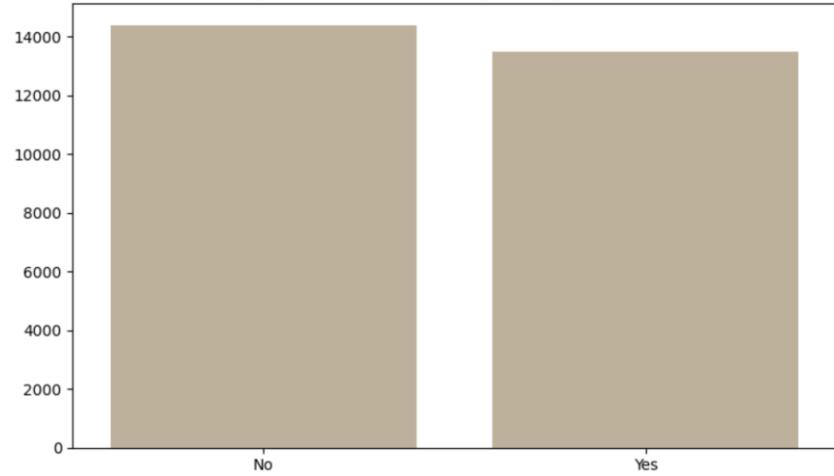
Гистограмма чистот для Sleep Duration



Гистограмма чистот для Dietary Habits

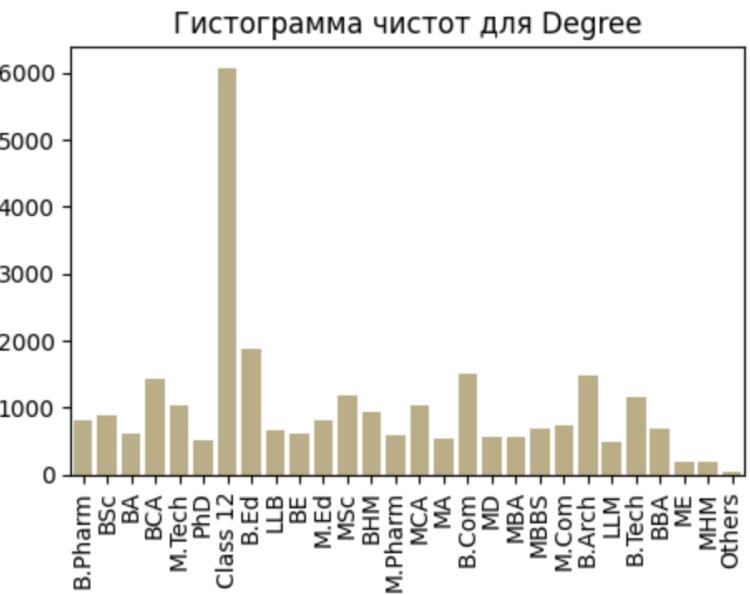
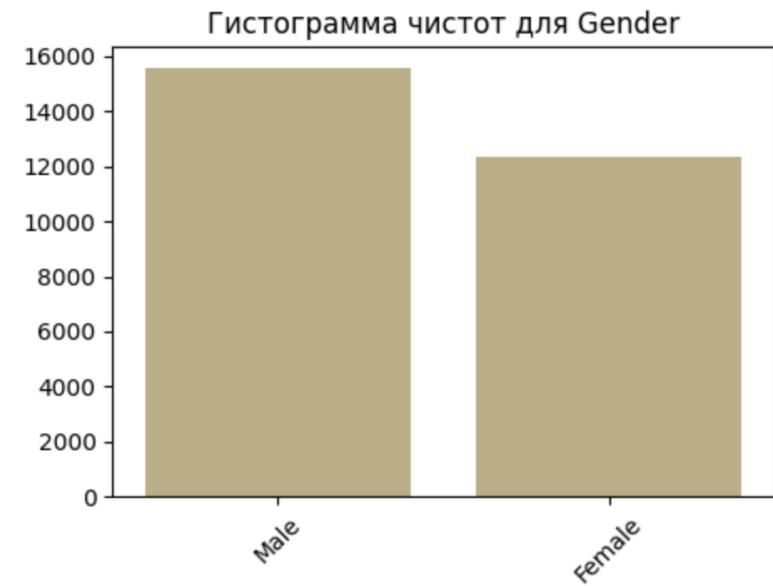
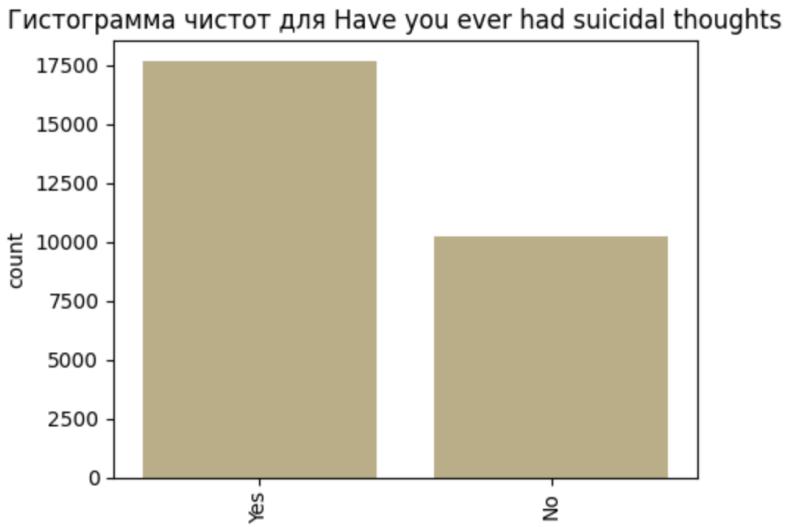


Гистограмма чистот для Family History of Mental Illness



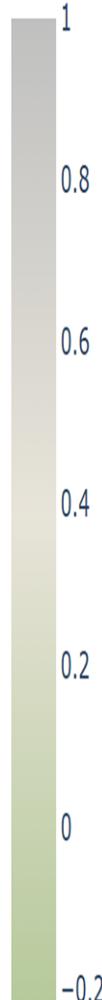
# Распределение

## категориальных переменных



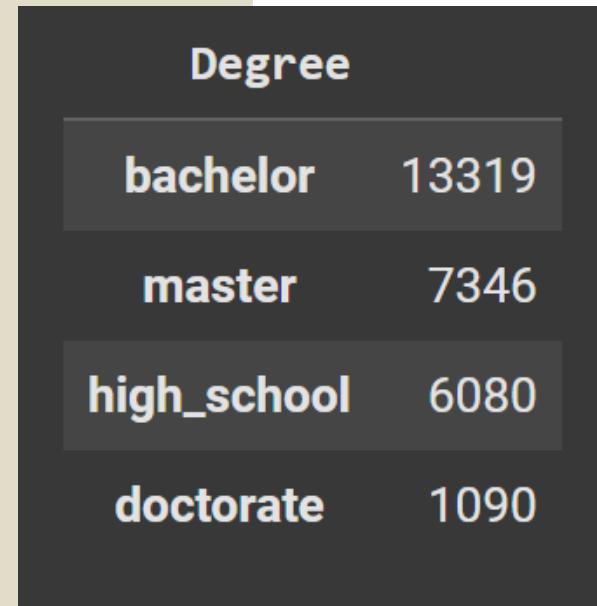
# Корреляционная матрица

	Age	Academic Pressure	CGPA	Study Satisfaction	Study Hours	Financial Stress	Depression
Age	1	-0.07587061	0.005167092	0.009131312	-0.03277535	-0.09525441	-0.2267377
Academic Pressure	-0.07587061	1	-0.02232527	-0.1109034	0.09629085	0.1517645	0.4748054
CGPA	0.005167092	-0.02232527	1	-0.04422728	0.002612736	0.005754528	0.02216402
Study Satisfaction	0.009131312	-0.1109034	-0.04422728	1	-0.03651004	-0.06522977	-0.168132
Study Hours	-0.03277535	0.09629085	0.002612736	-0.03651004	1	0.0754559	0.2090243
Financial Stress	-0.09525441	0.1517645	0.005754528	-0.06522977	0.0754559	1	0.363677
Depression	-0.2267377	0.4748054	0.02216402	-0.168132	0.2090243	0.363677	1



# Рассмотрение отдельных категорий

- Признак **Profession** имеет 14 уникальных категорий, в которой самая многочисленная – студенты (27870), оставшиеся 31 студент распределены по остальным 13 категориям, что в связи с незначительным объемом данных, позволяет не рассматривать данных студентов
- Признак **Degree** имеет 28 уникальных категорий, которые я распределила на 4 различных класса



Gender	0
Age	0
City	0
Academic Pressure	0
CGPA	0
Study Satisfaction	0
Sleep Duration	0
Dietary Habits	0
Degree	35
Have you ever had suicidal thoughts ?	0
Study Hours	0
Financial Stress	3
Family History of Mental Illness	0
Depression	0

# Работа с пропущенными значениями

- **Degree** – категориальный тип данных, я заполняла пропущенные значения модой по данному признаку
- **Financial Stress** – численный тип данных, я заполняла пропущенные значения с помощью **knn-imputer**

## Knn-imputer

это инструмент, который заполняет пропущенные значения. Он ищет k ближайших соседей (по другим доступным признакам), используя заданную метрику расстояния. Далее, пропущенное значение заполняется средним значением соответствующих соседей.

# Feature Scaling and One-Hot Encoding

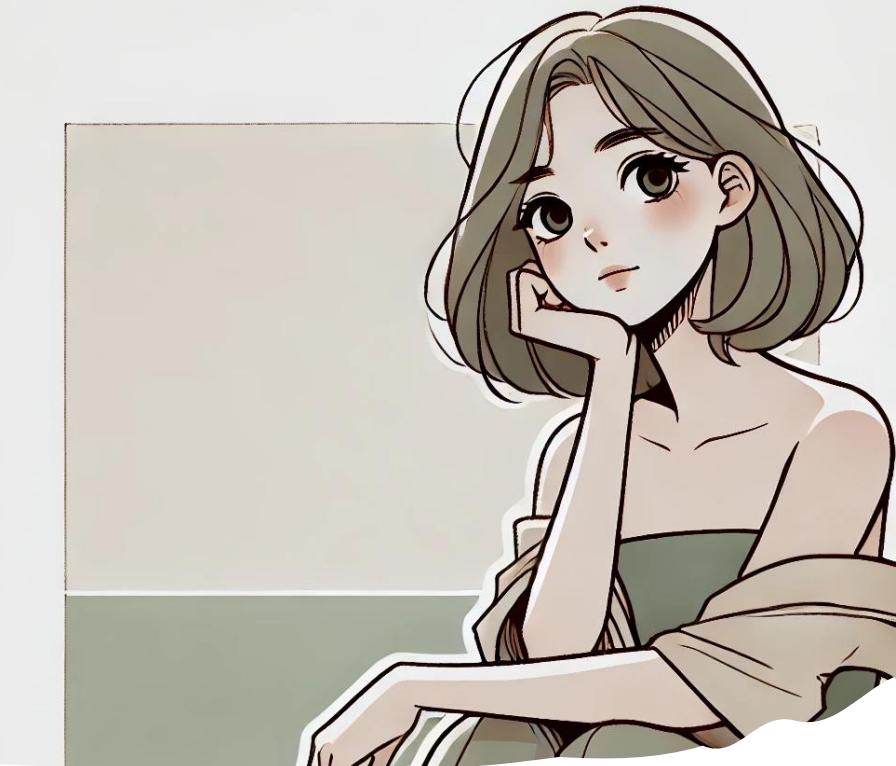
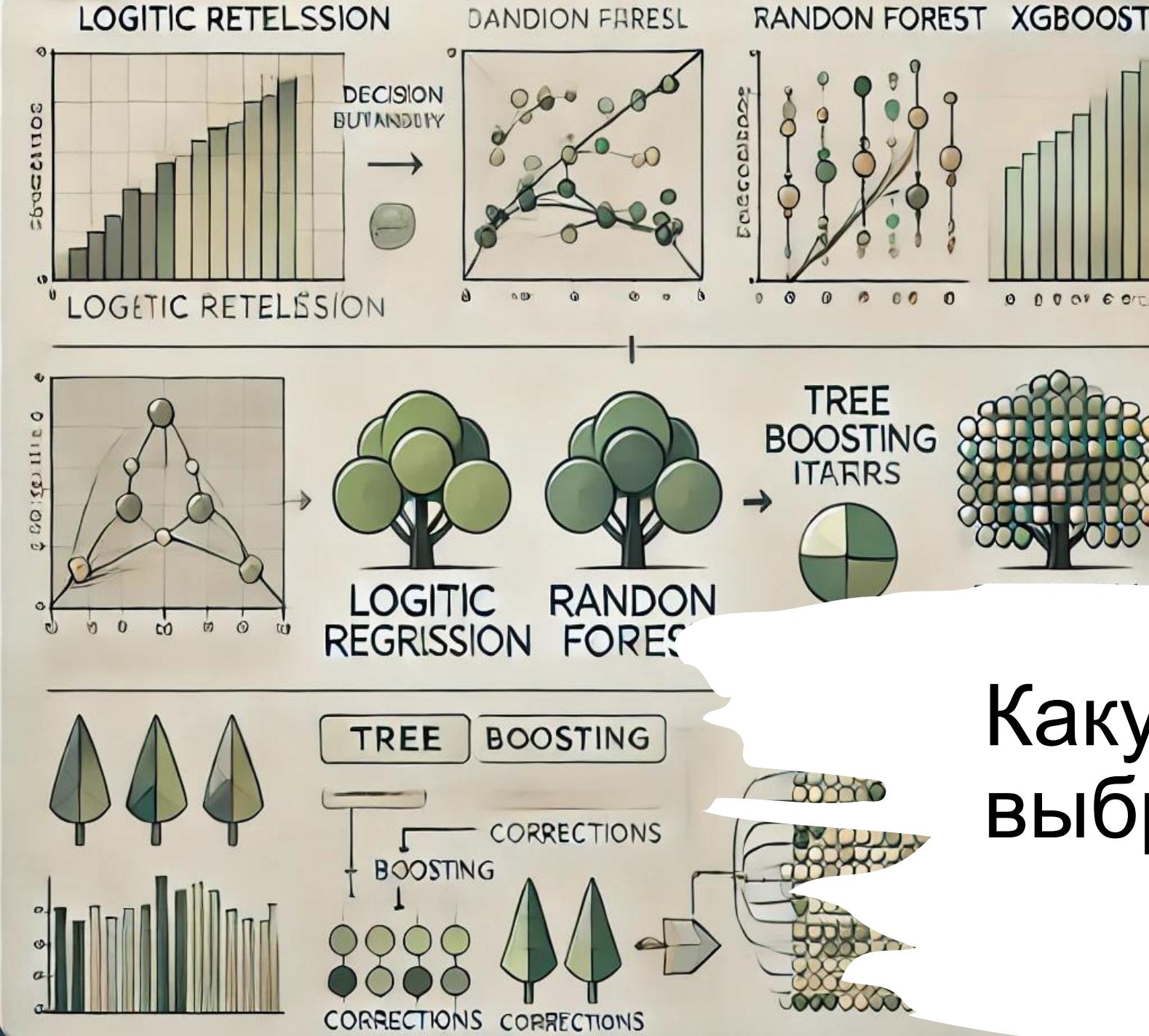
## Feature Scaling

(масштабирование признаков) — это процесс преобразования данных так, чтобы все признаки имели одинаковый масштаб.

## One-Hot Encoding

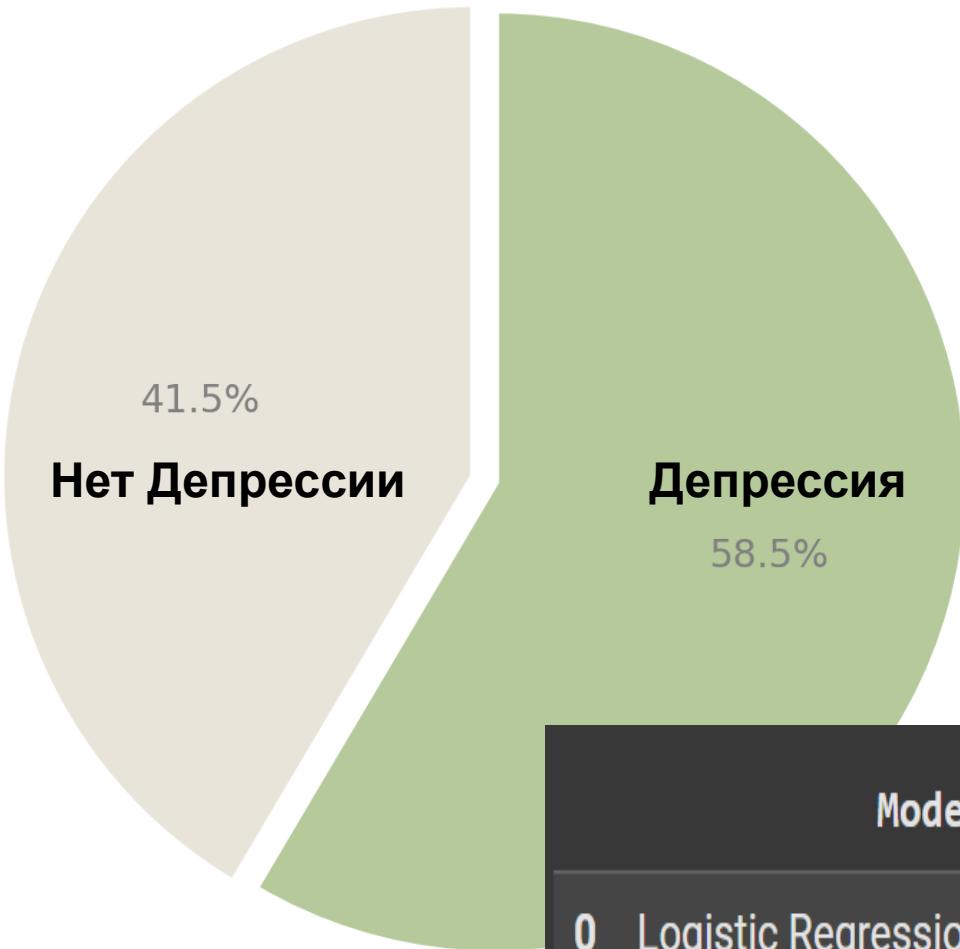
преобразует каждую категорию в отдельный столбец (или вектор), используя бинарное значение (0 или 1).

Если запись принадлежит категории, её значение становится 1, в противном случае — 0.



# Какую модель выбрать?

# Распределение Целевой переменной



Разница, между классами составляет 17 процентов. И такое распределение можно считать пости сбалансированным. Так что использование метрики

**Accuracy** вполне уместно.

Но если учесть, что мы пытаемся диагностировать болезнь, воспользуемся дополнительными метриками:

**ROC-AUC и F1-Score**

	Model	Accuracy	AUC	F1-Score	Best Params
0	Logistic Regression	0.847686	0.921601	0.871617	{'C': 0.1, 'penalty': 'l2'}
1	Random Forest	0.839792	0.917206	0.865815	{'max_depth': 20, 'min_samples_split': 10, 'n_estimators': 100, 'random_state': 42}
2	XGBoost	0.850556	0.923200	0.874416	{'learning_rate': 0.1, 'max_depth': 2, 'n_estimators': 100, 'random_state': 42}