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

# Лабораторная работа №4

## «ЛОГИЧЕСКИЕ МЕТОДЫ КЛАССИФИКАЦИИ»

Выполнил студент:

Володин Александр Сергеевич

Группа: ИА-232

Проверил: Брагин К.И.

Вариант: -

Новосибирск 2025

# Цели и задачи

***Цель лабораторной работы:***

*изучение принципов построения информационных систем с использованием логических методов классификации.*

***Основные задачи:***

- – освоение технологии внедрения алгоритмов на основе решающих списков в приложения;
- – освоение технологии внедрения алгоритмов на основе решающих деревьев в приложения;
- – изучение параметров логической классификации;
- – освоение модификаций логических методов классификации.

# Индивидуальное задание

Мной был выбран и согласован набор данных - “Land Mine

Данные загружены из файла diabetes_dataset.csv. Этот набор данных содержит 16 признаков, среди которых значения медицинских параметров, такие как уровень глюкозы, давление, индекс массы тела и другие, а также целевую переменную Outcome, которая принимает значения 0 (отсутствие диабета) или 1 (наличие диабета).

Признаки:

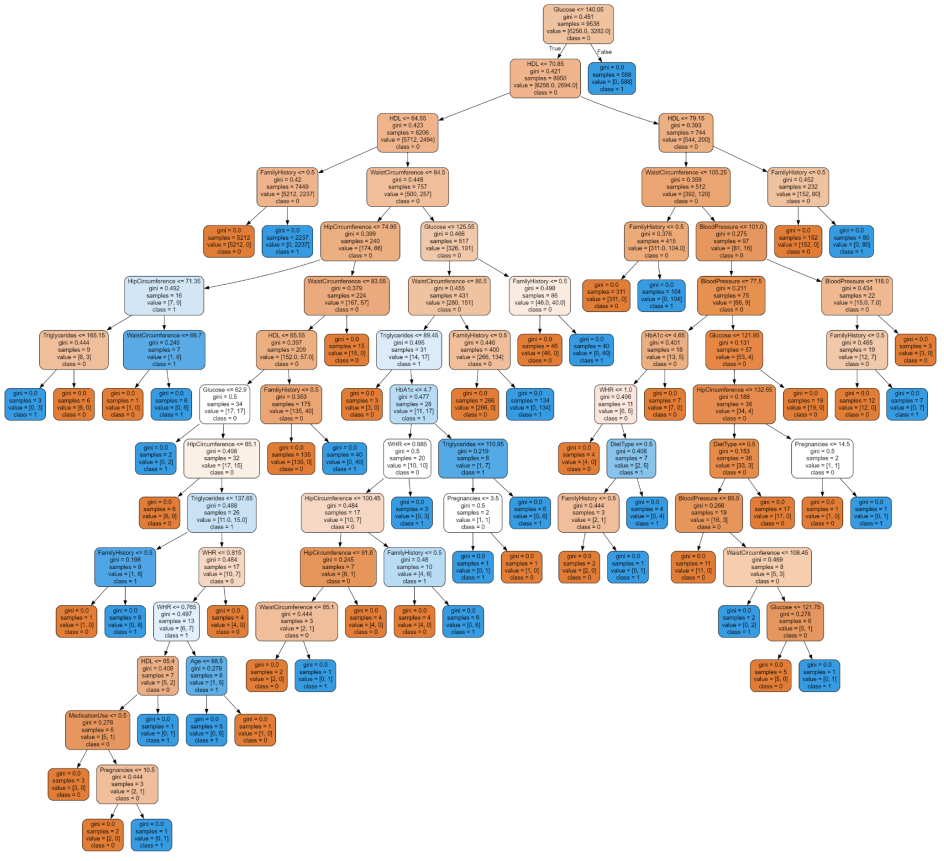
- Глюкоза,
- Давление,
- Индекс массы тела,
- Возраст и другие.

```
   Age  Pregnancies    BMI  Glucose  ...  FamilyHistory  DietType  Hypertension  MedicationUse
0   69            5  28.39    130.1  ...              0         0             0              1
1   32            1  26.49    116.5  ...              0         0             0              0
2   89           13  25.34    101.0  ...              0         0             0              1
3   78           13  29.91    146.0  ...              0         0             0              1
4   38            8  24.56    103.2  ...              0         1             0              0

[5 rows x 16 columns]
0    0
1    0
2    0
3    1
4    0
Name: Outcome, dtype: int64
Accuracy = 0.7211740041928721
Optimal max depth: [16]
Fitting 10 folds for each of 57 candidates, totalling 570 fits

Best parameters: {'max_depth': 17, 'max_features': 3}
Best cross-validation score: 0.9817600163667203
```

Обучение модели решающего дерева

## Визуализация данных

Для анализа данных были использованы следующие методы визуализации:



Матрица корреляции для оценки взаимосвязи между признаками:
sns.heatmap(data.corr(), cmap=plt.cm.Blues, annot=True, fmt='.2f')

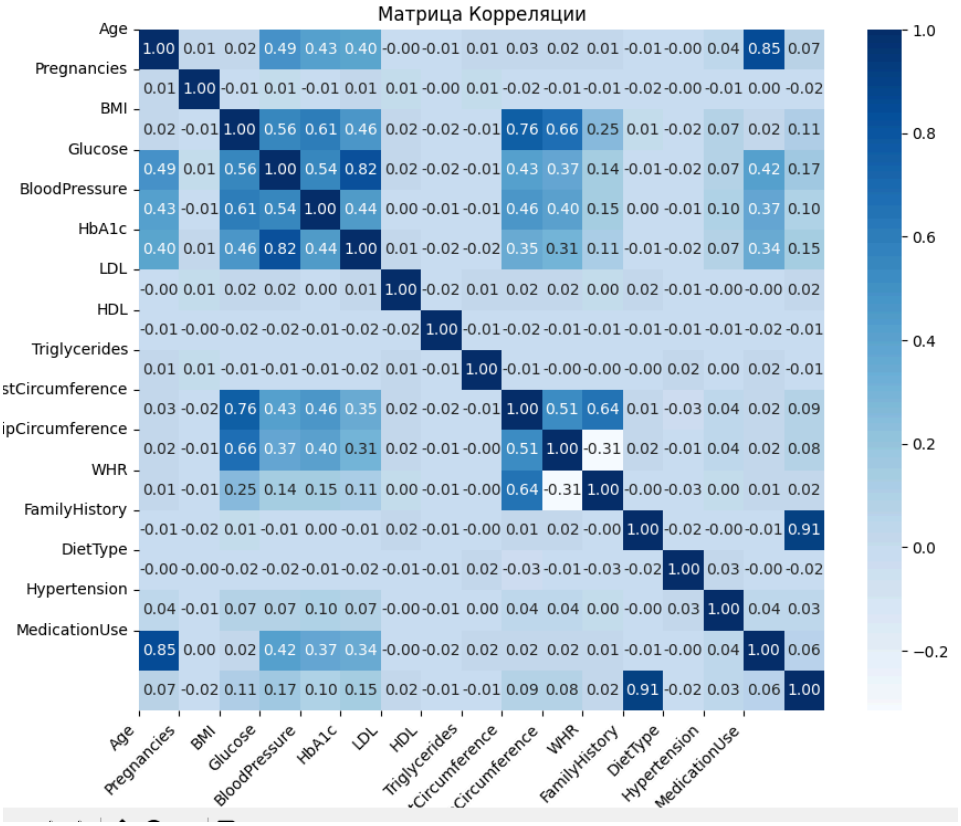
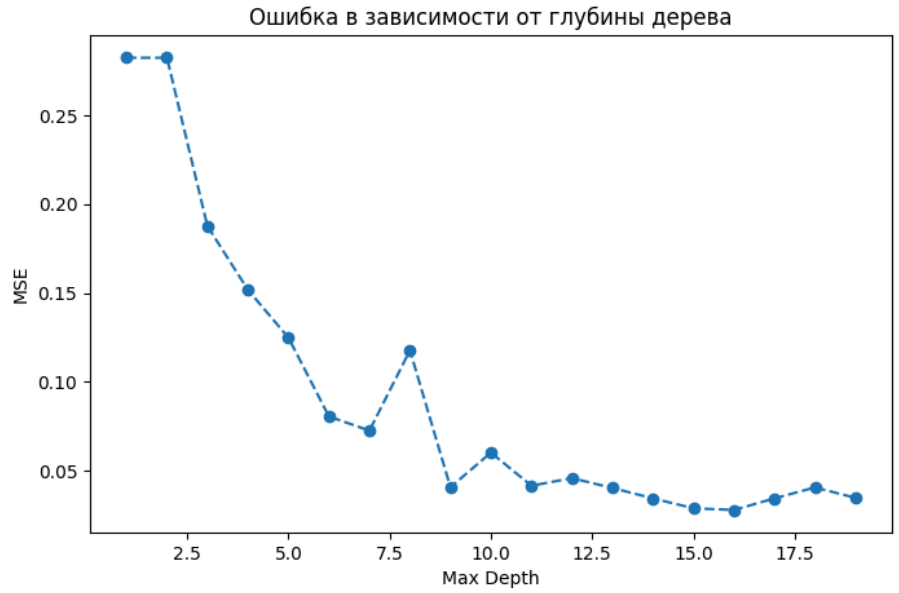



График ошибки классификации в зависимости от глубины дерева:
plt.plot(d_list, MSE, marker='o', linestyle='dashed')




Визуализация дерева решений с использованием export_graphviz:
export_graphviz(tree_grid.best_estimator_, feature_names=dataX.columns, class_names=[str(c) for c in dataY.unique()])

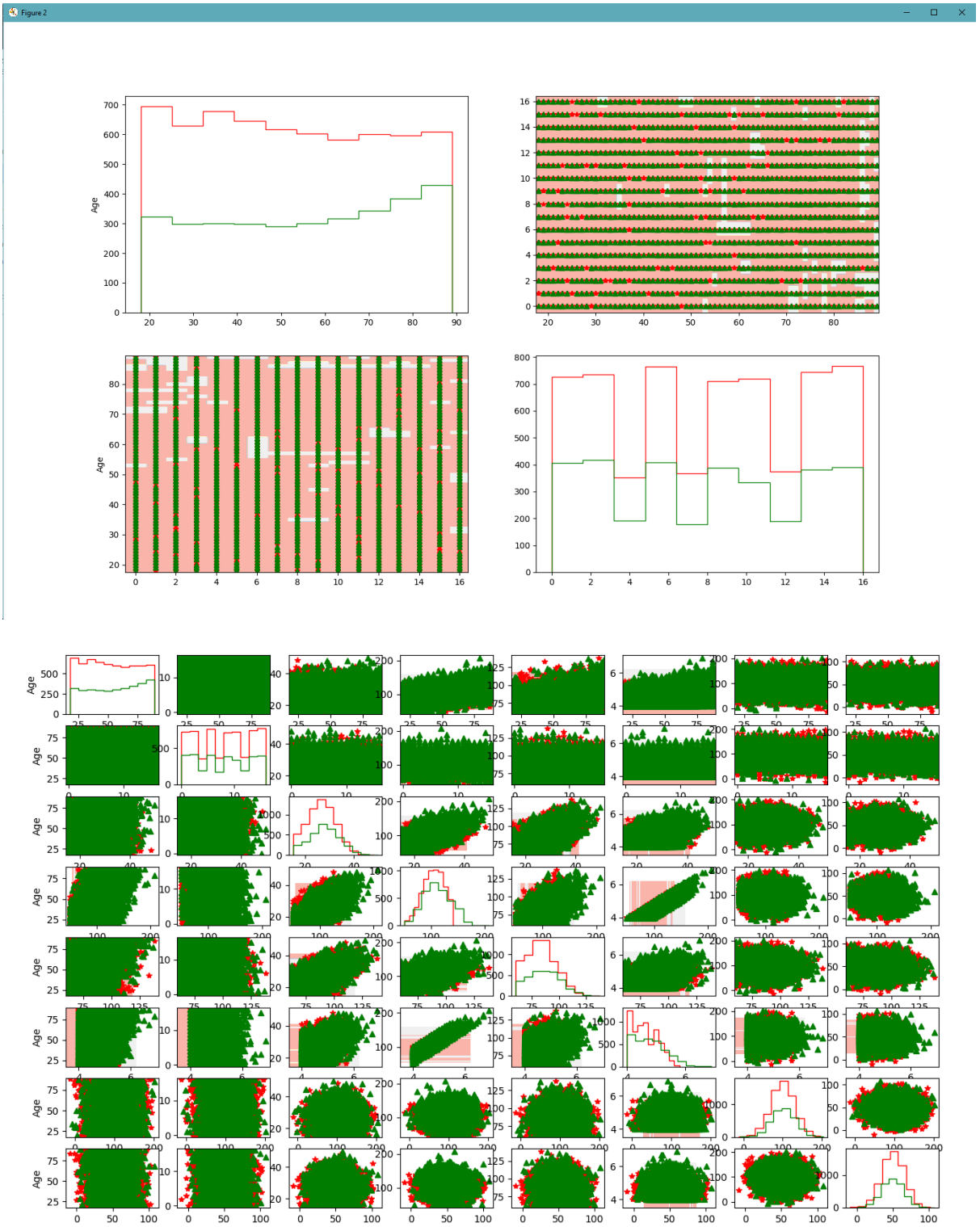
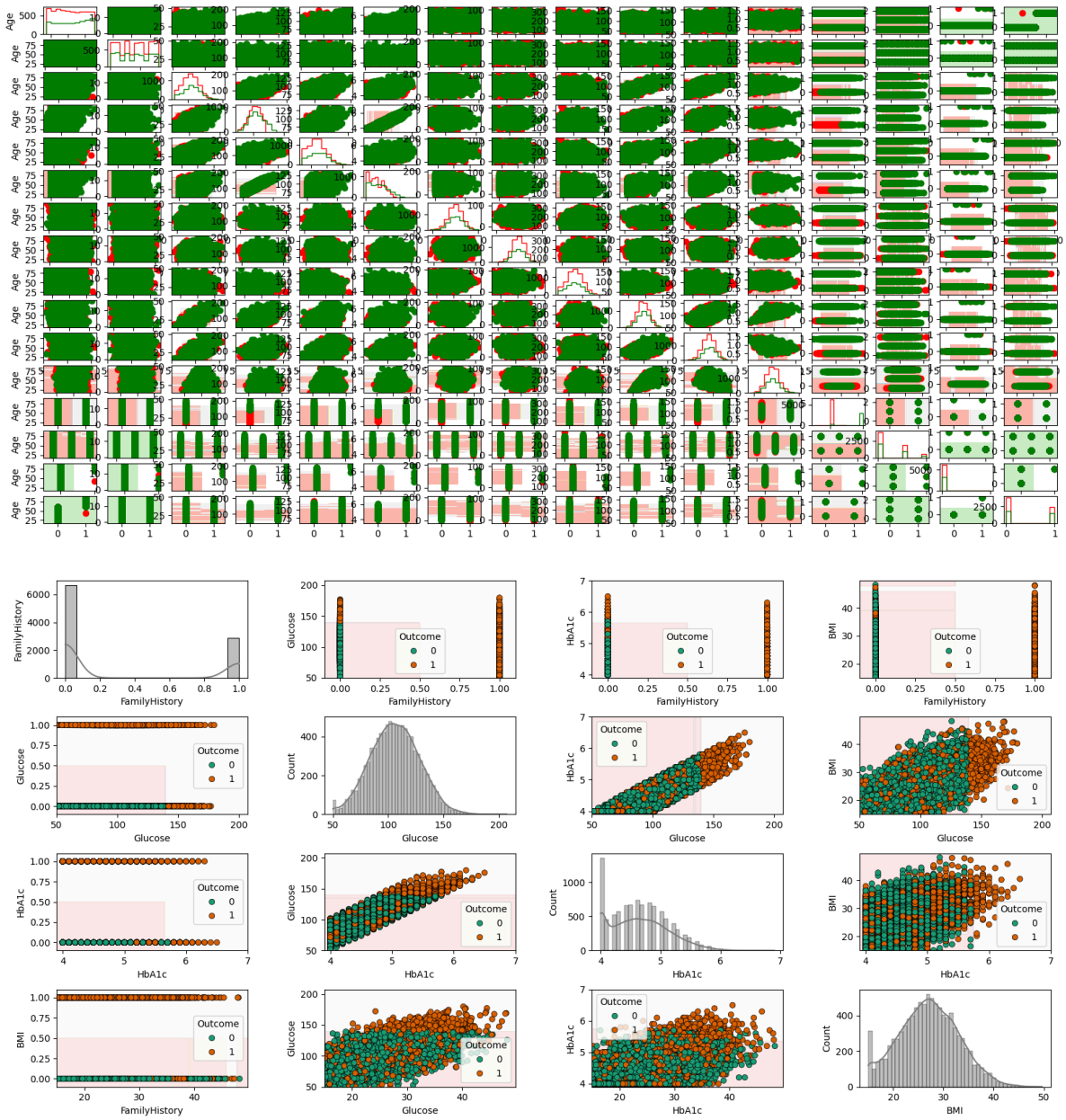


Figure 2
Age



## Выводы

- Модель решающего дерева показала точность 72% на тестовой выборке при параметре глубины дерева, равном 5.
- Проведена оптимизация глубины дерева и гиперпараметров с использованием кросс-валидации и GridSearchCV, что позволило найти оптимальные параметры.
- Визуализация данных, включая матрицу корреляции и графики ошибок, позволила наглядно понять зависимость точности модели от параметров.
- Результаты показали, что глубина дерева влияет на точность классификации, и оптимальное значение параметра было найдено на основе минимальной ошибки.

## Контрольные вопросы

**Ответы на контрольные вопросы:**

1. **Принцип построения дерева решений**: Дерево решений — это модель машинного обучения, которая используется для решения задач классификации и регрессии. Принцип построения дерева заключается в разделении исходного набора данных на подмножества с помощью простых логических условий (на основе признаков), с целью максимизации различий между классами в конечных листах дерева. Процесс построения дерева включает следующие шаги:
    - Выбирается признак, который наиболее эффективно разделяет данные на подмножества.
    - Для каждого подмножества рекурсивно выбирается новый признак для дальнейшего разделения.
    - Этот процесс продолжается до тех пор, пока не будет достигнут критерий остановки, такой как максимальная глубина дерева или минимальное количество объектов в узле.
2. **Статистическое определение информативности**: В статистике информативность признака или переменной определяется как мера того, насколько хорошо этот признак или переменная помогает разделить или предсказать целевую переменную. Одним из стандартных способов измерения информативности является использование **критерия информации** или **энтропии**, которая вычисляется для разных значений признаков. Чем выше различие между классами на основе данного признака, тем выше его информативность.
3. **Энтропийное определение информативности**: Энтропия — это мера неопределенности или беспорядка в данных. В контексте дерева решений, энтропия используется для измерения "нечистоты" узла. Чем меньше энтропия в узле, тем более информативным является признак, использованный для его разделения. Энтропия для набора данных вычисляется по формуле:
4. **Многоклассовая информативность**: Многоклассовая информативность — это мера, которая используется для оценки

способности признаков разделять данные на более чем два класса. В случае многоклассовой классификации мы должны искать признаки, которые наиболее эффективно различают большее количество классов. Многоклассовая информативность применяется в задачах, где целевая переменная может принимать более двух значений (например, классификация изображений по категориям, распознавание типов объектов и т.д.). В дереве решений для многоклассовой задачи применяется модификация критериев информативности, таких как энтропия или индекс Джини, которые учитывают несколько классов.

5. **Назначение и алгоритм бинаризации количественных признаков**: Бинаризация количественных признаков — это процесс преобразования непрерывных числовых признаков в бинарные (дискретные) переменные. Этот процесс используется для того, чтобы модель могла работать с такими признаками, превращая их в "да/нет" или "больше/меньше" форматы. Алгоритм бинаризации обычно заключается в установке порога для каждого признака, при котором все значения больше порога становятся 1, а значения, меньшие или равные порогу, становятся 0. Например:
    - Признак "Возраст" может быть бинаризован порогом 40 лет, где 1 — если возраст больше 40 лет, и 0 — если меньше или равен 40. Это используется, чтобы упростить обработку данных и создать более понятные и интерпретируемые признаки для модели.
6. **Порядок поиска закономерностей в форме конъюнкций**: Поиск закономерностей в форме конъюнкций обычно происходит в процессе построения логических моделей, таких как решающие деревья или алгоритмы на основе правил. Конъюнкция — это логическая операция, которая объединяет два или более условия. Алгоритм построения таких закономерностей предполагает поэтапное объединение признаков с учетом их значений и формулировки условий для каждого класса. Например, для классификации пациентов с диабетом, модель может искать закономерности в форме:
    - Если уровень глюкозы больше 120 и возраст больше 50 лет, то вероятно, что пациент имеет диабет. Порядок поиска заключается в том, чтобы на каждом шаге объединять

признаки, которые наилучшим образом разделяют данные на классы, проверяя их соответствие для каждого условия. Этот процесс продолжается до тех пор, пока не будут найдены все значимые закономерности.