Уважаемый кандидат!

Мы рады пригласить вас пройти тестовое задание, которое поможет нам оценить ваши навыки и опыт в области машинного обучения и анализа данных с акцентом на хемоинформатику. Это задание позволит вам продемонстрировать ваше понимание работы с химическими структурами, навыки разработки моделей и умение решать задачи, связанные с предсказанием биологической активности молекул.

В рамках этого задания вам предстоит построить классификационные модели, используя графовые нейросети и алгоритм XGBoost. Вам будет необходимо разделить данные, провести подбор гиперпараметров и оценить точность моделей на тестовом наборе данных. Мы рассчитываем, что вы выполните задание в течение 1 недели, однако, даже частично выполненное задание даст нам представление о вашем подходе к решению задачи.

Мы ценим не только правильность выполнения задания, но и ваш подход к анализу данных, выбор моделей, их оптимизацию и документирование процесса. Полное и четко структурированное решение будет рассматриваться как значительное преимущество.

Если у вас возникнут вопросы по ходу выполнения задания, не стесняйтесь обращаться к нам. Мы с удовольствием предоставим вам необходимую поддержку и ответим на все вопросы.

Желаем вам удачи и с нетерпением ждем вашего решения!

Описание задачи:

У вас есть датасет с двумя колонками:

- **SMILES** (строковое представление химической структуры)
- **Activity** (бинарный целевой признак).

Ваша задача — построить классификационную модель для предсказания активности молекул. Для этого вам нужно обучить несколько моделей, включая графовые нейросети (GCN, GAT, GIN) и модель XGBoost на основе молекулярных отпечатков (ECFP4).

Шаги выполнения:

1. **Разделение данных:**
    - Разделите исходный датасет на три части в соотношении 3:1:1:
        - **Трейн** (60% данных): для обучения моделей.
        - **Валидация** (20% данных): для подбора гиперпараметров.
        - **Тест** (20% данных): для оценки финальной модели.
2. **Предобработка данных:**
    - Для графовых нейросетей преобразуйте SMILES в графовые представления.
    - Для XGBoost создайте молекулярные отпечатки (2048-битные ECFP4).
3. **Модели для обучения:**
    - **Графовые нейросети**: обучите модели GCN, GAT, GIN на графовых представлениях молекул.
    - **XGBoost**: обучите модель XGBoost на 2048-битных ECFP4 отпечатках.
4. **Подбор гиперпараметров:**

- Для каждой модели проведите подбор гиперпараметров на валидационном датасете. Вы можете использовать grid search или random search для поиска оптимальных значений.
- **Основные гиперпараметры для подбора включают:**
  - **Графовые нейросети (GCN, GAT, GIN):**
    - **Число слоев**: [2, 3, 4] + 1 линейный слой (в конце).
    - **Размер скрытого слоя** (hidden_dim): [64, 96, 128, 256].
    - **Dropout**: [0, 0.2].
    - **Learning Rate**: [0.001, 0.003, 0.01].
    - **Функция активации**: ELU.
    - **Оптимизатор**: Adam.
    - **Количество MLP слоев GIN**: [1, 2, 3]
  - **XGBoost:**
    - **Число деревьев** (n_estimators): [300, 500, 1000, 1500, 2000].
    - **Максимальная глубина дерева** (max_depth): [3, 5, 7, 10].
    - **Минимальное количество образцов на лист** (min_child_weight): [1, 5, 10].
    - **Learning Rate**: [0.1, 0.01, 0.001].
    - **Параметр регуляризации L2** (lambda): [0, 0.2, 1, 5].
    - **Параметр регуляризации L1** (alpha): [0, 0.2, 1, 5].
    - **Colsample bytree**: [0.5, 0.8, 1].
    - **Subsample**: [0.5, 0.7, 1.0].

5. **Оценка модели:**
   - После подбора гиперпараметров, обучите модель на трейн датасете с использованием оптимальных настроек и оцените её на тестовом датасете.
   - Замерьте следующие метрики, важные для классификации (особенно в условиях дисбаланса классов):
     - **Accuracy**: точность.
     - **Precision**: точность по положительному классу.
     - **Recall**: полнота по положительному классу.
     - **F1-score**: гармоническое среднее precision и recall.
     - **ROC-AUC**: площадь под ROC кривой.
     - **PR-AUC**: площадь под PR кривой (precision-recall).
6. **Результаты:**

- Сравните результаты всех моделей по вышеуказанным метрикам. Сформируйте таблицу, где по оси Y расположены 4 модели (3 графовые + XGBoost), по оси X метрики.
- Сформируйте выводы о том, какая модель показала наилучшие результаты и почему.

Требования к оформлению:

- Предоставьте отчет, включающий описание использованных подходов, гиперпараметров, результаты по каждой модели и их анализ.
- Код должен быть предоставлен в Jupyter Notebook или в виде Python скрипта.
- Убедитесь, что код хорошо задокументирован и может быть запущен для проверки.