Идеи решения задачи предсказания ССЗ

Группа 1

1. Дополнительный признак по категории людей (физическая активность, употребление алкоголя, курение), категории:
2. Результаты измерения: сконструировать общий признак (например, группа: норм. давление и нормальные анализы по холестерину и уровню глюкозы).
3. Зависимость от возраста.
4. Зависимость от веса.
5. Общий признак от веса и возраста – ввести признак индекс массы тела.
6. Попробовать найти зависимость по разнице верхнего и нижнего давления.

Группа 2

1) удалить id
2) возраст перевести в года/ разделить на возрастные группы / использовать для новых признаков (формула давления)
3) пол – 2 мужской и 1 женский / использовать для новых признаков (имп)
4) даваление

4.1) обработка: модуль значения / определить границы нормального давления / обработать выбросы, заменив их либо общим медианным занчением, либо медианным значением для соответсвующей возрастной группы
4.2) использовать для новых признаков: верхнее = 109 + 0.5 * age + 0.1 * weight / нижнее = 63 + 0.1*age + 0.15 * weight

5) курение – None = 0 / предсказание
6) алкоголь None = Smoke / предсказание

Группа 3

1. Индекс массы тела, категоризировать

2. Ввести рамки давления на категории в зависимости от связки ap hi/ap_lo.

Новый признак в зависимости от их взаимосвязи на 4 категории

Поискать в интернете

3. Данные по алкоголю, курению, активности инвертированы

4. Глюкоза, холестерин являются вспомогательными признаками, не основными

5. Общая идея: восстановить значения аномально заполненных признаков по значениям других признакам.

Группа 4

1. Есть ли в данных пропуски? Если да, чем их можно заполнить?

- **Пропуски в Test в столбцах Smoke и Alco единицами, а Active нулями.**

2. Есть ли категориальные признаки? Нужно ли их использовать? Если да, то как их закодировать?

- **Нет (но Ксения так не считает)**

2. Какие новые признаки можно придумать для решения данной задачи?

- **Smoke и Alco сложить**
- **Создать комбинации Smoke, Alco & Ap_lo, Ap_hi, а также с Gluc и Cholesterol**
- **Вычислить пол, посмотреть статистику по толстым мужчинам**

3. Есть ли в данных выбросы? Как их найти? Что с ними делать (удалять/нет) и почему?

- **Выбросы есть (ищем по границам)**
- **Некорректные данные в давлении удалить**
- **Некорректные данные в весе и росте заменить средним по возрасту**

5. Какую или какие модели использовать для решения данной задачи?

- **Voting (XGboost + LightGBM)**
- **CatBoost**