Здравствуйте!

Меня зовут Воронкина Дарья

**1)** И я представляю доклад на тему: Прогнозирование времени дожития пациентов с боковым амиотрофиическим склерозом

**2)** БАС - прогрессирующее, неизлечимое дегенеративное заболевание центральной нервной системы, приводящее к параличам и последующей атрофии мышц. Болезнь, которая затронула таких популярных людей как Стивен Хокинг и Дмитрий Шостакович.

В большом кол-ве случаев болезнь связана с мутациями гена супероксиддисмутазы-1. Изучением мутаций в данном гене занимаются ученые из ИЦИГ, которые предоставили нам данные для построения модели. Данные содержат 182 признака водородных связей фермента SOD1 для 72 пациентов.

**3)** Мое исследование началось с изучения предоставленных данных, так как признаки водородных связей представлены донорно-акцепторным взаимодействием с описывающим весом этого взаимодействия, то для каждого пациента можно рассматривать фермент в виде графа.

**4)** Предполагается, что течение болезни влияет на структуру водородных связей в ферменте SOD1. Идея заключается в выявлении наличия зависимости между временем дожития пациента и изменением структуры фермента для этого будем считать характеристики графа. Рассматриваемое уравнение выглядит следующим образом, а модель строится с помощью метода наименьших квадратов.

**5)** Первый этап — это построение графа. Донор и акцептор представлен следующим образом: позиция аминокислотного остатка, имя субъединицы и названии атома. Учитывая такую сложную структуру, вершину можно рассматривать тремя различными способами:

**6)** 1) Вершина представлена в виде полной водородной связи, пример того как выглядит граф вы можете видеть на рисунке 1.

**7) 8)**

**9)** Следующий этап — это построение обучающих выборок. Для того чтобы построить обучающие выборки, для каждого из графов, которые принадлежат пациентам, были посчитаны характеристики этих графов. Я брала характеристики, описывающие в целом граф, вы можете видеть их на слайде

**10)** Построенные модели показали маленький коэффициент детерминации, что означает плохое описание данных. Это обусловлено тем, что Характеристики для всего графа не учитывают структуру конкретных подграфов, в которых может возникнуть мутация.

**11)** Поэтому посоветовавшись с учеными из ИЦИГ было принято решение рассматривать граф, где вершины представлены в виде полной водородной связи. Так как характеристики, посчитанные в целом для всего графа не дали результатов, было принято решение считать характеристики для каждой вершины этого графа, но выяснилось что так как в графе присутствует очень много изоморфных друг другу подграфов, например подграфы с двумя вершинами, в обучающей выборке возникает коллинеарность, что ведет к отсутствию значимых моделей и невозможности построить корректную модель.

**12)** Мы обсудили это с учеными из ИЦИГ и они сказали нам что в действительности изоморфные друг другу подграфы не несут нужной нам информации и так как нам интересна зависимость изменения структуры графа мы можем рассматривать подграфы только с уникальной структурой. То есть в модель для каждого пациента войдут только подграфы с уникальной структурой.

**13)**  Хочу заметить что так как пациентов 72 и все они разные люди с разной степенью тяжести течения болезни, то и структура фермента SOD1 у них будет разная, следовательно и уникальные подграфы будут разные. Давайте рассмотрим это на примере. Допустим у пациента 1 есть подграф структуры 1, но нету подграфа структуры 2, а также есть подграф структуры 3, но в отличии от пациента 2, наименование вершин этого подграфа другое. То есть может быть такая ситуация, что у пациента 1 есть подграф структуры 3 и имя его ABCD, а у пациента 2 имя GHNT, все это учитывается при построении модели и в общей обучающей выборке, вместо минуса будет стоять 0, а вместо + конкретное значение подсчитанной характеристики.

**14)** Далее для графа, состоящего из уникальных подграфов, обучающая выборка составлялась следующим образом: Для каждого подграфа считалась одна из представленных характеристик и на основе этой информации составлялась обучающая выборка. Я отобрала только выделенные жирным характеристики, так как остальные характеристики связаны между собой и из-за них возникает эффект мультиколлиниарности.

**15)** Таких обучающих выборок три. Для каждой из обучающих выборок строится модель. Построенные модели показали следующие результаты. Количество регрессоров это количество посчитанных характеристик для уникальных графов после удаления нулевых и линейно зависимых столбцов. Количество значимых регрессоров – это кол-во регрессоров с уровнем значимости меньше 0.05

**16)**Очевидно, модели с коэффициентом детерминации меньше 0.5 плохо описывают обучающую выборку, поэтому, чтобы улучшить качество модели я объединила три обучающие выборки, исключила линейнозависимые столбцы и получила модель с коэфициентом детерминации 0.515

**17)** Для исследования возможности использования данной модели была посчитана кросс валидация **и она равна 6.71.** Результат показал, что на данный момент использовать модель еще нельзя, но она уже гораздо лучше описывает данные.

**18)** Таким образом: прочитать то что в выводе написано

**19)** На данном этапе еще проводятся исследования. Сейчас передо мной стоит цель найти новый подходящий характеристики чтобы улучшить модель и рассмотреть другие методы машинного обучения

**20)** По данной работе была сделана публикация, которую вы можете видеть на слайде

**21)** Спасибо за внимание, я готова ответить на ваши вопросы