ТЕСТОВОЕ ЗАДАНИЕ

Предлагается рассмотреть две работы, в которых исследовали экспрессии генов пациентов, больных множественной миеломой:

1. GSE19784 (282 пациента)
2. GSE24080 (559 пациентов)

Более подробную информацию об исследованиях можно найти на сайте <https://www.ncbi.nlm.nih.gov/> (The National Center for Biotechnology Information). Также прилагаются аннотации по всем пациентам – пол, возраст, стадия заболевания. Целевая переменная в данном случае – HR\_FLAG – относится ли пациент к группе риска или нет.

Задание состоит в том, чтобы обучить предиктор, предсказывающий, относится ли пациент к группе риска, и повалидировать этот предиктор. В качестве обучающей выборки используйте датасет GSE24080, в качестве валидационной - GSE19784. При этом в процессе Вы наверняка столкнетесь с такими проблемами, как:

* Как отобрать наиболее важные фичи (гены)
* Как чистить данные, убирать шумы
* Как сравнивать два разных датасета. Несмотря на то, что в обоих работах данные получены в идентичных уловиях, они все же могут отличаться.
* Как использовать клинические данные из аннотаций

Предложенное решение ожидается в виде кода на Python, по возможности с комментариями. Предпочтительнее использовать jupyter notebook.

Индексы генов в датасетах - это Entrez ID. Их можно переводить в Hugo Symbol либо вручную - через базу данных NCBI, либо автоматически с помощь библиотеки mygene (http://mygene.info/)