НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
“КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ”
ФІЗИКО-ТЕХНІЧНИЙ ІНСТИТУТ
КАФЕДРА ПРИКЛАДНОЇ ФІЗИКИ

Звіт

з науково-дослідницької практики

Виконав:
студент 6-го курсу групи ФФ-41м
Редчук Богдан
Керівник практики від вузу:
Монастирський Г.Є.
Керівник практики від підприємства:
Строй Д.О.

Київ – 2016

Науково-дослідну практику було пройдено в Інституті фізіології ім. О.О. Богомольця НАН України в лабораторії генетики, яка відноситься до загальної та молекулярної патофізіології. Головними предметами досліджень даної лабораторії є дослідження кореляцій генетичних та негенетичних факторів із ризиками розвитку патологій організму.

Керівником практики було поставлене завдання на науково-дослідну практику, яке полягало в опануванні необхідними навичками та методиками аналізу надвеликих масивів даних та первинним аналізом результатів генетичного скрінінгу, необхідних для підготовки магістерської дисертації за темою «Використання методів машинного навчання для визначення факторів ризику мультифакторних патологій».

Теоретична довідка:

З моменту суттєвого здешевшення технологій генетичного тестування (скрінінгу) питання збору великих масивів даних для подальшого аналізу остаточно перейшло до розряду таких, що обмежені не технологією збору даних, а лише об’ємами даних, що потрібно зібрати.

Це призвело до накопичення надвеликих масивів даних- так званих, «Big Data». **Big Data** — це набори даних такого об'єму, що традиційні інструменти не здатні здійснювати їх охоплення, управління та обробку за помірний час. Важливо також відзначити те, що під терміном Big Data у різних контекстах можуть мати на увазі дані великого об'єму, технології їх обробки, проекти, компанії, які активно використовують дану технологію.

Основною (загальною) методикою роботи з «надвеликими даними» є технології машинного навчання. Машинне навчання — розділ штучного інтелекту, має за основу побудову та дослідження систем, які можуть самостійно навчатись з даних.

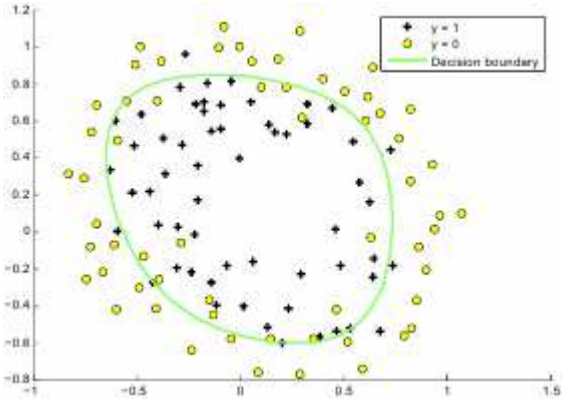
Методи досліджень даних:

Дерево прийняття рішень (класифікаційні дерева). Структура дерева містить такі елементи: «листя» і «гілки». На ребрах («гілках») дерева прийняття рішення записані атрибути, від яких залежить цільова функція, в «листі» записані значення цільової функції, а в інших вузлах — атрибути, за якими розрізняються випадки. Щоб класифікувати новий випадок, треба спуститися по дереву до листа і видати відповідне значення. Кожен лист являє собою значення цільової змінної, зміненої в ході руху від кореня по аркушу. Кожен внутрішній вузол відповідає одній з вхідних змінних. Дерево може бути також «вивчено» поділом вихідних наборів змінних на підмножини, що засновані на тестуванні значень атрибутів. Це процес, який повторюється на кожному з отриманих підмножин. Рекурсія завершується тоді, коли підмножина в вузлі має ті ж значення цільової змінної, таким чином, воно не додає цінності для пророкувань. Процес, що йде «згори донизу», індукція дерев рішень (TDIDT), є прикладом поглинаючого «жадібного» алгоритму, і на сьогоднішній день є найбільш поширеною стратегією дерев рішень для даних, але це не єдина можлива стратегія.

Його безпосереднім розвитком є використання ***методу випадкового лісу («Random forest»)*** — алгоритму машинного навчання,що полягає у використанні комітету (ансамблю) вирішальних дерев. Алгоритм поєднує в собі дві основні ідеї: метод беггінга Брейман і метод випадкових підпросторів, запропонований Tin Kam Ho. Алгоритм застосовується для задач класифікації, регресії і кластеризації.

Третім методом, що використовується у нашому аналізі масиву даних при пошуку закономірностей (і, як виявилося, мало прийнятний сам по собі, у якості безпосереднього інструменту аналізу первинних даних; див. нижче) є **Логістична регресія** або **логіт регресія** англ. *logit model* — статистичний регресійний метод, що використовується у випадку коли пояснювана змінна може набувати тільки двох значень (чи, більш загально, скінченну множину значень).

Результати досліджень масиву даних:



На даному рисунку графічно відображена дія логістичної регресії і отримана в результаті гіпотеза, яка реалізована у вигляді обмежувальної лінії.

Двома параметрами були обрані рівень цукру у крові та спадкова схильність до серцево-судинних захворювань. «Хрестиками» позначені особини, які не мають серцево-судинних патологій, а «кружечками» -- страждають ними.

Для прикладу обрано лише два параметри, щоб можна було наочно зобразити результат даного алгоритму машинного навчання. У роботі буде ж проведене дослідження даних з використанням до 40-50 різних характеристик.

В подальшому заплановано розгляд методу опорних векторів, комбінування всіх вище згаданих методів, для емпіричного визначення оптимального алгоритму вирішення проблеми поставленої за ціль магістерської дисертації.