МЕТОДИ DATA MINING ДЛЯ ПОСТАНОВКИ ДІАГНОЗУ, ОСНОВАНОГО НА СКАРГАХ ПАЦІЄНТА

Філіпенко Я. КА-44

Науковий керівник: проф. Бідюк П.І.

Мета роботи

- Метою роботи є вивчення та практичне застосування методів аналізу тексту та великих масивів даних для постановки діагнозу, основаного на скаргах пацієнтів, розробка програмного забезпечення щоб поліпшити працю лікарів та створити додаткові послуги для клієнтів.

Актуальність дослідження

- На даний момент приватні поліклініки кожному пацієнтові виділяють 30 хвилин на прийом. Тобто в день один лікар може прийняти 16-20 людей. Це число можна збільшити, зменшивши час прийому.

- Також це можливість розробити додаткові сервіси для клієнтів. Тобто звільнити менеджерів від частини роботи.

Об’єкт дослідження

- Об’єктом дослідження є велика вибірка даних, яка складається зі скарг пацієнтів, поставлених лікарями, діагнозів та їх відповідних кодів у МКХ-10.

- Вибірка включає в себе близько 70000 записів, які треба привести до одного вигляду та проаналізувати.

Предмет дослідження

- Нейронна мережа Keras
- Класифікатор Random Forest
- Наївний байєсів класифікатор

Постановка задачі

- Після отримання скарг від пацієнта в текстовому вигляді провести їх аналіз та отримати в результаті первинний діагноз.

- Розробити програмне забезпечення для поліклінік, яке слугувало б як система підтримки прийняття рішень для постановки первинного діагнозу. Дані беруться відповідно до кодів J00-J99 з МКХ-10.

Дані

- Дані подані у таблиці Excel, яка має 4 стовпчики:

- Номер запису
- Скарги
- Код хвороби по МКХ-10
- Діагноз

Використані методи дослідження

- Під час досліджень було використано 3 методи:

- Нейронна мережа
- Класифікатор Random Forest
- Наївний баєсів класифікатор

Data-mining

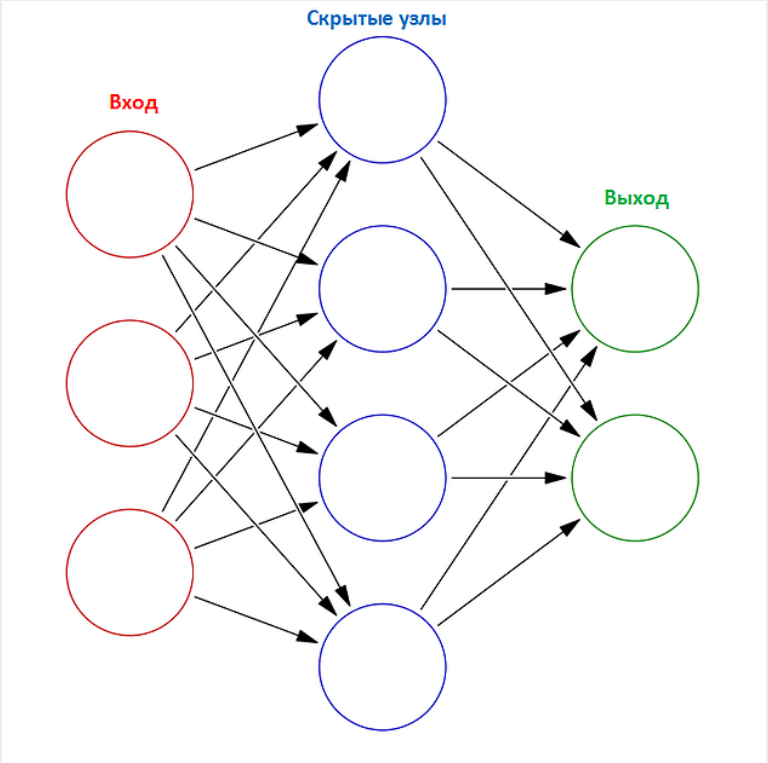
- Це процес виявлення в сирих даних раніше невідомих, нетривіальних, практично корисних і доступних інтерпретацій знань, необхідних для прийняття рішень у різних сферах людської діяльності.

- Це процес сортування великого набору даних для розпізнавання моделей та встановлення відношення між змінними для розв’язання задач за допомогою вибраних методів аналізу даних

Нейронна мережа

- Штучний інтелект і нейронні мережі зараз в медіа-тренді, проте в них немає якогось особливого чарівництва більшого, ніж диво математики. Це математична модель втілена в програмному або апаратному забезпеченні, що виконує обробку даних.

- Найпоширенішими застосуваннями нейронних мереж є:
- Класифікація
- Передбачення
- Розпізнавання



Нейромережева бібліотека Keras

- Keras є високорівневим нейромережевим API, написаним на R і може працювати поверх TensorFlow, CNTK або Theano. Розробка спрямована на можливість швидкого експериментування з використанням даних.

Нейромережева бібліотека Keras

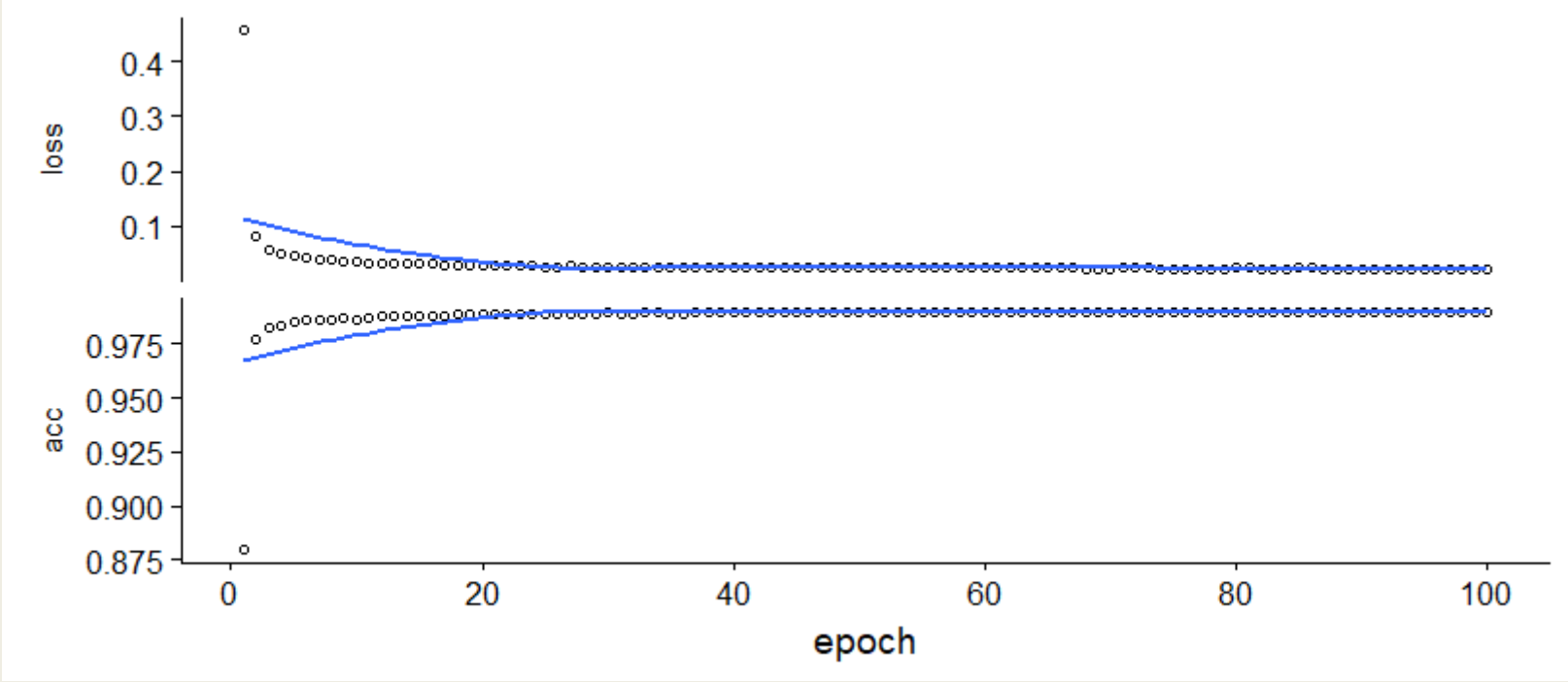
- Керівні принципи:

- Зручність у використанні. Keras - API, призначений для людей, а не машин. Це робить користувальницький досвід фронтом та центром. Keras пропонує послідовні та прості API, мінімізує кількість дій користувача, необхідних для звичайних випадків використання, і забезпечує чіткий та дієвий відгук щодо помилок користувача.
- Модель розуміється як послідовність або автономних, повністю налаштовуваних модулів, які можуть бути підключені разом із якомога меншими обмеженнями.
- Нові модулі легко можна додавати (як нові класи та функції), а існуючі модулі наводять достатньо прикладів.
- Робота з R та Python.

Нейромережева бібліотека Keras

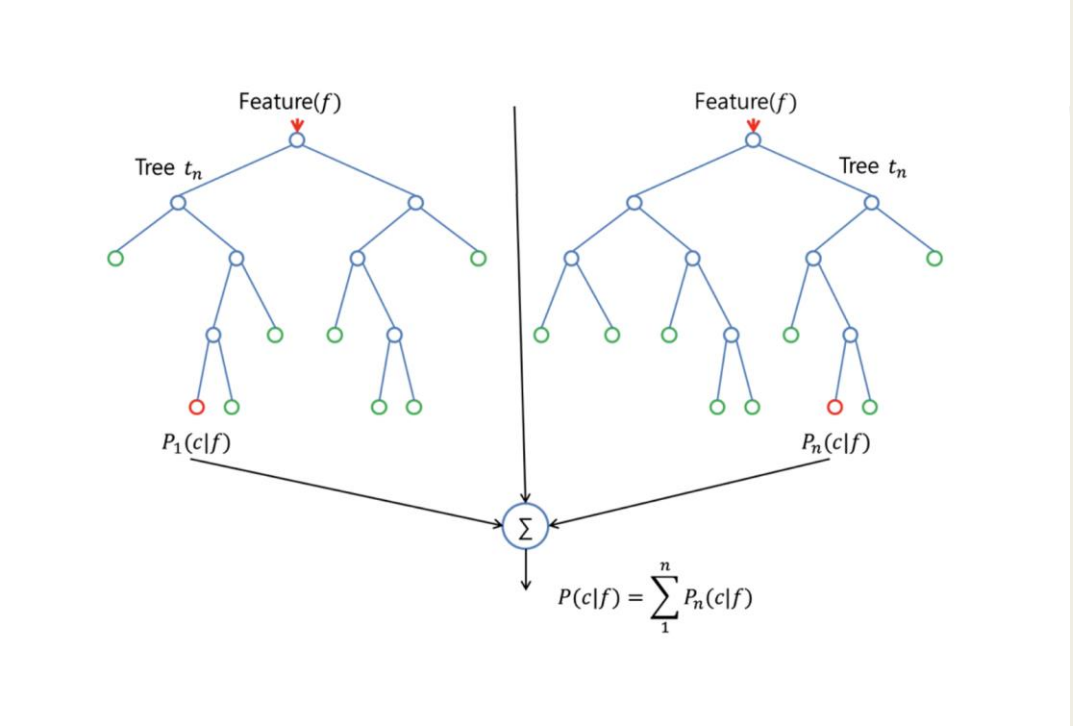
- Час навчання нейронної мережі – 202 секунди на вибірці біля 70000 записів

Графік навчання нейронної мережі



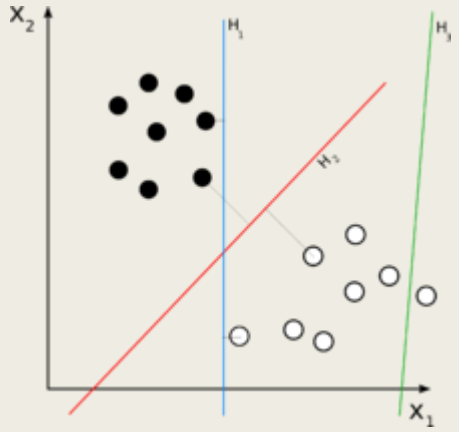
Random Forest

- Random Forest – гнучкий, простий у використанні алгоритм машинного навчання.
- Одна з великих переваг випадкового лісу полягає в тому, що така модель мо бути використана як для класифікації, так і для регресійного аналізу, які склад більшість існуючих систем машинного навчання.



Наївний байєсів класифікатор

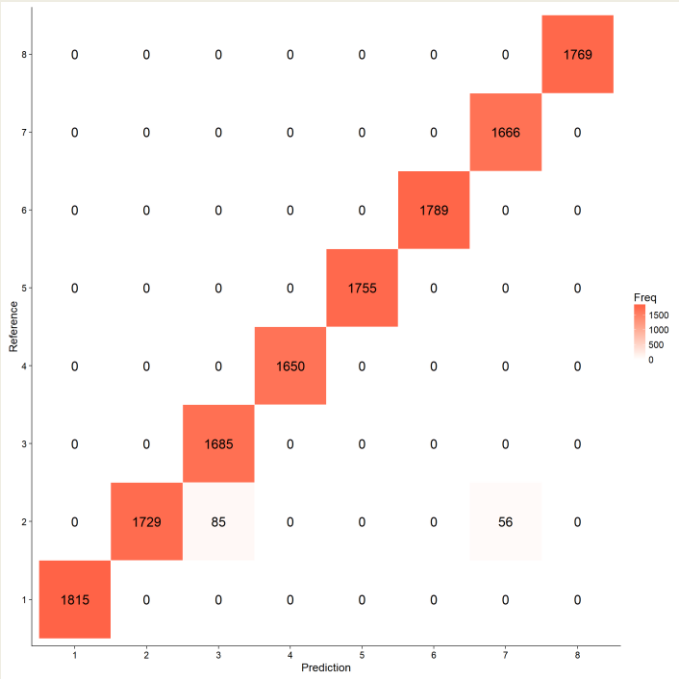
- Статистично-ймовірнісний метод класифікації, який використовується в машинному навчанні та інтелектуальному аналізі.
- Перевагою цього підходу є те, що вимоги до розміру вибірки скорочуються від експоненційних до лінійних. Недолік — те, що модель є точною лише у випадку, коли виконується припущення про незалежність факторів впливу.
- В іншому випадку, строго кажучи, обчислені ймовірності вже не є точними



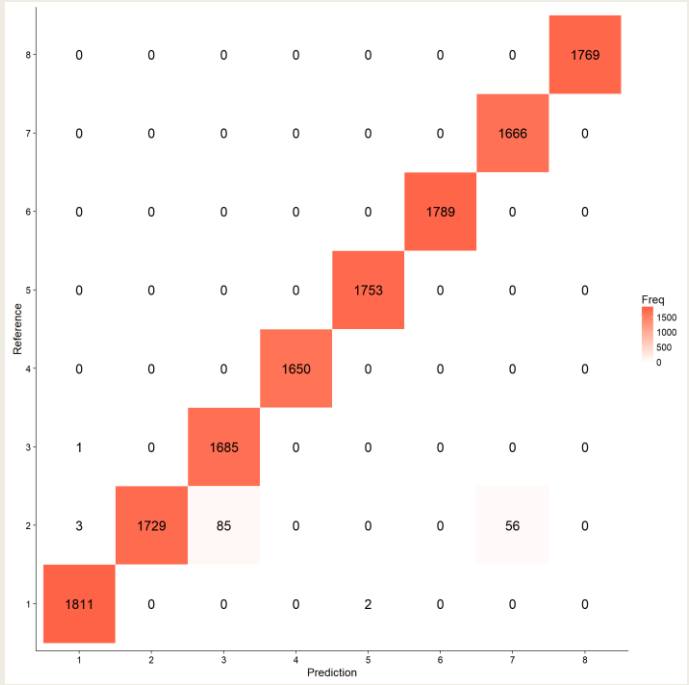
$$c = \arg\max_{c \epsilon C} P(c) \prod P(o_i|c)$$

Результати обчислювальних експериментів

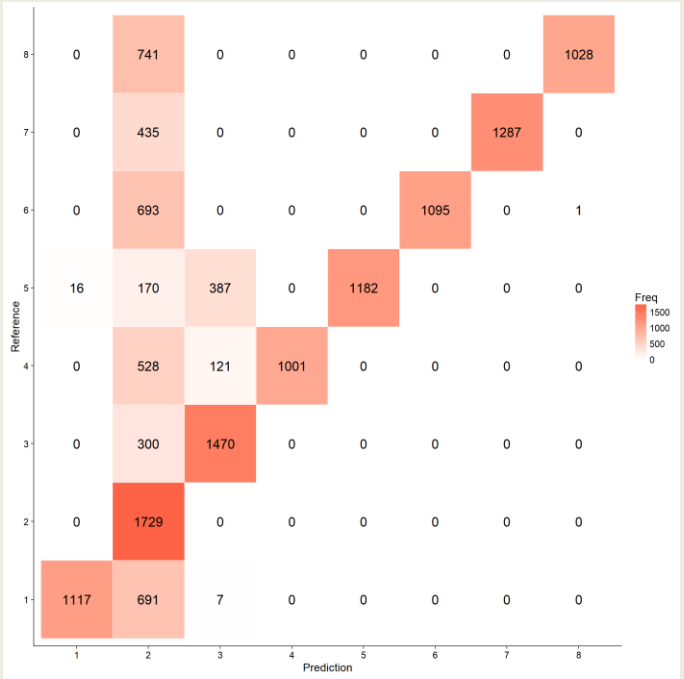
Результати показують, наскільки точно працює кожен із методів



Точність: 0.9899279
Нейронна мережа Keras

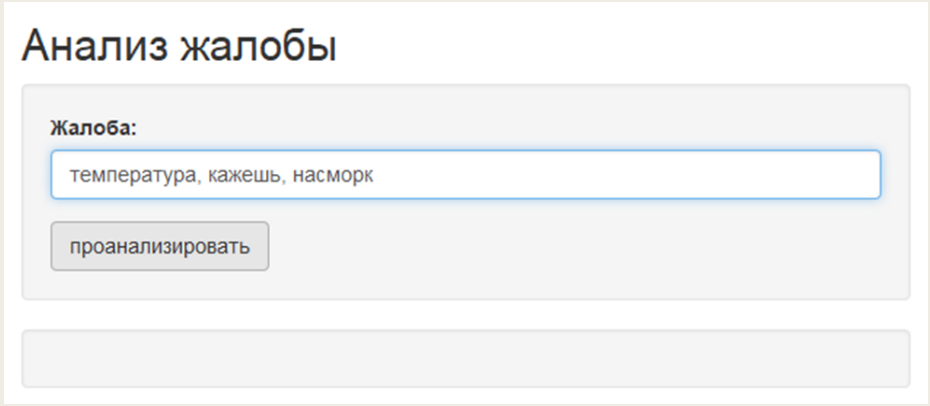


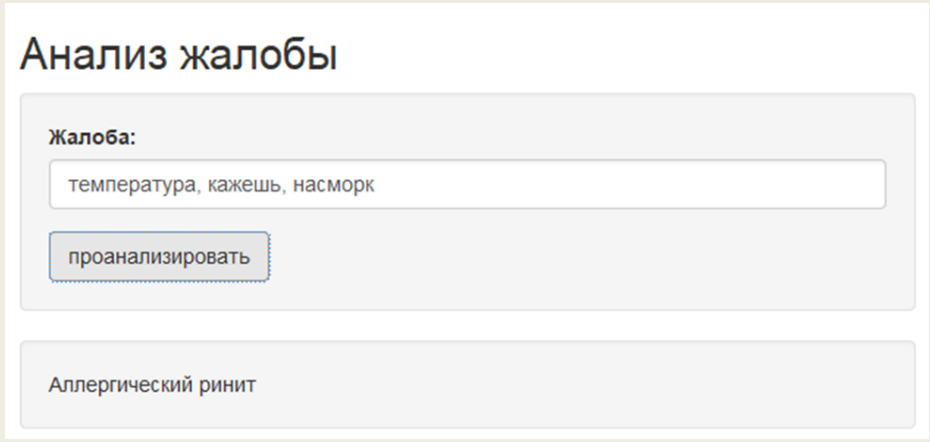
Точність: 0.9894992
Random Forest



Точність: 0.7078363
Наївний байєсів класифікатор

Інтерфейс програмного продукту





Висновок

- За результатами досліджень виявилося, що для аналізу текстових полів краще всього працює метод на основі нейронних мереж. Він дає більшу ймовірність отримати правильне рішення, ніж інші методи.
- Також розроблено систему підтримки прийняття рішення на мові R, за допомогою якої можна зменшити час прийому лікаря з досягненням високого результату стосовно попереднього діагностування.
- Також систему можна розвивати збільшенням кількості хвороб у базі, голосовим введенням тексту, введенням правил для формування діагнозу та ін.