ХЕРСОНСЬКИЙ НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ
ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ ТА ДИЗАЙНУ
КАФЕДРА ІНФОРМАТИКИ ТА КОМП'ЮТЕРНИХ НАУК

Пояснювальна записка

до дипломної бакалаврської роботи

на тему:

«Застосування комбінованого підходу для розв'язання задачі розрахунку прогнозних значень пептид-протеїнового зв'язування на основі байєсівської лінійної регресії»

Виконав: студентка 4 курсу, групи 4КН

спеціальності 122 «Комп'ютерні науки»

Макарова В.О.

Керівник: д.т.н. проф. Литвиненко В.І.

Рецензент: д.т.н. проф. Рудакова Г.В.

Херсон – 2021 р.

Факультет	**Інформаційних технологій та дизайну**
Кафедра	**Інформатики та комп'ютерних наук**
Рівень вищої освіти	**перший (бакалаврський) рівень**
Галузь підготовки	**12 «Інформаційні технології»**
	(шифр і назва)
Освітньо-професійна програма	**Комп'ютерні науки**
	(назва)
Спеціальність	**122 «Комп'ютерні науки»**
	(шифр і назва)

ЗАТВЕРДЖУЮ
Завідувач кафедри ІКН,
професор
_____В.І. Литвиненко
«____» _____ 2021 року

З А В Д А Н Н Я

НА ДИПЛОМНУ РОБОТУ СТУДЕНТА

Макарова Владислава Олексіївна

(прізвище, ім'я, по батькові)

1. Тема роботи: Застосування комбінованого підходу для розв'язання задачі розрахунку прогнозних значень пептид-протеїнового зв'язування на основі байєсівської лінійної регресії.

керівник роботи Литвиненко Володимир Іванович, доктор технічних наук, професор, завідуючий кафедри інформатики і комп'ютерних наук

(прізвище, ім'я, по батькові, науковий ступінь, вчене звання)

затверджені наказом ХНТУ від 26.11.2020р. № 644-с

2. Строк подання студентом роботи

04.06.2021

3. Вихідні дані до роботи

4. Зміст розрахунково-пояснювальної записки (перелік питань, які потрібно розробити): Вступ. Розділ 1. Аналітичний огляд літературних та інших джерел. Розділ 2. Теоретична частина. Аналіз предметної області. Розділ 3. Практична частина. Розділ 4. Охорона праці.

5. Перелік графічного матеріалу (з точним зазначенням обов'язкових креслень)

Таблиць – 22,

Формул – 9,

Рисунків – 34.

6. Консультанти розділів роботи

Розділ	Прізвище, ініціали та посада консультанта	Підпис, дата	
		завдання видав	завдання прийняв
4	к.с.-г.н., доцент Малєєв В.А.		

7. Дата видачі завдання 08.02.2021

КАЛЕНДАРНИЙ ПЛАН

№ з/п	Назва етапів дипломної роботи	Строк виконання етапів роботи	Примітка
1	Збір інформації. Аналітичний огляд літературних та інших джерел.	08.02.2021-01.03.2021	
2	Огляд середовища розробки RStudio та мови R.	02.03.2021-16.03.2021	
3	Огляд методів прогнозування.	17.03.2021-31.03.2021	
4	Вибір методів для вирішення задачі.	1.04.2021-30.04.2021	
5	Розробка алгоритму для розрахунку прогнозних значень.	01.05.2021-28.05.2021	
6	Охорона праці.	29.05.2021-01.06.2021	
7	Оформлення роботи.	02.06.2021-04.06.2021	

Студент _____ В.О. Макарова

(підпис) (прізвище та ініціали)

Керівник роботи _____ В.І. Литвиненко

(підпис) (прізвище та ініціали)

ЗМІСТ

РЕФЕРАТ

Дана кваліфікаційна бакалаврська робота присвячена застосуванню комбінованого підходу для розрахунку прогнозних значень пептид-протеїнового зв’язування.

Пояснювальна записка дипломного проекту складається з 4 розділів та 4 додатків. У тексті присутні: 9 таблиць, 34 формули, 22 рисунка. При написанні дипломного проекту було використано 23 джерела.

Проаналізовано методи та підходи, які застосовуються в сучасному прогнозуванні даних, розроблено підхід до вирішення задачі розрахунку прогнозних значень пептид-протеїнового значення. Для розробки була використана мова програмування R та сторонні пакети.

Дана робота присвячена комбінованому підходу з використанням байесовської регресії для розв'язання задачі прогнозування в сфері біоінформатики.

Ключові слова: аналіз даних, регресія, прогнозування, пептид-протеїнове зв'язування.

ABSTRACT

This qualifying bachelor's thesis is devoted to the application of a combined approach to calculate the predicted values of peptide-protein binding. The explanatory note of the diploma project consists of 4 sections and 4 appendices. The text contains: 9 tables, 34 formulas, 22 figures. 23 sources were used in writing the thesis project. The methods and approaches used in modern data forecasting are analyzed, the approach to solving the problem of calculating the predicted values of peptide-protein value is developed. The programming language R and R-packages were used for development. This paper is devoted to a combined approach using Bayesian regression to solve the problem of forecasting in the field of bioinformatics.

Key words: recognition, regression, prediction, peptide-protein.

ПЕРЕЛІК УМОВНИХ СКОРОЧЕНЬ

Скорочення, термін, позначення	**Пояснення**
ОС	Операційна система.
ПЗ	Програмне забезпечення.
МНК	Метод найменших квадратів
MSE	Mean Squared Error
MAE	Mean Absolute Error
IDE	Integrated Development Environment; Інтегроване середовище розробки - система, яка створена для розробки програмного забезпечення.

ВСТУП

На сьогоднішній день дуже стрімко розвиваються комп'ютерні технології. Вони застосовуються в багатьох галузях науки, таких як медицина, економіка, астрономія та інших.

Оскільки тепер збір та накопичування даних відбувається швидко, з'явилися такі проблеми як обробка та аналіз даних. Чим більше даних – тим більше часу потрібно для їх обробки.

Зараз стоїть питання про пошук та розробку алгоритмів, методів, які дозволять швидко та без втрат обробляти дані, маніпулювати ними.

Необхідною умовою сучасного аналізу даних є використання технологій, а точніше – комп'ютерних програм. Від їх функціональної повноти та алгоритмічної продуманості залежить кінцева інтерпретація результатів дослідження та надійність зроблених висновків.

Прогнозування є важливою концепцією успіху. Це передбачення майбутнього на підставі досвіду та припущень відносно нього. Використання методів прогнозування стало поширеним в багатьох сферах. Їх застосовують в економетриці, медицині, маркетингу, в аналізі соціальних, інформаційних мереж.

Актуальність теми. З розвитком технологій актуальним залишається завдання розробки ефективних алгоритмів та підходів до вирішення різноманітних задач, які стосуються обробки даних. Одна з найважливіших тем на сьогоднішній день є прогнозування значень.

Раніше більшість методів не використовували через потреби витрат великої кількості ресурсів та часу. Але тепер, з розвитком технологій, такі методи стали більш доступними для використання.

Причиною стрімкого розвитку машинного навчання в останній час полягає в великій кількості даних, які збираються та доступними. Оскільки набори даних все більше, потрібно тримати під контролем кількість

характеристик, які враховуються обираючи «найкращі», на яких моделі будуть навчатися.

Машинне навчання та глибинне навчання почали широко використовувати в медицині. Системи підтримки прийняття рішень, які допомагають експертам в прийнятті рішень, були розроблені з використанням алгоритмів машинного кавчання. З використанням таких систем лікарі діагностують різні хвороби та визначають, які потрібні ліки пацієнту [2]. Наявність історії хвороби відіграє важливу роль в діагностиці та лікуванні. Системи прийняття рішень обробляють історію хвороби та попереднє лікування, а це в свою чергу може бути використано як довідник та для прогнозування лікування таких випадків [3].

Перший геном людини був секвенсован більше десяти років тому та став доступним для наукових досліджень. Це стало великим відкриттям і завершена послідовність містила більше трьох мільярдів пар основ. В проекті використовувалися не лише передові методи молекулярної біології, але й обчислювальні методи, від яких залежав успіх, особливо на завершальній фазі. Отже, одним з наслідків цього проекту є те, що біологічні дослідження за допомогою комп'ютера є та будуть необхідними.

Завершення роботи над послідовністю геному людини означало початок нової ери наукових досліджень, яку називають пост-геномною ерою. Досягнення в області геномних досліджень дали величезну кількість даних. Для відкриття біологічних знань та отримання клінічної інформації з цих даних потрібен був інтенсивний аналіз. Розвиток складних біологічних проблем та геномних технологій з великою кількістю даних вимагає об'єднання обчислювальних методів та наук про життя. Перспективні рішення та підходи були запропоновані алгоритмами, які призначені для вирішення конкретних проблем в біологічних системах. Проте, дані отримані за допомогою цих технологій змушують розробляти нові стратегії для кращого аналізу та моделювання інформації, а також інтеграції їх з біологічними системами.

Потреба в більш ефективному аналізі та отриманні цінної інформації з наборів біологічних даних привела до розвитку області біоінформатики. Це міждисциплінарна область досліджень, яка направлена на покращення аналізу наборів біологічних даних за допомогою обчислювальних методів та відповідних програмних інструментів для вирішення складних біологічних проблем.

В пост-геномну епоху набори даних біоінформатики часто є складними завданнями. Вони є не тільки великими, але й часто вимірювані дані є неповними та містять невизначеність. Тому обчислювальні методи, що розробляються, спрямовані на зменшення шуму, розмірності, а також на вирішення проблеми неповноти та невизначеності в таких наборах даних.

Прогнозування афінності (спорідненості) зв'язування є однією з прикладних областей біоінформатики, де дані часто бувають складними, невизначеними та багатовимірними. Людське мислення має змогу в основному обробляти невеликі набори даних порівняно з комп'ютерами, які здатні обробляти великі об'єми даних. Комп'ютерне прогнозування спорідненості зв'язування має велике значення для ефективного аналізу наборів біологічних даних. Прогнозування спорідненості зв'язування пептидів вважається однією найскладніших проблем моделювання в біоінформатиці.

Мета і завдання дослідження. Дипломна робота присвячена дослідженню існуючих методів та підходів до питання прогнозування, створення програмного забезпечення для розрахунку прогнозних значень з використанням байєсіської регресії за допомогою мови програмування R та її сторонніх пакетів.

Об'єктом дослідження є процес обирання та розробки моделей для обробки даних, а також розрахунку прогнозних значень.

Предметом дослідження є підходи та методи до питання прогнозування в умовах сучасного розвитку технологій.

Наукова новизна отриманих результатів полягає в отриманні програмного забезпечення, яке ефективно прогнозує значення, використовуючи комбінований підхід до задачі.