3. Data Structures & Algorithms in Swift: Implementing practical data structures with Swift 4, 2018. – 600 с.
4. Swift Apprentic: “Beginning Programming with Swift 4” 3rd Edition, 2017. – 600 с.
5. Core Data by Tutorials: “iOS 10 and Swift” 3rd Edition, 2016. – 600 с.

УДК 004.4:314.117.11:(94) (043.2)

Ємельянова А.О., студентка 4 курсу спеціальності «Комп’ютерні науки»
Нескородєва Т. В., канд. техн. наук, доцент, завідувач кафедри комп’ютерних наук та інформаційних технологій

АНАЛІЗ НАРОДЖУВАНОСТІ В АВСТРАЛІЇ ЗАСОБАМИ МОВИ R

Донецький національний університет імені Василя Стуса, м. Вінниця

Народжуваність є одним із основних чинників зростання населення (або падіння). Аналіз народжуваності у сполученні із смертністю та міграціями дозволяє оцінювати, як популяція буде зростати, спадати або стабілізуватися із плином часу. Даний аналіз також може передбачати й інші демографічні зміщення, як от майбутні вікові розподіли серед населення.

Актуальність полягає в тому, що коефіцієнти народжуваності впливають на коефіцієнт заміщення, який, в свою чергу, визначає вік населення певної країни і від розміру показника якого залежить безліч інших не менш важливих аспектів.

Розглянемо приклад такого аналізу на основі даних, взятих із сайту Австралійського бюро статистики [3], який містить безліч статистичних даних, зокрема датасет, який включає у собі дані по народжуваності серед жінок різних вікових категорій у період з 1935 по 2019 роки. Даний набір даних містить 85 спостережень із 9 змінними, а саме:

- Year – період часу, за який проводилась статистика;
- 7 вікових категорій жінок з 15 до 49 років із кроком 4;
- Avg – власноруч створена змінна для обрахунку середнього коефіцієнту народжуваності по усім віковим категоріям за кожен рік.

У наборі зібрана та наведена статистика у вигляді вікових коефіцієнтів народжуваності, що і є об’єктами спостереження. Віковий коефіцієнт народжуваності визначається річною кількістю новонароджених у жінок певного віку або певної вікової групи поділеної на 1000 жінок цієї вікової групи.

На першому етапі створюється описова статистика показників набору «Age-specific fertility rates 1935-2019» і будується матриця розсіювань змінних (рис. 1).

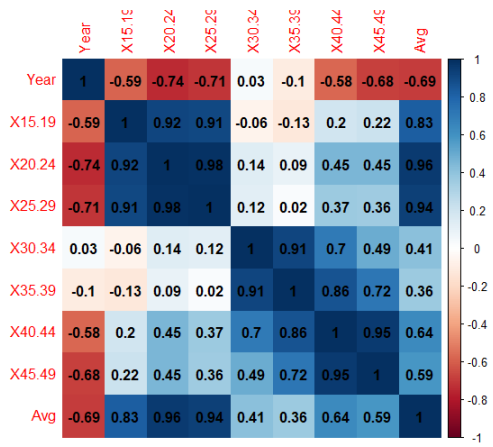

Рисунок 1 – Побудова матриць діаграм розсіювання

Із побудованої матриці діаграм розсіювання можна зробити висновок, що у більшості випадків найбільша кореляція відбувається між тими віковими категоріями, які межують між собою, тобто є сусідами.

Далі важливим і необхідним кроком є створення візуалізації даних для того, щоб краще розуміти дані набору та створювати подальші припущення на основі цих даних (рис. 2).

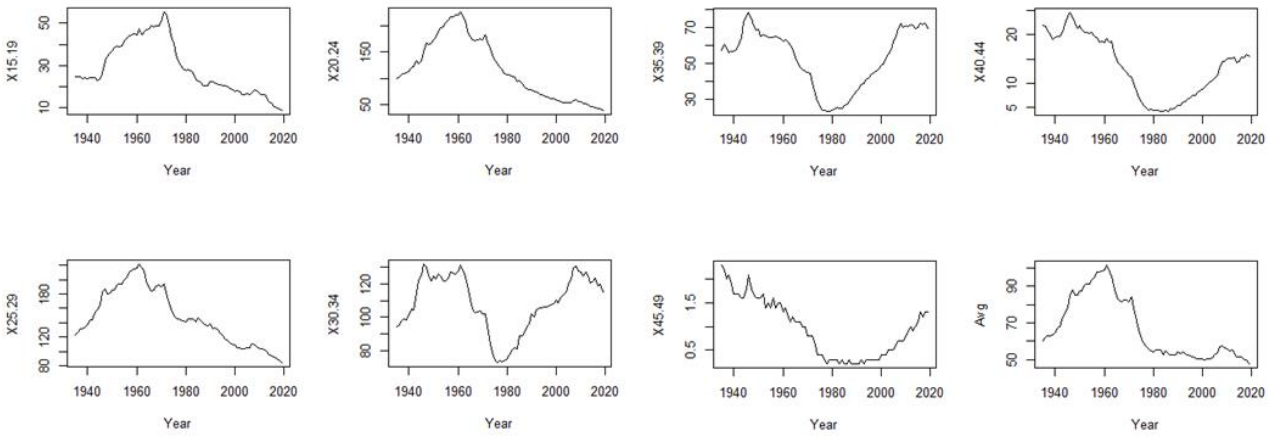

Рисунок 2 – Побудова графіків для аналізу руху народжуваності

Згідно аналізу побудованих графіків можна зробити висновок, що коефіцієнт народжуваності у кожній віковій категорії жінок останніми роками дещо знизився. Також 1980-роки негативно вплинули на народжуваність жінок віком від 30 до 49 років. А ще оцінюючи графік по Avg, виходить, що середній коефіцієнт народжуваності по усім віковим категоріям почав знижуватись починаючи із 2005 року.

Далі для прогнозування руху народжуваності потрібно створити якісну змінну Direction, яка матиме лише два значення: позитивне «1» або негативне «0», що означають збільшення або зменшення коефіцієнту відповідно. Дана змінна створюється на основі середнього значення Avg за кожен рік. Тобто якщо Avg буде більшим за середнє значення своєї категорії, то матиме значення «1», інакше – значення «0».

Наступним етапом є створення GLM, LDA, QDA і KNN моделей [4] із Direction у якості відгуку та віковою категорією 20-24, яка має найбільші коефіцієнти народжуваності порівняно з іншими, у якості предиктора (рис. 3).

```
> table(glm.pred, Direction.test)
        Direction.test
glm.pred  0  1
       0 15  0
       1 11 17
> mean(glm.pred != Direction.test)
[1] 0.255814
> library(MASS)
> fit.lda = lda(Direction ~ X20.24, data = age, subset = train)
> lda.pred = predict(fit.lda, age.test)
> table(lda.pred$class, Direction.test)
    Direction.test
      0  1
  0 25  3
  1  1 14
> mean(lda.pred$class != Direction.test)
[1] 0.09302326
```

```
> qda.fit = qda(Direction ~ X20.24, data = age, subset = train)
> qda.pred = predict(qda.fit, age.test)
> table(qda.pred$class, Direction.test)
    Direction.test
      0  1
  0 25  3
  1  1 14
> mean(qda.pred$class != Direction.test)
[1] 0.09302326
> library(class)
> train.X = as.matrix(X20.24[train])
> test.X = as.matrix(X20.24[!train])
> train.Direction = Direction[train]
> set.seed(1)
> knn.pred = knn(train.X, test.X, train.Direction, k = 1)
> table(knn.pred, Direction.test)
        Direction.test
knn.pred  0  1
       0 25  2
       1  1 15
> mean(knn.pred != Direction.test)
[1] 0.06976744
```

Рисунок 3 – Звіти по GLM, LDA, QDA і KNN моделях

Зрештою, кожна модель має наступний відсоток помилки точності прогнозування:

- GLM – 25,6%;
- LDA – 9,3%;
- QDA – 9,3%;
- KNN (при кількості найближчих сусідів K = 1) – 7%;
- KNN (при кількості найближчих сусідів K = 20) – 7%;
- KNN (при кількості найближчих сусідів K = 40) – 39,5%.

Здійснений аналіз дає змогу дійти висновку, що найточніше прогнозування руху коефіцієнту народжуваності дає KNN модель при невеликій кількості сусідів, оскільки чим ближча кількість сусідів до довжини вибірки, тим більшою буде помилка.

Отже, аналіз даних відіграє важливу роль у будь-якій сфері життя, оскільки дозволяє шукати, виявляти та передбачати закономірності у різних об’ємах даних. Датасет, що розглядається у даній роботі, дає можливість проаналізувати і передбачити, у якому напрямку рухатиметься крива коефіцієнту народжуваності в майбутньому, а також представити результати досліджень у вигляді різноманітних візуальних елементів за допомогою засобів мови програмування R.

Список літературних джерел

1. Грущенко В.Ю., Нескородєва Т.В. Аналіз даних ринку персональних комп’ютерів. Матеріали всеукраїнської науково-практичної конференції "Комп’ютерні технології обробки даних" (4 грудня 2020 року) – Вінниця: ДонНУ імені Василя Стуса. С.14-16.
2. Білич А.О., Нескородєва Т.В. Аналіз даних перегляду фільмів користувачами Платформи IMDB. Матеріали всеукраїнської науково-практичної конференції "Комп’ютерні технології обробки даних" (4 грудня 2020 року) – Вінниця: ДонНУ імені Василя Стуса. С.135-138.

3. Births, Australia. Australian Bureau of Statistics. 2020. [Електронний ресурс]. Режим доступу до ресурсу: https://www.abs.gov.au/statistics/people/population/births-australia/latest-release.
4. Джеймс Г., Уиттон Д., Хасти Т., Тибширани Р. Введение в статистическое обучение с примерами на языке R. Пер. с англ. С. Э. Мастицкого - М.: ДМК Пресс, 2017. – 456 с.

УДК 004.021

Засько Б. В., студент 2 курсу СО Магістр
Штовба С. Д., професор кафедри інформаційних технологій

ПІДБІР РЕЦЕНЗЕНТІВ НА ОСНОВІ АНАЛІЗУ КЛЮЧОВИХ СЛІВ

Донецький національний університет імені Василя Стуса, м. Вінниця

Важливим етапом процесу рецензування наукових робіт є знаходження людини, яка фахово може оцінити. Для підбору рецензента зазвичай застосовується «ручний» підхід, коли кандидата на роль рецензента шукають серед пулу науковців, або ж звертаються до певної наукової установи, а в самій уже установі все так само «ручним» способом шукають підходящу людину поміж своїх співробітників. Такий спосіб є простим, але потребує значних витрат часу, має великі ризику упередженого підбору кандидатур та витоку конфіденційної інформації.

Альтернативний підхід до підбору рецензента – автоматизований, який і розглядається в даній роботі. За цього підходу рецензована робота завантажується до деякої інформаційної системи, і в результаті аналізу відповідних метаданих, система генерує рейтинговий список кандидатів у експерти. І хоча автоматизований підбір рецензентів виглядає привабливішим, але реалізувати його непросто – поточні наукові бази даних не пристосовані для підбору рецензентів.

Нами на рис. 1 пропонується загальний алгоритм підбору рецензентів для роботи. Першим кроком користувач надає ключові слова наукової роботи, для якої відбувається пошук рецензента. Далі відбувається аналіз введених ключових слів. Якщо аналіз пройшов успішно, останнім кроком відбувається розрахунок подібності роботи з кожним з кандидатів із пулу рецензентів, і користувачу повертається упорядкований перелік потенційних рецензентів.