**КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ ІМЕНІ ТАРАСА ШЕВЧЕНКА**

**Факультет інформаційних технологій**

Кафедра технологій управління

**ЗВІТ**

**ПРО ПРОХОДЖЕННЯ НАУКОВО-ДОСЛІДНОЇ ПРАКТИКИ**

Студентки 2 курсу магістратури групи ІАВ21

спеціальності Інформаційна аналітика та впливи

Сокол Олена Олександівна

(прізвище, ім’я, по батькові)

Термін практики з 21.01 по 4.03.2019р.

База практики АТ «Альфа-Банк»

Науковий керівник магістерської роботи

Професор кафедри технологій управління, доктор технічних наук

(науковий ступінь, вчене звання керівника)

Білощицький Андрій Олександрович

(прізвище, ім’я, по батькові)

Керівник практики від кафедри

Асистент кафедри технологій управління, кандидат економічних наук

(науковий ступінь, вчене звання керівника)

# Мезенцева Ольга Олексіївна

(прізвище, ім'я, по батькові)

**Київ-2019р.**

1. Окрім питань, передбачених програмою виробничої практики, розглянуто таке завдання:

Вдосконалити методи аналізу даних для прогнозування задля покращення кредитної ситуації в Україні.

Здобуті техніки аналізу даних під час проходження науково-дослідницької практики в компанії АТ «Альфа-Банк» дасть змогу використати такі завдання:

а) Провести аналіз основних технік «Data pre-processing».

б) Реалізувати алгоритм кластеризації FOREL.

2. Рекомендована література та інші матеріали.

1. Patricia B. Cerrito «Introduction to Data Mining»
2. Daniel Fasulo «An Analysis Of Recent Work on Clustering Algorithms»
3. Н. Паклин «Алгоритми кластеризації на службі Data Mining»

**ІНДИВІДУАЛЬНИЙ ГРАФІК ПРОХОДЖЕННЯ НАУКОВО-ДОСЛІДНОЇ ПРАКТИКИ**

студентом 2 курсу магістратури групи ІАВ21 денної форми навчання спеціальності 122 «Комп’ютерні науки», програма «Інформаційна аналітика та впливи»

|  |  |  |  |
| --- | --- | --- | --- |
| Завдання за планом | Термін виконання | Фактичне виконання | Підписи наукового керівника та керівника від  кафедри |
| Розробка індивідуального графіку проходження практики. |  |  |  |
| Формулювання теми магістерський досліджень,  визначення предмету та об’єкту дослідження. |  |  |  |
| Ознайомлення з науковими напрямами роботи установи  на якій проходять практику. |  |  |  |
| Ознайомлення з іноземними та вітчизняними науково- інформаційними джерелами за спеціалізацією, обрання  наукової проблематики та формування бібліографії. |  |  |  |
| Збір та обробка відповідними методами фактичного,  фактологічного та статистичного матеріалу щодо стану об’єкту дослідження. |  |  |  |
| Ознайомлення з нормативно-правовою документацією за  обраною проблематикою та формування напрямів удосконалення щодо стану об’єкту дослідження. |  |  |  |
| Виконання індивідуального завдання з обраної проблеми  досліджень |  |  |  |
| Оформлення звіту з практики |  |  |  |

Узгоджено: дата

**Науковий керівник магістерської роботи**

К.т.н., професор Білощицький А.О.,

(науковий ступінь, вчене звання керівника,

прізвище, ім'я, по-батькові)

**Керівник практики від кафедри**

К.е.н., асистент Мезенцева О.О.,

(науковий ступінь, вчене звання керівника,

прізвище, ім'я, по-батькові)

Зміст

[Вступ 5](#_Toc489132440)

[Теоретична частина 7](#_Toc489132441)

[Загальні відомості про](#_Toc489132442)  АТ "Альфа-банк" [. 7](#_Toc489132442)

Напрями та характеристика наукової роботи [8](#_Toc489132443)

Літературний огляд [9](#_Toc489132441)

[Практична (індивідуальна) частина 11](#_Toc489132444)

Розробка проекту проведення аналітичної діяльності [11](#_Toc489132445)

Методи та засоби обробки інформації  [11](#_Toc489132445)

[Кластерний аналіз](#_Toc489132445). K-means16

[Проблеми](#_Toc489132446) 17

[Завдання](#_Toc489132447) 19

[Етапи](#_Toc489132448) 19

[Методи](#_Toc489132449) 20

[Метод *FOREL*](#_Toc489132450) 21

[Опис алгоритму](#_Toc489132451) 21

[Принцип дії](#_Toc489132452) 23

[Переваги і недоліки](#_Toc489132453) 24

[Застосування](#_Toc489132454) 24

[Висновок](#_Toc489132455) 25

[Список використаної літератури:](#_Toc489132456) 26

# 

# Вступ

**Мета.** Метою виробничої практики є ознайомлення, та в подальшому, закріплення, поглиблення і систематизація знань аналітичної діяльності відділу «Моделювання та портфельного аналізу» в «Альфа-Банку». Також метою виробничої практики є придбання первинного практичного досвіду.

**Завданням** є набуття теоретичних знань і практичних навичок щодо методів і засобів обробки інформації, задля вдосконалення моделювання даних, які будуть використані у проекті дипломної роботи.

**Місце виробничої практики**. «Альфа-Банк» — українська філія російського банку, заснована 1993 року з головним офісом у Києві. За розмірами активів та часткою на роздрібному ринку входить в десятку найбільших банків України.

Загальні активи Альфа-Банку, на кінець 2015 року, становили понад 42 мільярди гривень. Роздрібна мережа обслуговування включає в себе близько 100 відділень та 200 банкоматів.

**Головною ціллю** практики є вивчення отриманих матеріалів, аналіз, систематизація і закріплення знань за весь період навчання.

Цілі практики були досягнуті завдяки вирішенню наступних задач:

* розглянути надіслані теоретичні матеріали, засвоїти новий матеріал;
* здобути практичний досвід роботи з обробкою первинних даних, алгоритмами кластерного аналізу.

В результаті вивчення навчальної дисципліни мною були здобуті такі знання**:**

- правила техніки безпеки, протипожежної безпеки та виробничої санітарії на підприємстві АТ «Альфа-Банк»;

- органiзацiйна i технологiчна структура АТ «Альфа-Банк»;

- схеми функціональних структур керування аналітикою,

- економiчнi показники пiдприємства, органiзацiю працi.

Зачас проходження виробничої практики я закріпила навички**:**

- використовувати набутi в університеті знання в практичній діяльності;

- опрацьовувати науково-технiчну лiтературу i проектну документацію,

організовувати, проектувати i розробляти програмні засоби;

- проводити тестування окремих програм, які є складовою частиною

інформаційної системи;

- програмування алгоритмів і методів, сортування i опрацювання інформації на тему кластерного аналізу, користування науково-технiчною літературою, проектною технiчною i планово-нормативною документацією.

**Зв’язок з іншими дисциплінами**. Виробнича практика надала можливість для засвоєння багатьох дисциплін професійної підготовки, а насамперед таких дисциплін, як , «Інтелектуальний аналіз даних», «Аналіз ринкових процесів», «Методи пошуку та обробки інформації», «Бази даних».

# 

# Теоретична частина

## Загальні відомості про «Альфа-Банк»

«Альфа-Банк» є одним з найбільших українських комерційних банків з міжнародним капіталом. Банк входить в приватну інвестиційну холдингову компанію ABH Holdings SA (ABHH) з головним офісом в Люксембурзі і інвестиціями у фінансові установи в Україні, Білорусі, Казахстані, Нідерландах і Росії, а також представництвами на Кіпрі і в Великобританії. Заснований банк в 1993 році. З 2001 року він здійснює діяльність під брендом Альфа-Банк Україна. Банк займає провідні позиції у всіх сегментах банківського ринку, є одним з найбільш стійких і надійних банків України і входить в Топ-10 фінустанов в країні за обсягом активів, згідно з даними Національного банку України.

За період 2008—2011 капітал Банку був збільшений на $559 млн, а також надано субординовані кредити на суму $248 млн. У I половині 2014 планується збільшення капіталу на $100 млн. Частка ринку по обслуговуванню підприємств великого корпоративного сегменту — 7 %. Розмір інвестицій банку в українські підприємства в 2013 році — $1,290 млрд.

## Напрям та характеристика відділу «Моделювання та портфельного аналізу»

Продуктові кредитні політики, невід'ємною частиною яких є узгоджені ризик-профілі і сам процес схвалення заявки, - це основа успішної моделі роздрібного кредитування в Альфа-Банку Україна. Конкурентна перевага в роздрібному кредитуванні, особливо в беззаставному, ґрунтується на якісному прийнятті рішень. Для досягнення цих цілей Банк використовує найкращі ІТ-рішення (Siebel, Experian), централізований і маневрений процес прийняття рішень в залежності від рівня кредитного ризику, власні та актуальні інструменти прийняття рішень (скорингові карти, системи протидії шахрайству і верифікації, ранє реагування). Роздрібний ризик-менеджмент для управління портфелем використовує статистичні та математичні моделі, на постійній основі проводить моніторинг ефективності продуктів і скорингових моделей, а також впроваджує ефективні способи перевірки клієнтів.

Створюються і щодня оновлюються звіти за показниками ефективності продуктів, на підставі яких контролюється система тригерів в усіх напрямках управління кредитним ризиком роздрібного кредитування.

Напрями та характеристика наукової роботи

Під час наукової практики були розглянуті такі процеси:

* ETL / ELT - процес отримання даних з різнорідних джерел з подальшою їх обробкою та фільтрацією;
* зберігання інформації в базі даних;
* формування аналітичних звітів ;
* середовище для проведення глибокого аналізу даних, описового та прогнозного моделювання, прогнозування часових рядів, оптимізації тощо.
* аналіз клієнтської бази включає управління маркетинговими кампаніями, поведінкова сегментація клієнтів;
* управління маркетинговими ресурсами;
* управління ризиками (операційними, кредитними, ринковими та ін.);
* система оцінки кредитоспроможності особи, що базується на статистичних методах;
* прогнозування попиту.

# Літературний огляд

# «Tidy data» by Hadley Wickham

# Є поширена думка, що до 80% процесу аналізу даних - це час, витрачений на їх підготовку. У даній книзі розглянуто тільки один аспект підготовки даних до аналізу - структурування та приведення в порядок наборів даних, т.зв. «data tidying». Перекласти термін «tidy data» на українську мову можна як «акуратні дані» або «впорядковані дані»

# Підготовка даних - це не лише перший крок, а й багато повторних кроків(алгоритмів) протягом аналізу, коли з'являються нові проблеми або з'являються нові дані.

# Незважаючи на кількість часу, який потрібно витратити, на даний момент не так багато літератури про те, як добре очистити дані. Частиною виклику є широта діяльності, яку вона охоплює, від перевірки відхилень, до розбору дат, пропущених значень і т.д.

# Я скористалася цим посібником, щоб отримати готові алгоритми очистки даних структурування наборів даних для полегшення аналізу.

# Використання алгоритмів полегшує чищення початкових даних, оскільки вам не потрібно починати з нуля і винаходити колесо кожного разу.

# «Practical Data Science with R» Джон Маунт

# Дана книга визначає, що таке наука про дані, чому вона має значення для організацій, а також про те, як створити успішну команду. Представлено широкий аспект підходів, особисті поради та прийоми, а також реальні приклади. Стає більш глибоким розуміння концепцій, що лежать в основі науки про дані.

# Книга представляє багато ідей та їх практичне застосування по обробці та моделюванні даних.

# Н. Паклин «Алгоритми кластеризації на службі Data Mining»

# Кластеризація - об'єднання в групи схожих об'єктів - є однією з фундаментальних завдань в галузі аналізу даних і Data Mining. Список прикладних областей, де вона застосовується, широкий: сегментація зображень, маркетинг, боротьба з шахрайством, прогнозування, аналіз текстів і багато інших. На сучасному етапі кластеризація часто виступає першим кроком при аналізі даних. Після виділення схожих груп застосовуються інші методи, для кожної групи будується окрема модель.

# Дане джерело допомагає краще зрозуміти принципи кластеризації даних та способи застосування.

# Практична (індивідуальна) частина

## 1. Розробка проекту проведення аналітичної діяльності.

## Скорингова модель - це методика оцінки кредитного ризику, яка дозволяє, оцінивши набір ознак, що характеризують позичальника, визначити, чи варто надавати йому кредит.

## В основі скорингової системи лежить припущення, що люди зі схожими соціальними показниками поводяться однаково. Суть скорингу полягає в тому, що кожному параметру, що характеризує позичальника, надається реальна оцінка в балах.

## Побудова скорингової моделі проходить такі етапи:

## C:\Users\User\Desktop\практика\Схема.PNG

## Оскільки практична робота має бути зосереджена на управлінні та обробці інформації, тому надалі розглянемо процеси етапу «Data Preparation».

## 2. Методи та засоби обробки інформації.

## a) portfolio analysis, data gathering, data cleaning.

## Дано репрезентативну вибірку клієнтів банку(1000 клієнтів, 12 змінних).

## \* процес «Data Preparation» реалізований на мові R, скрипт знаходиться у додатку до звіту.

## Дані для вибірки можуть мати неточності, «викривлені» значення та ін. Щоб зробити модель більш точною наведемо кілька прикладів вдосконалення перинної структури даних.

## - пропущені значення

## Часто буває, що у клієнтів відсутні значення у деякій змінні, наприклад, є відсутні значення у змінної «age\_in\_yrs»(вік клієнта, який вимірюється у роках), тоді є кілька варіантів:

## якщо змінна має занадто багато відсутніх значень, то видаляємо її;

## Якщо у змінної дуже мало пропущених значень, то краще видалити клієнтів, у яких є пропущені значення;

## Якщо видалення клієнтів з відсутніми значеннями зменшить суттєво розмір вибірки, то заміняємо прощущені значення середнім значенням або медіаною.

## -викиди (надто малі або надто великі значення)

## якщо виключити викиди із змінної, то це призведе до змін середнього значення та дисперсії. Враховуючи, що викиди також вважаються значеннями, виключення їх з аналізу робить цей підхід неадекватним для «регулювання» змінної;

## типовою стратегією є встановлення всіх відхилень до заданого процентилу даних; наприклад, 98% - це всі дані без викидів. Остальні: 1% - дані з надто малими значеннями, інший 1% - з надто великими значеннями. Ці дані з надто малими та великими значеннями замінюємо на найменші та ,відповідно, найбільші значення 98% даних без викидів.

## b) data binning;

## В скоринговії моделі можуть використовуватися в якості незалежних змінних категоріальні і кількісні предиктори.

## Багато розробників скорингових систем використовують завжди метод категоризації кількісних змінних(binning).

## C:\Users\User\Desktop\практика\binning.PNGКатегоризація кількісних змінних дозволяє досягти таких основних переваг при побудові скорингової карти: полегшити обробку викидів та екстремальних значень кількісних змінних; спростити інтерпретацію скорингової карти; відобразити складні нелінійні зв'язки.

## Розглянемо основні методи бінінгу, які застосовуються у скорингу на основі змінної «age\_in\_yrs».

## Розглянемо емпіричний розподіл змінної(без бінінгу):

## C:\Users\User\Desktop\практика\w_o_binning.PNG

## Методи бінінгу:

## *Equal-width (distance):*

## ділить діапазон на N інтервалів однакового розміру;

## якщо A і B є найнижчими і найвищими значеннями змінної, то ширина інтервалів буде: W = (B-A) / N;

## чутливий до відхилень, нерівні дані не обробляються добре.

## Поділимо змінну «age\_in\_yrs» за допомогою методу «Equal-width (distance)»:

## C:\Users\User\Desktop\практика\eq_width.PNG

## *Equal-depth (frequency):*

## Він ділить діапазон на N інтервалів, кожен з яких містить приблизно однакову кількість спостережень;

## Гарне масштабування даних.

## C:\Users\User\Desktop\практика\eq_depth.PNG

## *“Jenks Natural Breaks” method:*

## метод почергово перебирає всі можливі варіанти класів, рахуючи при цьому міжкласову та міжгрупову дисперсію, отже, вимагає затратів по часу;

## ідея методу полягає у мінімізації дисперсії в класах і максималізації міжкласової дисперсії.

## C:\Users\User\Desktop\практика\jenks.PNG

## Ще одним методом бінінгу, метод кластерного аналізу, - *K-means.*

## Jenks та K-means відрізняються тим, як вони мінімізуються в межах відстаней групи.

## Jenks використовує той факт, що одновимірні дані сортуються, що робить його більш швидким алгоритмом для одновимірних даних.

## K-means є більш загальними алгоритмом, він ефективний із даними великої розмірності.

## Надалі розглянемо методи обробки інформації за допомогою методу *K-means* та застосуємо метод на наших даних.

## 

## Кластерний аналіз. K-means

Інформація, що росте в колосальних обсягах породжує потребу в обробці великих обсягів даних (Big Data). В цьому напрямку важливе місце відведено інтелектуальному аналізу даних (Data Mining). Цей напрямок включає в себе методи, відмінні від класичного аналізу, засновані на моделюванні, імовірнісному асоціюванні і пошуку закономірностей. Розвиток цієї дисципліни спричинений проникненням в сферу аналізу даних ідей, що виникли в теорії штучного інтелекту.

Головним завданням кластерного аналізу є виділення необхідного числа груп об'єктів схожих між собою всередині групи і максимально відмінних від елементів інших класів. Подібний аналіз широко застосовується в інформаційних системах для пошуку закономірностей.

Кластеризація має величезне і широке застосування: це робота з зображеннями, маркетингові дослідження, прогнозування, текстовий аналіз, виявлення шахрайських дій і багато інших. Приклади застосування:

* сегментація клієнтів(розділення клієнтів на групи за певними характеристиками клієнтів) ;
* позиціонування(визначення характеристик, якими повинен володіти товар для того, щоб потрапити в цільовий сегмент);
* аналіз поведінки споживача(побудова однорідних груп споживачів з метою отримання уявлення про те, як поводиться клієнт з кожного сегмента, які драйвери визначають його поведінку).

Завдання кластеризації формуються в різних наукових напрямках: в статистиці, оптимізації, машинному навчанні, що породило безліч різних синонімів поняття кластера - клас, таксон, згущення. Сьогодні вже застосовуються кілька десятків алгоритмів розбиття і безліч їх модифікацій. Найчастіше дані з якими доводиться працювати цим алгоритмам обумовлені такими особливостями як висока розмірність і величезний обсяг даних, так само зміст великої кількості числових (numerical), які здатні упорядочиваться в просторі, і категорійних (categorical) атрибутів. Атрибути отримують значення відповідно до типу шкали, вибір якої є окремим завданням.

У своїй основі алгоритми кластеризації порівнюють між собою об'єкти на основі різних видів подібності або близькості. Наприклад для числових атрибутів часто застосовується евклідова відстань.

## Проблеми

Незважаючи на велику кількість досліджень в області кластерного аналізу, в ході роботи, можна зіштовхнутися з деякими актуальними проблемами. Можна виділити основні з них:

* обгрунтування якості результатів. Проблема полягає в тому, що один і той же об'єкт може бути класифікований в різні групи незалежно від його внутрішніх властивостей, а в зв'язку з різними експертними даними або різним побудовою системи. Для уникнення цього необхідно розробляти і вводити актуальні критерії якості.
* аналіз великої кількості різнотипних даних породжує методологічну проблему вибору метрик. Так само збільшення числа об'єктів навіть однотипних даних може спричинити за собою непомітність відстаней.
* нелінійність взаємозв'язків. Класичні методи зниження розмірності в кластерному аналізі спрямовані на лінійний взаємозв'язок між змінними. Для пошуку більш складних залежностей необхідно переходити до ядерних методів.
* проблема пошуку глобального екстремуму функції критерію якості. Критерій якості, як правило, є функцією, яка залежить від великої кількості факторів, нелінійним, що володіє безліччю локальних екстремумів. Для знаходження кластерів необхідно вирішити складну комбінаторних завдання пошуку оптимального варіанта класифікації. Тому алгоритм повного перебору варіантів має трудомісткість, експоненціально залежить від розмірності. Якщо число груп заздалегідь невідомо, то переборного завдання стає ще складніше. Таким чином, при збільшенні розмірності таблиць даних відбувається «комбінаторний вибух». Класичні алгоритми кластерного аналізу здійснюють спрямований пошук в порівняно невеликому підмножині простору рішень, використовуючи різного роду апріорні обмеження (на число кластерів або їх форму, на порядок включення об'єктів в групи і т.д.). При цьому знаходження суворо-оптимального рішення не гарантується. Для пошуку оптимального рішення застосовуються більш складні методи, такі як генетичні (еволюційні) алгоритми, нейронні мережі і т.д. Існують експериментальні дослідження, що підтверджують переваги таких алгоритмів перед класичними алгоритмами. Однак і при використанні еволюційних методів виникають проблеми, пов'язані зі специфікою розв'язуваної задачі кластер-аналізу: з труднощами інтерпретації використовуваних операторів рекомбінації і кросовера.
* нестійкість результатів кластеризації. Найчастіше результати угруповання можуть сильно змінюватися в залежності від вибору початкових умов, порядку об'єктів, параметрів роботи алгоритмів. Різними авторами пропонуються способи підвищення стійкості группіровочних рішень, засновані на застосуванні ансамблів алгоритмів. При цьому використовуються результати угруповання, отримані різними алгоритмами, або одним алгоритмом, але з різними параметрами налаштування, за різними підсистемами змінних і т.д. Після побудови ансамблю проводиться знаходження підсумкового колективного рішення.
* недостатня кількість знань про об'єкт. Трапляються області, що важко піддаються формалізації, в яких важко створити модель об'єкту. В такому випадку застосування алгоритмів, що грунтуються на уявленні класу, як набору розподілених в просторі змінних, стає неможливим.
* проблема представлення результатів. Крім гарної прогнозуючої здатності для будь-якого алгоритму аналізу даних важливо, як інтерпритуються і наскільки зрозумілими є його результати. Для покращення інтерпретації рішень можна використовувати логічні моделі. Такого роду моделі використовуються для вирішення завдань розпізнавання образів і прогнозування кількісних показників, наприклад, в методах побудови дерев рішень або логічних функцій.

## Завдання

Кластерний аналіз виконує наступні основні завдання:

* Розробка типології або [класифікації](https://uk.wikipedia.org/wiki/Класифікація).
* Дослідження корисних концептуальних схем групування об'єктів.
* Породження [гіпотез](https://uk.wikipedia.org/wiki/Гіпотеза) на основі дослідження даних.
* Перевірка [гіпотез](https://uk.wikipedia.org/wiki/Гіпотеза) або дослідження для визначення, чи дійсно групи, виділені тим чи іншим способом, присутні в наявних даних.

## Етапи

Незалежно від конкретної сфери, застосування кластерного аналізу передбачає наступні етапи:

* Відбір [вибірки](https://uk.wikipedia.org/wiki/Вибірка) для кластеризації.
* Визначення множини характеристик, по яких будуть оцінюватися об'єкти у вибірці.
* Обчислення значень тієї чи іншої [міри схожості](https://uk.wikipedia.org/w/index.php?title=Міра_схожості&action=edit&redlink=1) між об'єктами.
* Застосування одного з методів кластерного аналізу для створення груп схожих об'єктів.
* Перевірка достовірності результатів кластеризації.

Якщо кластерному аналізу передує [факторний аналіз](https://uk.wikipedia.org/wiki/Факторний_аналіз), то вибірка не потребує коректування — викладені вимоги виконуються автоматично самою процедурою факторного [моделювання](https://uk.wikipedia.org/wiki/Моделювання). В іншому випадку вибірку потрібно коректувати.

Методи

Об'єднання схожих об'єктів у групи може бути здійснене різними способами. Саме для цього етапу існує цілий ряд методів:

* [K-середніх (K-means)](https://uk.wikipedia.org/w/index.php?title=Кластеризація_зображення_методом_к-середніх&action=edit&redlink=1)
* [Нечітка кластеризація C-середніх (C-means)](https://uk.wikipedia.org/w/index.php?title=Нечітка_кластеризація_методом_с-середніх&action=edit&redlink=1)
* [Графові алгоритми](https://uk.wikipedia.org/w/index.php?title=Графові_алгоритми&action=edit&redlink=1) кластеризації
* [Статистичні алгоритми](https://uk.wikipedia.org/w/index.php?title=Статистичні_алгоритми&action=edit&redlink=1) кластеризації
* Алгоритми сімейства [FOREL](https://uk.wikipedia.org/w/index.php?title=FOREL&action=edit&redlink=1)
* [Ієрархічна кластеризація](https://uk.wikipedia.org/wiki/Ієрархічна_кластеризація) або [таксономія](https://uk.wikipedia.org/wiki/Таксономія)
* [Нейронна мережа Кохонена](https://uk.wikipedia.org/wiki/Нейронна_мережа_Кохонена)
* Ансамбль кластеризаторів
* Алгоритми сімейства [KRAB](https://uk.wikipedia.org/w/index.php?title=KRAB&action=edit&redlink=1)
* [ЕМ-алгоритм](https://uk.wikipedia.org/wiki/ЕМ-алгоритм)
* Метод просіювання

### Метод FOREL

Кластеризація методом k-середніх ([англ.](https://uk.wikipedia.org/wiki/Англійська_мова) *k-means clustering*) — популярний метод [кластеризації](https://uk.wikipedia.org/wiki/Кластеризація), — впорядкування множини об'єктів в порівняно однорідні групи. Винайдений в 1950-х роках математиком [Гуґо Штайнгаузом](https://uk.wikipedia.org/wiki/Гуґо_Штайнгауз) і майже одночасно Стюартом Ллойдом. Особливу популярність отримав після виходу роботи МакКвіна.

Мета методу — розділити *n* спостережень на *k* кластерів, так щоб кожне спостереження належало до кластера з найближчим до нього [середнім значенням](https://uk.wikipedia.org/wiki/Середнє_арифметичне). Метод базується на мінімізації суми квадратів відстаней між кожним спостереженням та центром його кластера, тобто функції

,

де d — метрика, xi{\displaystyle x\_{i}} — і-ий об'єкт даних, а {\displaystyle m\_{j}(x\_{i})}mj(xi) — центр кластера, якому на j-ій ітерації приписаний елемент {\displaystyle x\_{i}} xi{\displaystyle x\_{i}} .

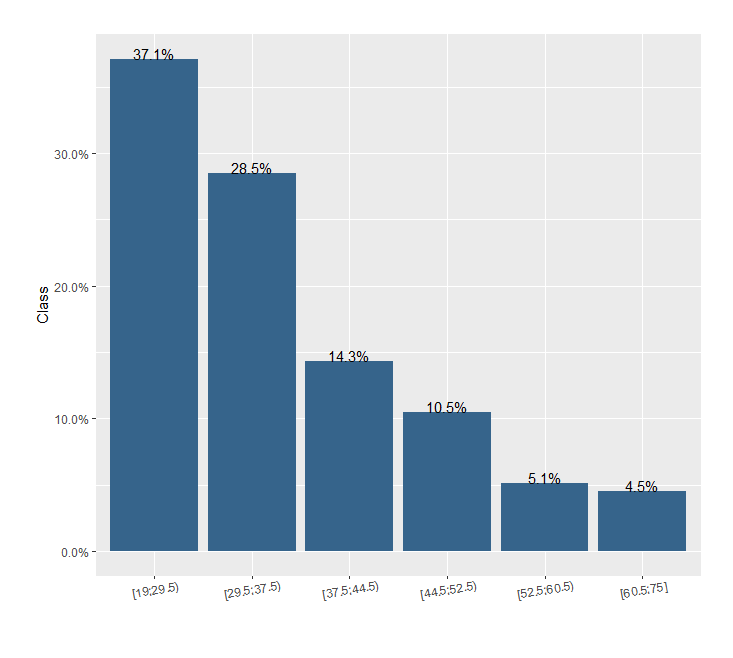
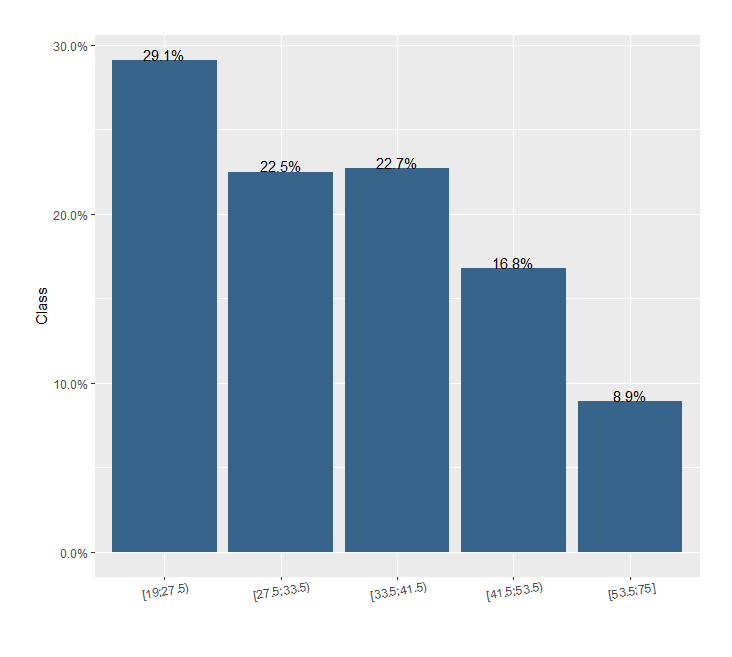
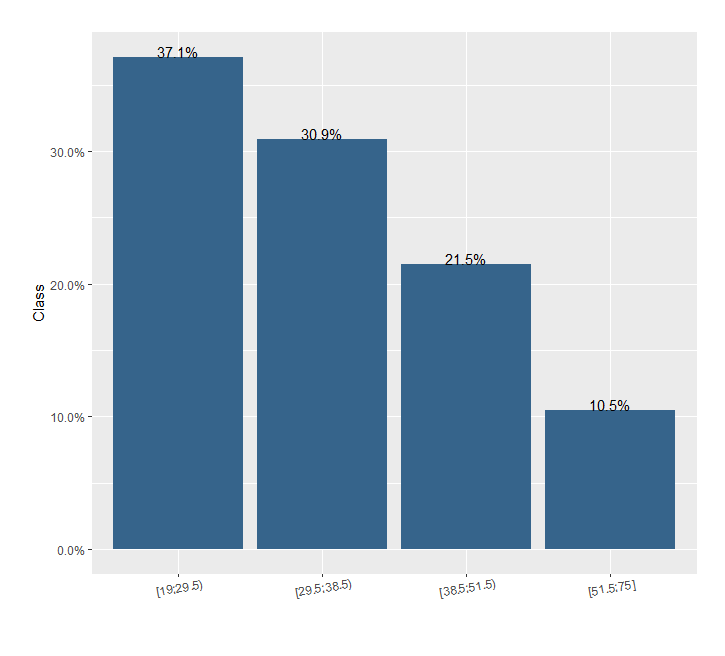
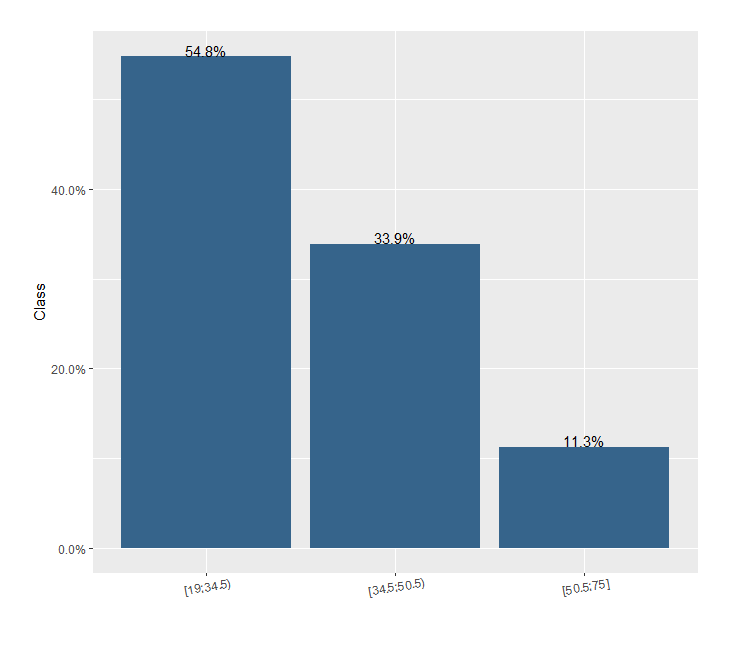
#### Опис алгоритму

Маємо масив спостережень (об'єктів), кожен з яких має певні значення по ряду ознак. Відповідно до цих значень об'єкт розташовується у багатовимірному просторі.

1. Дослідник визначає кількість кластерів, що необхідно утворити
2. Випадковим чином обирається k спостережень, які на цьому кроці вважаються центрами кластерів
3. Кожне спостереження «приписується» до одного з n кластерів — того, відстань до якого найкоротша
4. Розраховується новий центр кожного кластера як елемент, ознаки якого розраховуються як середнє арифметичне ознак об'єктів, що входять у цей кластер
5. Відбувається така кількість ітерацій (повторюються кроки 3-4), поки кластерні центри стануть стійкими (тобто при кожній ітерації в кожному кластері опинятимуться одні й ті самі об'єкти), дисперсія всередині кластера буде мінімізована, а між кластерами — максимізована

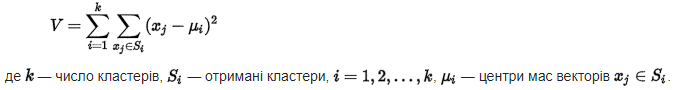
Вибір кількості кластерів відбувається на основі дослідницької гіпотези. Якщо її немає, то рекомендують створити 3 кластери, далі 4,5,6 порівнюючи отримані результати.

Застосуємо даний метод на наших даних.



#### Принцип дії

Принцип алгоритму полягає в пошуку таких центрів кластерів та наборів елементів кожного кластера при наявності деякої функції Ф(°), що виражає якість поточного розбиття множини на k кластерів, коли сумарне квадратичне відхилення елементів кластерів від центрів цих кластерів буде найменшим:



В початковий момент роботи алгоритму довільним чином обираються центри кластерів, далі для кожного елемента множини [ітеративно](https://uk.wikipedia.org/wiki/Ітерація) обраховується відстань від центрів з приєднанням кожного елемента до кластера з найближчим центром. Для кожного з отриманих кластерів обчислюються нові значення центрів, намагаючись при цьому мінімізувати функцію Ф(°), після чого повторюється процедура перерозподілу елементів між кластерами.

[Алгоритм](https://uk.wikipedia.org/wiki/Алгоритм) методу «Кластеризація за схемою к-середніх»:

* вибрати k інформаційних точок як центри кластерів поки не завершиться процес зміни центрів кластерів;
* зіставити кожну інформаційну точку з кластером, відстань до центра якого мінімальна;
* переконатися, що в кожному кластері міститься хоча б одна точка. Для цього кожний порожній кластер потрібно доповнити довільною точкою, що розташована «далеко» від центра кластера;
* центр кожного кластера замінити середнім від елементів кластера;
* кінець.

#### Переваги і недоліки

Головні переваги методу k-середніх — його простота та швидкість виконання. Метод k-середніх більш зручний для кластеризації великої кількості спостережень, ніж метод ієрархічного кластерного аналізу (у якому дендограми стають перевантаженими і втрачають наочність).

Одним із недоліків простого методу є порушення умови зв'язності елементів одного кластера, тому розвиваються різні модифікації методу, а також його нечіткі аналоги ([англ.](https://uk.wikipedia.org/wiki/Англійська_мова) *fuzzy k-means methods*), у яких на першій стадії алгоритму допускається приналежність одного елемента множини до декількох кластерів (із різним ступенем приналежності).

Незважаючи на очевидні переваги методу, він має суттєві недоліки:

1. Результат класифікації сильно залежить від випадкових початкових позицій кластерних центрів
2. Алгоритм чутливий до викидів, які можуть викривлювати середнє
3. Кількість кластерів повинна бути заздалегідь визначена дослідником

#### Застосування

Метод k-середніх є доволі простим і прозорим, тому успішно використовується у різноманітних сферах — маркетингових сегментаціях, геостатистиці, астрономії, сільському господарстві тощо.

# Висновок

При проходженні виробничої практики були розглянуті та впроваджені аналітичні процеси відділу «Моделювання та портфельного аналізу» АТ «Альфа-Банк». Завдяки здобутій практичній діяльності було впроваджено проект дипломної роботи та ознайомлення з новими методами та засобами обробки інформації для успішної реалізації проекту.

В результаті обраної теми при проходженні виробничої практики були розглянуті основні техніки «Data pre-processing», методи кластерного аналізу і їх застосування. Під час дослідження теми кластерного аналізу були розглянуті методи, що застосовуються в сучасних системах, з яких був обраний метод к-середніх.

Для прикладу була використана тестова вибірка даних. За початкові дані були взяті репрезентативна вибірка банку невеликого об’єму.

Реалізація методів була успішно проведена на мові R.

# 

# Список використаної літератури:

1. «Tidy data» - Hadley Wickham.
2. «Practical Data Science with R» Джон Маунт
3. Н. Паклин «Алгоритми кластеризації на службі Data Mining»
4. Patricia B. Cerrito «Introduction to Data Mining»
5. Dr. Winston W. Rovce ANAGING THE DEVELOPMENT OF LARGE SOFTWARE SYSTEMS
6. Х. Марманис, Д. Бабенко (пер. М. Низовец) Алгоритми інтелектуального Інтернету. Передові [методики збору](http://nethash.ru/oborudovanie-i-metodiki-nauchnogo-parka-sankt-peterburgskogo-g.html), аналізу і обробки даних (Algorithms of the Intelligent Web)
7. Daniel Fasulo «An Analysis Of Recent Work on Clustering Algorithms»