

РОЗВІДУВАЛЬНИЙ АНАЛІЗ ДАНИХ ДЛЯ ІНФОРМАЦІЙНОЇ ТЕХНОЛОГІЇ АНАЛІЗУ ФАКТОРІВ, ЩО ВПЛИВАЮТЬ НА РІВЕНЬ ЗАРОБІТНОЇ ПЛАТИ

Вінницький національний технічний університет

Анотація

У роботі проведено розвідувальний аналіз даних для подальшого використання в інформаційній технології аналізу факторів (демографічні характеристики, професійний досвід, рівень освіти, займана посада), що впливають на рівень заробітної плати працівників. Виконано попередню обробку даних, проаналізовано розподіли числових і категоріальних ознак, а також досліджено залежність заробітної плати від досвіду роботи з урахуванням соціально-демографічних факторів. Отримані результати дозволяють виявити ключові чинники, що впливають на формування доходів, та можуть бути використані для побудови прогнозних моделей у задачах аналізу ринку праці.

Ключові слова: інформаційні технології, аналіз даних, розвідувальний аналіз даних, заробітна плата, досвід роботи, рівень освіти, регресійний аналіз.

Abstract

The paper presents an exploratory data analysis for further use in information technology analysis of factors (demographic characteristics, professional experience, level of education, position held) that affect the level of employee wages. Data preprocessing was performed, distributions of numerical and categorical features were analyzed, and the relationship between salary and years of experience was investigated considering socio-demographic factors. The obtained results allow identifying key factors influencing income formation and can be used for building predictive models in labor market analysis tasks.

Keywords: information technology, data analysis, exploratory data analysis, salary, work experience, education level, regression analysis.

Вступ

Сучасні інформаційні технології відіграють важливу роль у дослідженні соціально-економічних процесів, зокрема у сфері аналізу ринку праці. Одним із ключових показників, що характеризує соціально-економічний стан населення, є рівень заробітної плати, який формується під впливом багатьох факторів, таких як досвід роботи, рівень освіти, професійна діяльність та демографічні характеристики.

Застосування методів аналізу даних та машинного навчання дозволяє виявляти приховані закономірності у великих наборах соціально-економічної інформації та будувати моделі для прогнозування доходів. Важливим етапом створення таких моделей є розвідувальний аналіз даних (Exploratory Data Analysis), який дає змогу оцінити структуру набору даних, дослідити розподіли ознак та встановити взаємозв'язки між ключовими змінними [1].

Отже, метою роботи є проведення розвідувального аналізу даних щодо рівня заробітної плати з метою виявлення ключових факторів, що впливають на її формування, та підготовки даних для подальшого застосування методів регресійного аналізу та машинного навчання в інформаційній технології аналізу ринку праці.

Розвідувальний аналіз

Для проведення розвідувального аналізу було обрано набір даних «Salary Data» [2], який містить інформацію про заробітні плати працівників залежно від демографічних характеристик, рівня освіти, професійної діяльності та досвіду роботи. Даний датасет є загальнодоступним та використовується для досліджень у сфері аналізу даних і соціально-економічного моделювання. Він включає 6698 записів та 6 атрибутів, що описують основні фактори, які можуть впливати на рівень заробітної плати.

Цільовою змінною у даному дослідженні є показник Salary, який має числовий характер і відображає річний рівень заробітної плати. Рівень заробітної плати може залежати від таких факторів, як вік працівника (Age), кількість років професійного досвіду (Years of Experience), рівень освіти (Education Level), професійна посада (Job Title), а також статі працівника (Gender).

Серед ознак наявні як числові атрибути — вік, досвід роботи та заробітна плата, так і категоріальні — стать, рівень освіти та професійна посада. Така структура даних дозволяє провести комплексний розвідувальний аналіз (EDA), дослідити розподіли ознак, а також виявити залежності між досвідом роботи, соціальними факторами та рівнем оплати праці. На рисунку 1 представлено фрагмент датасету «Salary Data», що містить основні атрибути.

	Age	Gender	Education Level	Job Title	Years of Experience	Salary
0	32.0	Male	Bachelor's	Software Engineer	5.0	90000.0
1	28.0	Female	Master's	Data Analyst	3.0	65000.0
2	45.0	Male	PhD	Senior Manager	15.0	150000.0
3	36.0	Female	Bachelor's	Sales Associate	7.0	60000.0
4	52.0	Male	Master's	Director	20.0	200000.0

Рис. 1. Приклад атрибутів, що містить набір даних

На етапі попередньої обробки даних було виконано очищення набору від пропущених значень, зокрема для ознак віку та досвіду роботи [1, 3]. Після цього числові змінні було приведено до цілочисельного формату. Для оцінки загальної структури даних побудовано гістограми розподілу віку, років досвіду та заробітної плати, а також діаграми частот для категоріальних ознак (рис. 2) [4, 5].

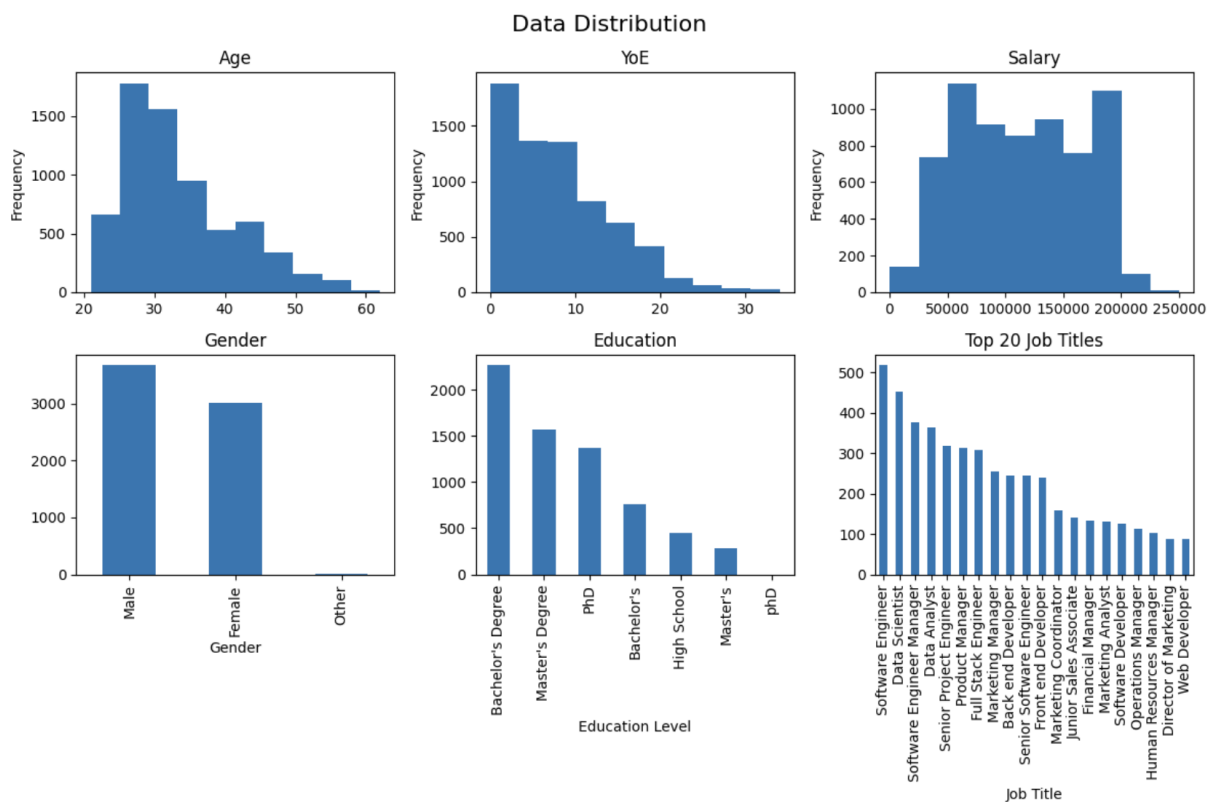


Рис. 2. Розподіл основних ознак набору даних

З отриманих графіків видно, що розподіл віку та років досвіду має правосторонню асиметрію, що свідчить про переважання працівників молодого та середнього віку з помірним рівнем професійного досвіду. Розподіл заробітної плати характеризується значною варіативністю та наявністю високих значень, що може вказувати на суттєву різницю в оплаті праці залежно від професійної діяльності та рівня кваліфікації. Аналіз категоріальних змінних показав домінування окремих рівнів освіти та професійних посад у вибірці.

Для детальнішого аналізу було досліджено 20 найпопулярніших професій, представлених у наборі даних, та відповідні діапазони заробітної плати. Для цього побудовано діаграми розмаху (boxplot), які дозволяють оцінити медіанні значення заробітної плати [4] та її варіативність для кожної професійної посади (рис. 3).

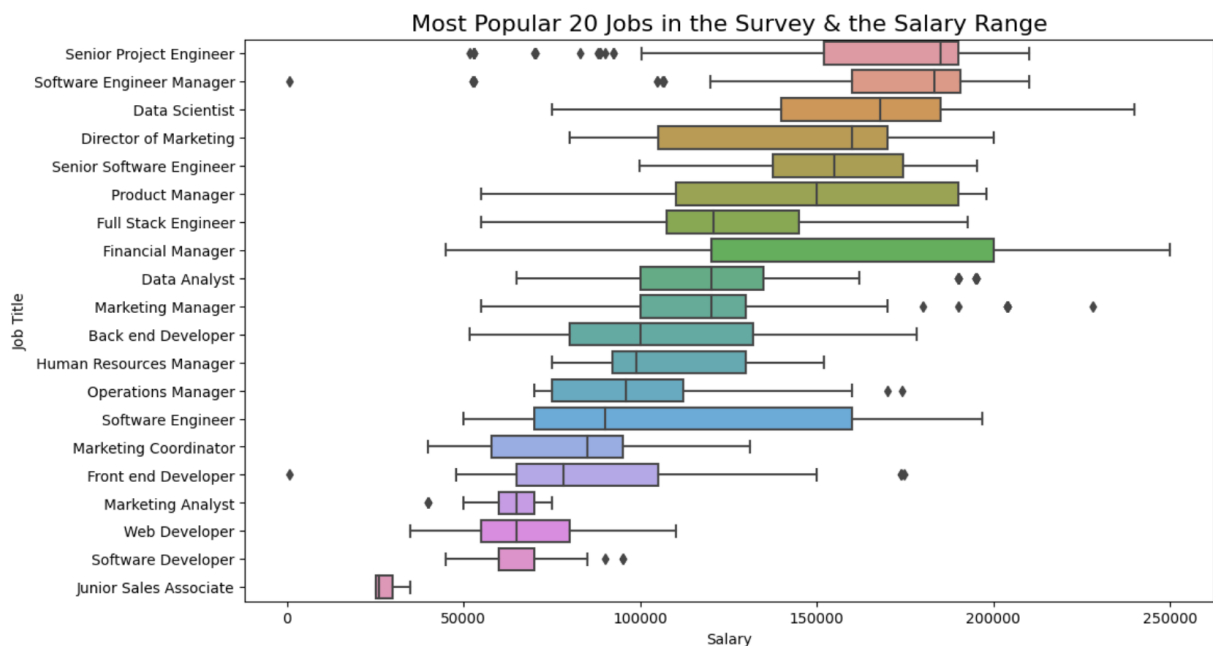


Рис. 3. Найпопулярніші професії та діапазон заробітної плати

З графіку, представленого на рисунку 3, видно, що різні професійні посади мають суттєві відмінності як за рівнем оплати праці, так і за розкидом значень, що свідчить про значний вплив професійного напрямку на рівень доходу.

Подальший аналіз було зосереджено на дослідженні взаємозв'язку між заробітною платою та досвідом роботи з урахуванням статі, рівня освіти та професійної діяльності. Побудовані регресійні графіки показали позитивну залежність між кількістю років досвіду та рівнем заробітної плати, причому характер цієї залежності може відрізнятися для різних соціально-демографічних груп (рис. 4).

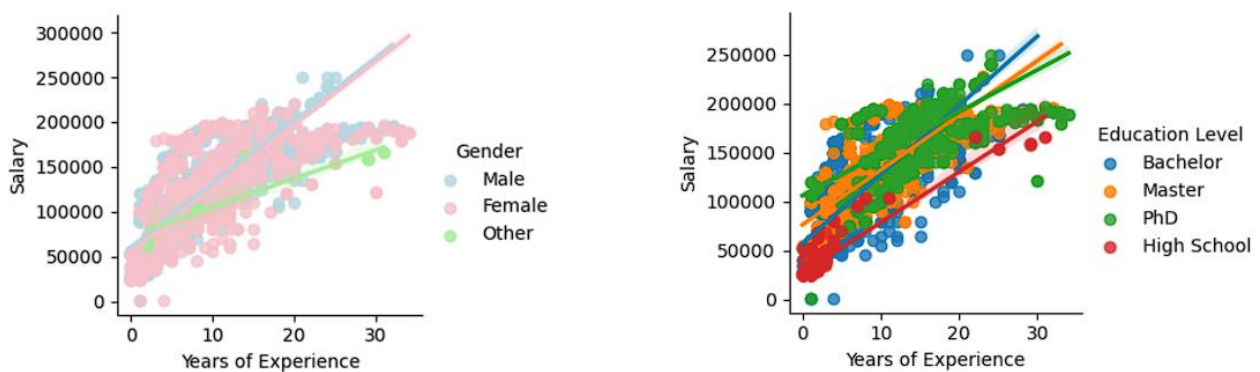


Рис. 4. Залежність заробітної плати від досвіду роботи з урахуванням статі та рівня освіти

Крім того, було виконано лінійний регресійний аналіз для найпопулярніших професій, що дозволило оцінити швидкість зростання заробітної плати залежно від досвіду роботи для кожної професійної групи. Отримані результати показують, що для окремих професійних посад зростання доходу з досвідом є більш вираженим, ніж для інших, що може бути використано для подальшого прогнозування та аналізу ринку праці.

Отримані результати розвідувального аналізу підтверджують значний вплив досвіду роботи, рівня освіти та професійної діяльності на формування заробітної плати та можуть бути використані для побудови моделей регресії та прогнозування доходів у подальших дослідженнях.

Висновки

У роботі проведено розвідувальний аналіз даних, спрямований на дослідження факторів, що впливають на рівень заробітної плати працівників. Виконано попередню обробку даних, яка включала очищення набору від пропущених значень, приведення ознак до уніфікованого формату та перевірку їхньої коректності, що забезпечило підготовку якісної вибірки для подальшого аналізу.

У процесі аналізу було досліджено розподіли основних числових і категоріальних змінних, зокрема віку, професійного досвіду, рівня освіти та займаної посади. Побудовані графічні візуалізації дозволили виявити асиметричність розподілів та значну варіативність рівня заробітної плати, що свідчить про неоднорідність досліджуваної сукупності та наявність істотних відмінностей між окремими професійними групами.

Встановлено наявність позитивної залежності між кількістю років професійного досвіду та рівнем заробітної плати, що узгоджується з загальними економічними закономірностями формування доходів. Водночас виявлено, що характер цієї залежності відрізняється для різних рівнів освіти та професійних посад, що свідчить про комплексний вплив соціально-демографічних і професійних чинників на формування заробітної плати.

Проведений аналіз дозволив ідентифікувати ключові фактори, що мають найбільший вплив на рівень доходу, а також виявити відмінності у динаміці зростання заробітної плати для різних професійних груп. Отримані результати можуть бути використані як основа для побудови регресійних та прогнозних моделей, а також для розроблення інформаційних технологій підтримки прийняття рішень у сфері аналізу ринку праці.

Таким чином, виконаний розвідувальний аналіз підтвердив ефективність застосування сучасних методів аналізу даних для дослідження соціально-економічних показників та продемонстрував можливість подальшого використання отриманих результатів у задачах прогнозування, моделювання та оцінювання факторів, що впливають на рівень заробітної плати.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Мокін В.Б., Дратованій М.В. Наука про дані: машинне навчання та інтелектуальний аналіз даних. Навчальний посібник. м. Вінниця, 2024 р. 263 с. – Режим доступу: <https://iq.vntu.edu.ua/repository/getfile.php/8163.pdf>
2. Salary_Data Dataset. Kaggle. [Електронний ресурс] – Режим доступу: <https://www.kaggle.com/datasets/mohithsairamreddy/salary-data>
3. Pandas Tutorial. [Електронний ресурс] – Режим доступу: <https://www.w3schools.com/python/pandas/default.asp>
4. Matplotlib Pyplot Documentation. [Електронний ресурс] – Режим доступу: https://matplotlib.org/3.5.3/api/as_gen/matplotlib.pyplot.html
5. A Complete Guide to Data Visualization in Python With Libraries & More. [Електронний ресурс] – Режим доступу: <https://www.simplilearn.com/tutorials/python-tutorial/data-visualization-in-python>

Лавренюк Євгеній Михайлович – студент групи СА-236, факультет інтелектуальних інформаційних технологій та автоматизації, Вінницький національний технічний університет, м. Вінниця. e-mail: lavrenyuk0629@gmail.com

Войцеховська Ольга Олександрівна – PhD, доцент кафедри системного аналізу та інформаційних технологій, Вінницький національний технічний університет, м. Вінниця, e-mail: olgav1085@gmail.com

Lavreniuk Yevhenii M. – student of Faculty of Intellectual Information Technologies and Automation, SA-23b, Vinnytsia National Technical University, Vinnytsia, e-mail lavrenyuk0629@gmail.com

Voitsekhovska Olha O. – PhD, Associate Professor of the Department of System Analysis and Information Technologies, Vinnytsia National Technical University, Vinnytsia, e-mail: olgav1085@gmail.com