

ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ АНАЛІЗУ ТА ПЕРЕДБАЧЕННЯ ЗАРОБІТНОЇ ПЛАТИ У СФЕРІ DATA SCIENCE У 2023 РОЦІ

Вінницький національний технічний університет

Анотація

Для розробки інформаційної технології аналізу та передбачення заробітної плати у сфері Data Science у 2023 році було здійснено збір та обробку відповідних даних, які стосуються динаміки заробітної плати в даному сегменті. Використовуючи передові інструменти аналізу даних та технології обробки інформації, проведено детальний аналіз параметрів, що впливають на зміни в рівні оплати праці у галузі Data Science. Розроблена система надає можливість глибокого вивчення факторів, що визначають заробітну плату в цій сфері, що сприяє більш точному розумінню ринкових умов та прийняттю ефективних управлінських рішень.

Ключові слова: Інформаційна технологія, аналіз даних, передбачення, заробітна плата, Data Science.

Abstract

For the development of information technology for the analysis and prediction of wages in the field of Data Science in 2023, the collection and processing of relevant data related to the dynamics of wages in this segment was carried out. Using advanced data analysis tools and information processing technologies, a detailed analysis of the parameters affecting changes in the level of remuneration in the field of Data Science was carried out. The developed system provides an opportunity for in-depth study of the factors that determine wages in this area, which contributes to a more accurate understanding of market conditions and the adoption of effective management decisions.

Keywords: Information Technology, Data Analysis, Forecasting, Salary, Data Science.

Актуальність дослідження

У сучасному інформаційному суспільстві, визначеному стрімким темпом технологічного розвитку, однією з ключових сфер виявляється Data Science. В контексті цієї області виникає нагальна потреба в розробці і вдосконаленні методів аналізу та передбачення заробітної плати для спеціалістів у цьому сегменті ринку праці. Актуальність дослідження полягає в тому, що зростаюча конкуренція та сталий розвиток технологій вимагають від компаній та фахівців у галузі Data Science не тільки високого рівня кваліфікації, але й адаптації до динамічного ринкового середовища. [1].

Постановка задачі

Метою дослідження є розробка та впровадження інформаційно-технологічних засобів для аналізу та передбачення заробітної плати у сфері Data Science на ринку праці у 2023 році. З цією метою передбачається вивчення актуальних тенденцій, визначення факторів, що впливають на заробітну плату, та розробка моделей для її передбачення.

Розробка інформаційної технології аналізу та передбачення заробітної плати

Для розробки інформаційної технології використовувалось середовище Kaggle та завантажений датасет із зібраними даними. На рис 1. зображено фрагмент з датасету.

```
[52]: # Load data
df = pd.read_csv(filepath_or_buffer=file_path)

# Quick look
df.head()
```

	work_year	experience_level	employment_type	job_title	salary	salary_currency	salary_in_usd	employee_residence	remote_ratio	company_location	company_size
0	2023	SE	FT	Principal Data Scientist	80000	EUR	85847	ES	100	ES	L
1	2023	MI	CT	ML Engineer	30000	USD	30000	US	100	US	S
2	2023	MI	CT	ML Engineer	25500	USD	25500	US	100	US	S
3	2023	SE	FT	Data Scientist	175000	USD	175000	CA	100	CA	M
4	2023	SE	FT	Data Scientist	120000	USD	120000	CA	100	CA	M

Рис.1 – Огляд даних в датасеті.

На основі зібраних вхідних даних розроблено UML-діаграму для розробки інформаційної технології (рис. 2)

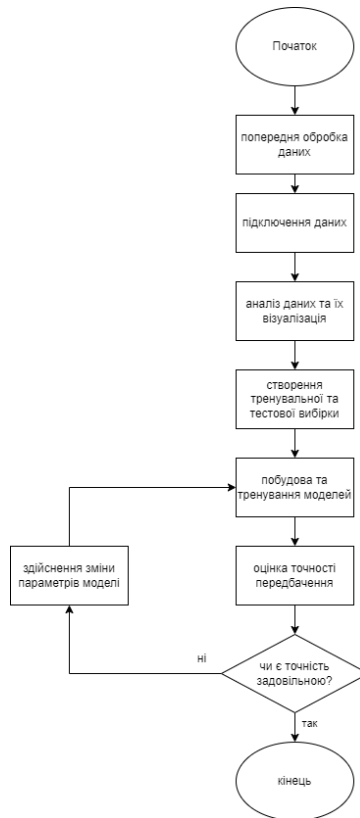


Рис.2 - UML-діаграма розробки інформаційної технології

Отримані результати

В результаті виконання роботи було виконано розвідувальний аналіз (EDA) та побудовано декілька діаграм на основі зібраних даних [4]. На рисунку 3 зображено діаграму розподілу виплат за роками.

Distribution of payments made over the years

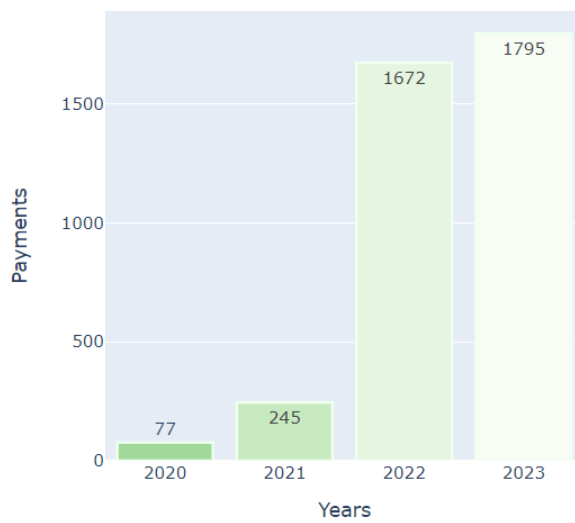


Рис. 3 - Діаграма розподілу виплат за роками.

На рисунку 4 зображено діаграму вивчення середньої заробітної плати за рік.

Examining the average salary per year

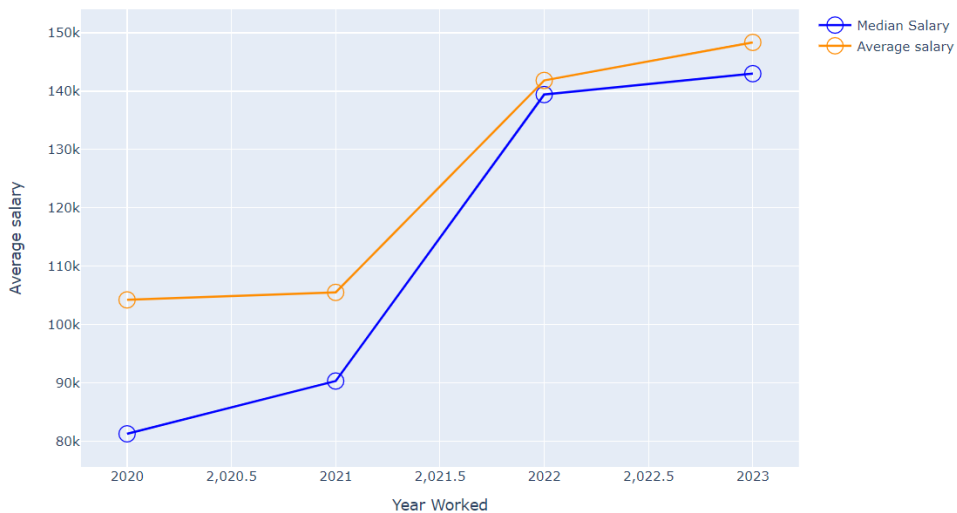


Рис. 4 - Діаграми вивчення середньої заробітної плати за рік.

Також було побудовано декілька моделей машинного навчання [2]. Шляхом очистки аномальних даних, Decision Tree Regressor має найвищу точність на тренувальних з показником 0.955492, та на тестових даних 0.976195 за метрикою R2 Score [3] (рис. 5).

[56]:	Модель	Точність на тренувальних даних	Точність на тестових даних
5	Decision Tree Regressor	0.955492	0.976195
4	XGBoost Regressor	0.953948	0.969675
3	Support Vector Regression	0.930543	0.944785
2	Random Forest Regressor	0.901650	0.912061
1	Lasso Regression	0.831197	0.857584
0	Linear Regression	0.722010	0.754830

Рис. 5 - Результат виконання моделей.

Висновки

У дослідженні було розроблено інформаційну технологію для аналізу та передбачення заробітної плати, вибравши оптимальні технології. Зібрані та проаналізовані дані, побудовано ряд візуалізацій у вигляді графіків та обрано оптимальну модель машинного навчання, що сприятиме ефективній адаптації до змін в ринковому середовищі для компаній та фахівців у галузі Data Science.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. CAO, Longbing. Data science: a comprehensive overview. ACM Computing Surveys (CSUR), 2021, 50.3: 1-42. [Електронний ресурс]: Режим доступу: <https://dl.acm.org/doi/abs/10.1145/3076253>
2. Machine Learning - Машинне навчання [Електронний ресурс]: Режим доступу: <https://www.it.ua/knowledge-base/technology-innovation/machine-learning>
3. Dolenko B. Kaggle Notebook «Salary Prediction» [Електронний ресурс]. Режим доступу: <https://www.kaggle.com/code/bohdandolenko/salary-prediction>
4. MATBOULI, Yasser T.; ALGHAMDI, Suliman M. Statistical machine learning regression models for salary prediction featuring economy wide activities and occupations. Information, 2022, 13.10: 495. [Електронний ресурс]. Режим доступу: <https://www.mdpi.com/2078-2489/13/10/495/pdf>

Варчук Ілона Вячеславівна – канд. техн. наук, доцент кафедри системного аналізу та інформаційних технологій, Вінницький національний технічний університет, Вінниця, e-mail: ilona.varchuk@gmail.com;

Доленко Богдан Анатолійович – студент групи 2ІСТ-22м, Факультет інтелектуальних інформаційних технологій та автоматизації, Вінницький національний технічний університет, Вінниця, e-mail: edubodya@ukr.net;

Varchuk Ilona – Cand. Sc. (Eng), Department of Systems Analysis and Information Technology, Vinnytsia National Technical University, Vinnytsia, e-mail: ilona.varchuk@gmail.com;

Dolenko Bohdan - student of 2IST-22m group, Faculty of Intelligent Information Technologies and Automation, Vinnytsia National Technical University, Vinnytsia, e-mail: edubodya@ukr.net.