МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ЧОРНОМОРСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ ІМЕНІ ПЕТРА МОГИЛИ

Соколюк Антон Вікторович

УДК 004.89

ДОСЛІДЖЕННЯ МЕТОДІВ МАШИННОГО НАВЧАННЯ ДЛЯ БІНАРНОЇ ІДЕНТИФІКАЦІЇ ЗАХВОРЮВАНЬ ПЕЧІНКИ

122 – Комп’ютерні науки

Автореферат
магістерської наукової роботи на здобуття освітньої кваліфікації
«Магістр комп’ютерних наук»

Миколаїв – 2020

Магістерська наукова робота є рукопис.

Робота виконана в Чорноморському національному університеті імені Петра Могили Міністерства освіти і науки України на кафедрі інтелектуальних інформаційних систем

Науковий керівник: к.т.н., доцент, доцент кафедри інтелектуальних інформаційних систем Сіденко Євген Вікторович

Рецензент: к.т.н., доцент, доцент кафедри комп'ютерної інженерії Крайник Ярослав Михайлович

Захист відбудеться «25» лютого 2020 р. о 9^{30} год. на засіданні екзаменаційної комісії (ауд. 2-403) у Чорноморському національному університеті імені Петра Могили за адресою: 54003, м. Миколаїв, вул. 68-ми Десантників, 10.

З магістерською науковою роботою можна ознайомитися в бібліотеці Чорноморського національного університету імені Петра Могили за адресою: 54003, м. Миколаїв, вул. 68-ми Десантників, 10.

Автореферат представлений «___» лютого 2020 р.

Секретар
екзаменаційної комісії,
к.пед.н., доцент Н. М. Болюбаш

ЗАГАЛЬНА ХАРАКТЕРИСТИКА РОБОТИ

Актуальність дослідження полягає у тому, що створення автоматизованої системи діагностики захворювань печінки може попередити хворих або потенційно хворих людей про наявні проблеми, і завчасне відвідування лікаря та отримання медичної допомоги може запобігти перетіканню хвороб у більш тяжку стадію. Також такі системи при наявності відповідної точності можуть допомогти спеціалістам з встановленням точного діагнозу при наявності розпливчатих даних.

Метою магістерської наукової роботи є полегшення праці лікарів шляхом створення автоматизованих систем діагностування захворювань печінки.

Об'єктом дослідження є сфера машинного навчання для задач медичної діагностики.

Предметом дослідження є методи машинного навчання для бінарної ідентифікації захворювань печінки.

Практичне значення даної магістерської наукової роботи полягає у оцінці ефективності методів машинного навчання для ідентифікації захворювань печінки.

Результати даної магістерської наукової роботи було надруковано у тезах XXI Всеукраїнської науково-методичної конференції «Могилянські читання – 2019» у секції Комп'ютерні науки, та у тезах Всеукраїнської науково-практичної конференції молодих вчених, аспірантів і студентів «Інтелектуальні інформаційні системи» 2020 року.

Магістерська наукова робота складається із вступу, 6 розділів, висновків, додатків. Загальний обсяг роботи складає 88 сторінки, 40 рисунків, 5 таблиць та 18 посилань на літературні джерела.

ОСНОВНИЙ ЗМІСТ РОБОТИ

У вступі магістерської наукової роботи обґрунтовано актуальність обраної теми, сформульовано мету і задачі дослідження, визначено предмет та об’єкт дослідження.

У першому розділі дано загальний опис захворювань печінки і важливості їх попереднього діагностування, видів машинного навчання (з вчителем, без вчителя, із підкріпленням), проаналізовані наявні дослідження по темі, обраний набір даних Indian Liver Patient Dataset з результатами біохімічний аналізів 583 індійських пацієнтів по 11 атрибутів. Також були сформовані кроки потрібні для проведення дослідження : вивчення і вибір доступних бібліотек машинного навчання, тестування різних алгоритмів машинного навчання, створення програмного додатку із використанням кращого алгоритму.

Для аналізу методів машинного навчання був обраний набір даних Indian Liver Patient Dataset (ILPD) що знаходиться у відкритому доступі у репозиторії для машинного навчання UCI .

Набір містить дані 583 індійських пацієнтів з району Andhra Pradesh по 11 атрибутам наведеним у табл. 1.1.

Таблиця 1. Атрибути набору даних

Атрибут	Тип
Стать	Категорія (Чоловіча, Жіноча)
Вік	Число
Загальний білірубін	Число
Прямий білірубін	Число
Загальні протеіни	Число
Альбумін	Число
Співвідношення альбуміну та глобуліну	Число
SGPT (Аланін амінотрансфераза)	Число

SGOT(Аспартат-амінотрансфераза);	Число
ALP (Лужна фосфатаза).	Число
Наявність захворювань	Категорія (1– хвора, 2– здорова людина)

У другому розділі дано опис бібліотекам машинного навчання що були обрані для аналізу: відносна нова бібліотека Microsoft ML.Net для вирішення задач машинного навчання за допомогою C#,відкрита бібліотека Skikit-learn на мові Python, та відкритої бібліотеки для створення нейронних мереж для глибокого навчання Keras що працює поверх бібліотеки для машинного навчання Google Tensorflow. Також наводиться необхідна теоретична база для аналізу алгоритмів класифікації, а саме про показники точності роботи алгоритмів та види попередньої обробки даних. Для демонстрації точності натренованих класифікаторів було вирішено використовувати показники точності, чутливості та специфічності, як найбільш прості для розуміння, а для визначення найбільш підходящих параметрів класифікаторів використовувалась зважена точність. Розглянуті відмінності між стандартизацією та minmax і maxabs масштабуванням. Також наведено поверхневий огляд анамблевих методів – беггінгу, бустингу і стакінгу.

Для оцінювання результатів роботи натренованих класифікаторів використовувалась перехресна перевірка з розбиттям на 5 частин.

Перехресна перевірка (англ. cross-validation) — метод оцінювання достовірності математичної моделі з метою перевірки, наскільки результати статистичного аналізу узагальнюються на незалежному наборі даних.

Одноразова перехресна перевірка передбачає розбиття вибірки на взаємодоповнювані під-вибірки з метою проведення аналізу на одній частині (що називається навчальним набором) і перевірки аналізу на іншій частині (що називається контрольним, або тестовим набором). Для зниження дисперсії здійснюється багаторазова перехресна перевірка із застосуванням різних розбиттів, і результати цих перевірок усереднюють.

Матриця помилок (confusion matrix) – у галузі машинного навчання, а конкретно у проблемі статистичної класифікації це специфічний макет таблиці, що дозволяє візуалізувати ефективність алгоритму, як правило, навчання із вчителем. Кожен рядок матриці представляє екземпляри в передбачуваному класі, тоді як кожен стовпець представляє екземпляри фактичного класу (або навпаки). Назва випливає з того, що це дозволяє легко зрозуміти, чи система плутає два класи (тобто зазвичай неправильно позначають один як інший).

Матриця помилок для бінарної класифікації зображена на рис. 1 показує чотири різні результати: вірні позитивні (true positive), хибні позитивні (false positive), вірні негативні (true negative) і хибні негативні (false negative). Фактичні значення формують стовпці, а передбачувані значення (мітки) утворюють рядки.

		Actual	
		Positive	Negative
Predicted	Positive	True Positive	False Positive
	Negative	False Negative	True Negative

Рисунок 1. Загальний вигляд матриці помилок

Для опису різних характеристик класифікатору існує значна кількість показників точності.

Точність (accuracy) визначає частку вірних результатів з усіх.

$$Accuracy = \frac{True\ Positive + True\ Negative}{Positive + Negative}$$

Чутливість (sensitivity, також recall і true positive rate) визначає частку позитивних результатів, які правильно визначені як такі.

$$Sensitivity = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

Специфічність (specificity, також true negative rate) визначає частку негативних результатів, які правильно визначені як такі.

$$Specificity = \frac{True\ Negative}{False\ Positive + True\ Negative}$$

Влучність (precision, також positive predictive value) визначає частку вірних позитивних результатів, з усіх визначених як позитивні.

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

Оцінка F1(F1 score) – це гармонійне середнє значення чутливості та влучності.

$$F_1 = 2 * \frac{Sensitivity * Precision}{Sensitivity + Precision}$$

Зважена точність (balanced accuracy, також Informedness або Youden's J statistic) – визначає частку вірних результатів відповідно кожного класу. У випадку бінарної класифікації визначається формулою:

$$J = Sensitivity + Specificity - 1$$

Для демонстрації точності натренованих класифікаторів було вирішено використовувати показники точності, чутливості та специфічності, як найбільш прості для розуміння.

Оскільки оцінка F1 не бере до уваги розміри класів, а в моєму випадку я мав справу із значно незбалансованим набором даних, для визначення найбільш підходящих параметрів класифікаторів використовувалась зважена точність.

У третьому розділі наведені результати аналізу машинного навчання для різних бібліотек.

Таблиця 2. Показники точності для різних класифікаторів ML.NET

Алгоритм	Accuracy	Sensitivity	Specificity
FastTree	0.69	0.44	0.8
FastForest	0.71	0.25	0.89
LightGBM	0.72	0.47	0.82
GAM	0.69	0.46	0.79
LinearSvm	0.64	0.51	0.70
SGD	0.64	0.71	0.62

При дослідженні ML.Net алгоритм FastForest дав прийнятну точність у визначені хворих пацієнтів але показав велику похибку при визначенні здорових. Алгоритм LightGBM дав найбільшу загальну точність але не кращу при визначені хворих.

Таблиця 3. Показники точності для різних класифікаторів Scikit-learn

Алгоритм	Accuracy	Sensitivity	Specificity
LogisticRegression	0.7	0.22	0.89
Perceptron	0.675	0.56	0.72
GaussianNB	0.55	0.95	0.39
KNeighbors	0.67	0.48	0.74
MLP	0.71	0.28	0.88
SGD	0.69	0.32	0.835
PassiveAggressive	0.69	0.22	0.88
DecisionTree	0.68	0.39	0.79
LinearDiscriminantAnalysis	0.675	0.44	0.77
NearestCentroid	0.6	0.67	0.58
SVC	0.7	0.38	0.83
NuSVC	0.635	0.64	0.635
AdaBoost+ SVC	0.7	0.06	0.96
AdaBoost+ Perceptron	0.685	0.29	0.84
GradientBoosting	0.72	0.4	0.845
XGBoost	0.7	0.4	0.81

При дослідженні Scikit-learn ансамблевий алгоритм Gradient Boosting дав найкращу точність відносно усіх інших, серед більш простих алгоритмів добре себе проявили метод опорних векторів та багатошаровий перцептрон. Загалом усі класифікатори дали недостатню точність для практичного використання.

Keras дав непоганий результат відносно інших, але трохи гірше ніж Gradient Boosting.

```
Accuracy : 0.7167530224525043
Sensitivity : 0.32727272727272727
Specificity : 0.8719806763285024
34.07726550102234
```

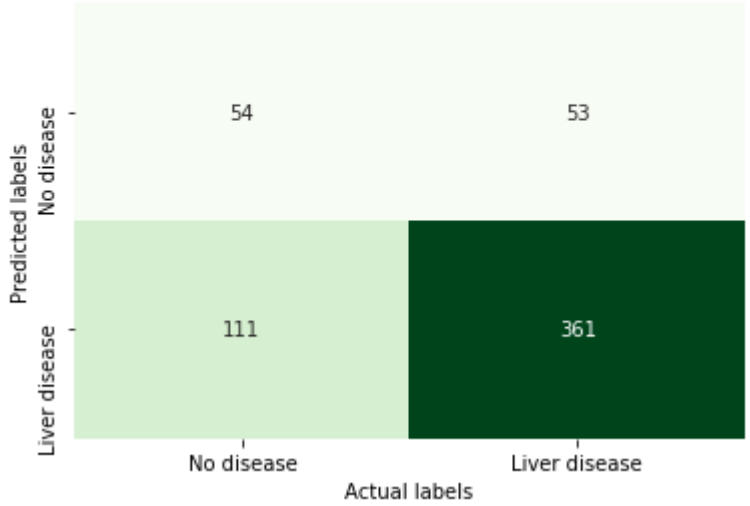

Рис. 3.23. Показники точності Keras

Проаналізувавши результати роботи класифікаторів була виявлена тенденція, відповідно з якою точність визначення хворих пацієнтів значно вища точності визначення здорових пацієнтів. Це можна пояснити значним домінуванням записів про хворих пацієнтів у наборі даних.

Було створено збалансований набір з однаковою кількість даних про хворих та здоровий пацієнтів(по 165) на якому будемо навчати класифікатори, а тестувати на повному наборі.

Таблиця 4. Показники точності для збалансованого набору даних

Алгоритм	Accuracy	Sensitivity	Specificity
KNeighbors	0.74	0.97	0.65
MLP	0.63	0.92	0.51
DecisionTree	0.66	0.8	0.6
AdaBoost+ SVC	0.58	0.91	0.45
AdaBoost+ DecisionTree	0.69	0.79	0.65
GradientBoosting	0.7	0.68	0.71
XGBoost	0.67	0.81	0.57

При однаковій кількості записів обох класів у наборі даних, ансамблеві методи на основі дерев рішень показують в середньому кращі результати. Алгоритм KNeighbors показав найкращу загальну точність з усіх тестів, але погано впорався з визначенням хворих, що є пріоритетом. Алгоритм GradientBoosting дав точність недостатню для точної діагностики, але прийнятну для автоматичного моніторингу, оскільки з 70% точністю правильно ідентифікує як хворих так і здорових, і цей результат можна вважати найкращим з усіх отриманих.

У четвертому розділі дано опис безкоштовного хмарного сервісу від Google Colab заснованого на Jupyter Notebooks що дає можливість створювати та ділитись блокнотами що складаються з блоків із скриптами на мові Python та створювати прості основні елементи графічного інтерфейсу Forms, що дають можливість зручної взаємодії із користувачем. Також наведені результати створеного простого додаток що приймає ввід користувача, і використовуючи найкращий натренований класифікатор повертає стан здоров’я пацієнта і вірогідність цього стану.

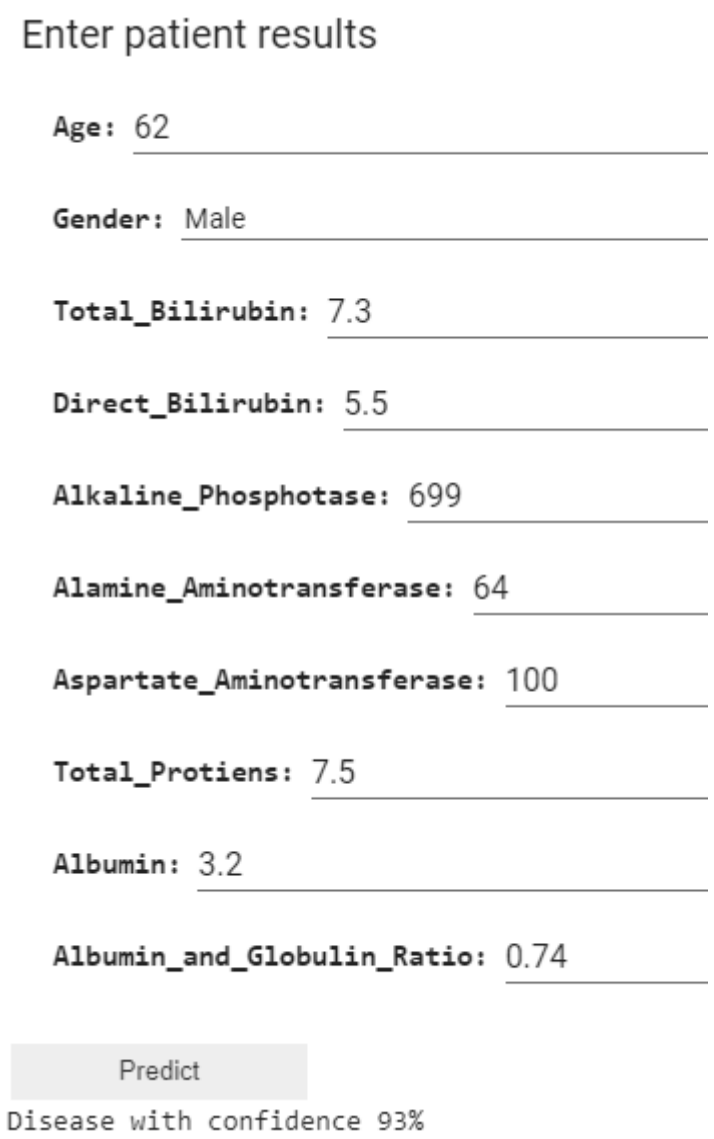

Рис. 3. Приклад роботи додатку

У методичній частині розроблено практичну роботу на тему «Класифікація з використанням бібліотеки Scikit-learn».

У спеціальній частині магістерської наукової роботи з «Охорони праці та безпеки життєдіяльності» описано стан робочого приміщення і робочих місць у ТОВ «ГлобалЛоджик Україна».В розділі виконано розрахунок загального рівномірного освітлення виробничого приміщення люмінесцентними лампами методом коефіцієнта використання світлового потоку. Встановлено 6 світильників із загальною кількістю ламп 24 та потужністю 40Вт кожна. Загальна потужність передбаченого штучного освітлення складає 9200Вт. Також приведений інструктаж з техніки безпеки при виникненні пожежі на підприємстві згідно чинних нормативно-правових актів з техніки безпеки та цивільного захисту.

ЗАГАЛЬНІ ВИСНОВКИ

Ця робота посвячена вивченню алгоритмів машинного навчання для бінарної класифікації що можуть бути використані при створенні автоматизованих систем діагностики захворювань печінки.

У першому розділі дано загальний опис захворювань печінки і важливості їх попереднього діагностування, видів машинного навчання (з вчителем, без вчителя, із підкріпленням), проаналізовані наявні дослідження по темі, обраний набір даних Indian Liver Patient Dataset з результатами біохімічний аналізів 583 індійських пацієнтів по 11 атрибутів. Також були сформовані кроки потрібні для проведення дослідження : вивчення вивчення і вибір доступних бібліотек машинного навчання, тестування різних алгоритмів машинного навчання, створення програмного додатку із використанням кращого алгоритму.

У другому розділі дано опис бібліотекам машинного навчання що були обрані для аналізу: відносна нова бібліотека Microsoft ML.Net для вирішення задач машинного навчання за допомогою C#,відкрита бібліотека Skikit-learn на мові Python, та відкритої бібліотеки для створення нейронних мереж для глибокого навчання Keras що працює поверх бібліотеки для машинного навчання Google Tensorflow. Також наводиться необхідна теоретична база для аналізу алгоритмів класифікації, а саме про показники точності роботи алгоритмів та види попередньої обробки даних. Для демонстрації точності натренованих класифікаторів було вирішено використовувати показники точності, чутливості та специфічності, як найбільш прості для розуміння, а для визначення найбільш підходящих параметрів класифікаторів використовувалась зважена точність. Розглянуті відмінності між стандартизацією та minmax і maxabs масштабуванням. Також наведено поверхневий огляд анамблевих методів – беггінгу, бустингу і стакінгу.

У третьому розділі наведені результати аналізу машинного навчання для різних бібліотек. При дослідженні ML.Net алгоритм FastForest дав кращу точність у визначені хворих пацієнтів але показав велику похибку при визначенні здорових. Алгоритм LightGBM дав найбільшу загальну точність але не кращу при визначені хворих. При дослідженні Scikit-learn ансамблевий алгоритм Gradient Boosting дав

найкращу точність відносно усіх інших, серед більш простих алгоритмів добре себе проявили метод опорних векторів та багатошаровий перцептрон. Загалом усі класифікатори дали недостатню точність для практичного використання.Keras дав непоганий результат відносно інших, але трохи гірше ніж Gradient Boosting.

При однаковій кількості записів обох класів у наборі даних, ансамблеві методи на основі дерев рішень показують в середньому кращі результати. Алгоритм KNeighbors показав найкращу загальну точність з усіх тестів, але погано впорався з визначенням хворих, що є пріоритетом. Алгоритм GradientBoosting дав точність недостатню для точної діагностики, але прийнятну для автоматичного моніторингу, оскільки з 70% точністю правильно ідентифікує як хворих так і здорових, і цей результат можна вважати найкращим з усіх отриманих.

У четвертому розділі дано опис безкоштовного хмарного сервісу від Google Colab заснований на Jupyter Notebooks що дає можливість створювати та ділитись блокнотами що складаються з блоків із скриптами на мові Python та створювати прості основні елементи графічного інтерфейсу Forms, що дають можливість зручної взаємодії із користувачем. Також наведені результати створеного простого застосунку, що приймає ввід користувача, і використовуючи найкращий натренований класифікатор повертає стан здоров’я пацієнта і вірогідність цього стану.

У методичній частині розроблено практичну роботу на тему «Ознайомлення з алгоритмами класифікації з бібліотеки для машинного навчання Scikit-learn».

У спеціальній частині магістерської наукової роботи з «Охорони праці та безпеки життєдіяльності» описано стан робочого приміщення і робочих місць у ТОВ «ГлобалЛоджик Україна».В розділі виконано розрахунок загального рівномірного освітлення виробничого приміщення люмінесцентними лампами методом коефіцієнта використання світлового потоку. Встановлено 6 світильників із загальною кількістю ламп 24 та потужністю 40Вт кожна. Загальна потужність передбаченого штучного освітлення складає 9200Вт.

Також приведений інструктаж з техніки безпеки при виникненні пожежі на підприємстві згідно чинних нормативно-правових актів з техніки безпеки та цивільного захисту.

АНОТАЦІЯ

Соколюк Антон Вікторович. Дослідження методів машинного навчання для бінарної ідентифікації захворювань печінки.

– На правах рукопису.

Магістерська наукова робота на здобуття освітньої кваліфікації «Магістр комп'ютерних наук». – Чорноморський національний університет імені Петра Могили, Миколаїв, 2020.

Дана магістерська наукова робота присвячена методів машинного навчання для бінарної ідентифікації захворювань печінки.

Метою магістерської наукової роботи є полегшення праці лікарів шляхом створення автоматизованих систем діагностування захворювань печінки.

Об'єктом дослідження є сфера машинного навчання для задач медичної діагностики.

Предметом дослідження є методи машинного навчання для бінарної ідентифікації захворювань печінки.

Фахова частина магістерської наукової роботи складається з наступних розділів: дослідження предметної області; опис технологій; аналіз методів машинного навчання; створення програмного застосунку.

Задачі, які були виконані в процесі роботи:

- вивчення теоретичної бази та аналіз попередніх досліджень на пов'язані теми;
- вивчення і вибір доступних бібліотек машинного навчання і теорії для їх використання;
- тестування різних алгоритмів машинного навчання для бінарної класифікації;
- створення програмного додатку для практичного використання найбільш підходящого алгоритму.

У методичній частині розроблено практичну роботу на тему «Ознайомлення з алгоритмами класифікації з бібліотеки для машинного навчання Scikit-learn».

У спеціальній частині магістерської наукової роботи з «Охорони праці та безпеки життєдіяльності» описано стан робочого приміщення і робочих місць у ТОВ «ГлобалЛоджик Україна». В розділі виконано розрахунок загального рівномірного освітлення виробничого приміщення люмінесцентними лампами методом коефіцієнта використання світлового потоку.Також приведений інструктаж з техніки безпеки при виникненні пожежі на підприємстві згідно чинних нормативно-правових актів з техніки безпеки та цивільного захисту.

Робота складається з 88 сторінок, 40 рисунків, 5 таблиць та 18 посилань на літературні джерела.

Ключові слова: *машинне навчання, класифікація, захворювання печінки, медицина.*

ABSTRACT

Sokoliuk Anton. Study of machine learning algorithms for binary classification of liver disease– On the rights of the manuscript.

Master's scientific work for obtaining an educational qualification "Master of Computer Science". – Petro Mohyla Black Sea National University, Mykolaiv, 2020.

The master’s research paper is devoted to research of machine learning methods for binary classification of liver sickness.

The purpose of this paper is lowering burden on doctor by developing automatized system of liver disease diagnostic.

The object is the field of machine learning for the tasks of medical diagnostics.

The subject of research is machine learning algorithms for binary classification of liver disease.

The professional part of master's research paper consists of the following sections: research of the subject area; description of technologies; analysis of machine learning algorithms; designing software application.

Tasks that were completed during the process:

- study of theoretical basis and analysis of previous research on related topics;
- study and selection of available machine learning libraries and theories for their use;
- testing different machine learning algorithms for binary classification;
- creating a software application for practical use of the most suitable algorithm.

In the methodical part practical work on the course of machine learning basis on the theme "Acquaintance with the classification algorithms from the Scikit-learn machine learning library" was developed.

In the special part of the master's research paper "Occupational safety and security in emergency situations" an analysis of the working conditions at workplaces in GlobalLogic Ukraine LLC is described. The section calculates the total uniform illumination of the production room with fluorescent lamps by the method of the coefficient of use of light flow. Safety instruction in case of fire at the enterprise in accordance with the current regulations on safety and civil protection is also given.

The work consists of 88 pages, 40 figures, 5 tables and 18 references to literary sources.

Keywords: machine learning, classification, liver disease, medicine.