|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Лабораторна робота №2**  **ПОРІВНЯННЯ МЕТОДІВ КЛАСИФІКАЦІЇ ДАНИХ**  ***Мета роботи:*** використовуючи спеціалізовані бібліотеки та мову програмування Python дослідити різні методи класифікації даних та навчитися їх порівнювати  **Завдання №2.1 – 2.14:**  Опис набору вхідних даних:   1. Age (вік): 17 – 90 років 2. Workclass (зайнятість): Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked 3. Fnlwgt (Final weight, ваговий коефіцієнт): числове значення 4. Education (освіта): Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc- acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool 5. Education-num (числове преставлення рівня освіти): числове значення 6. Marital-status (сімейний стан): Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse 7. Occupation (професія): Tech-support, Craft-repair, Other-service, Sales, Exec- managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed- Forces 8. Relationship (відносити): Wife, Own-child, Husband, Not-in-family, Other- relative, Unmarried 9. Race (раса): White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black. 10. Sex (стать): Female, Male | | | | | | | | | | |
|  |  |  |  |  | *ДУ «Житомирська політехніка».23.121.4.000 – Лр2* | | | | | |
|  |  |  |  |  |
| *Змн.* | *Арк.* | *№ докум.* | *Підпис* | *Дата* |
| *Розроб.* | | *Решетнюк Д.Р.* |  |  | Звіт з лабораторної роботи | *Літ.* | | | *Арк.* | *Аркушів* |
| *Перевір.* | | *Голенко М.Ю.* |  |  |  |  |  | *1* | *29* |
| *Керівник* | |  |  |  | *ФІКТ Гр. IПЗ-20-2[1]* | | | | |
| *Н. контр.* | |  |  |  |
| *Зав. каф.* | |  |  |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 1. Capital-gain (капіталовигода): числове значення 2. Capital-loss (капіталовтрата): числове значення 3. Hours-per-week (години на тиждень): числове значення 4. Native-country (Батьківщина): United-States, Cambodia, England, Puerto- Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand- Netherlands. 5. Клас доходу: >=50К чи <= 50К.   **2.1:**    Рис. 1. Код програми | | | | | | |
|  |  | *Решетнюк Д.Р.* |  |  | *ДУ «Житомирська політехніка».23.121.4.000 – Лр2* | *Арк.* |
|  |  | *Голенко М.Ю.* |  |  | *2* |
| *Змн.* | *Арк.* | *№ докум.* | *Підпис* | *Дата* |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Рис. 2. Код програми | | | | | | |
|  |  | *Решетнюк Д.Р.* |  |  | *ДУ «Житомирська політехніка».23.121.4.000 – Лр2* | *Арк.* |
|  |  | *Голенко М.Ю.* |  |  | *3* |
| *Змн.* | *Арк.* | *№ докум.* | *Підпис* | *Дата* |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Рис. 3. Код програми    Рис 4. Результат виконання Тестова точка належить до класу ‘<=50K’ | | | | | | |
|  |  | *Решетнюк Д.Р.* |  |  | *ДУ «Житомирська політехніка».23.121.4.000 – Лр2* | *Арк.* |
|  |  | *Голенко М.Ю.* |  |  | *4* |
| *Змн.* | *Арк.* | *№ докум.* | *Підпис* | *Дата* |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **2.2 (поліномінальне ядро):**    Рис 5. Код програми | | | | | | |
|  |  | *Решетнюк Д.Р..* |  |  | *ДУ «Житомирська політехніка».23.121.4.000 – Лр2* | *Арк.* |
|  |  | *Голенко М.Ю.* |  |  | *5* |
| *Змн.* | *Арк.* | *№ докум.* | *Підпис* | *Дата* |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Рис 6. Код програми | | | | | | |
|  |  | *Решетнюк Д.Р.* |  |  | *ДУ «Житомирська політехніка».23.121.4.000 – Лр2* | *Арк.* |
|  |  | *Голенко М.Ю.* |  |  | *6* |
| *Змн.* | *Арк.* | *№ докум.* | *Підпис* | *Дата* |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Рис 7. Код програми    Рис 8. Результат виконання  Degree = 2, адже навіть зміна пріоритетності задачі в диспетчері не допома- гала виконати код. Не вистачає потужності пристрою. | | | | | | |
|  |  | *Решетнюк Д.Р.* |  |  | *ДУ «Житомирська політехніка».23.121.4.000 – Лр2* | *Арк.* |
|  |  | *Голенко М.Ю.* |  |  | *7* |
| *Змн.* | *Арк.* | *№ докум.* | *Підпис* | *Дата* |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **2.2 (гаусове ядро):**    Рис. 9. Код програми | | | | | | |
|  |  | *Решетнюк Д.Р..* |  |  | *ДУ «Житомирська політехніка».23.121.4.000 – Лр2* | *Арк.* |
|  |  | *Голенко М.Ю.* |  |  | *8* |
| *Змн.* | *Арк.* | *№ докум.* | *Підпис* | *Дата* |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Рис. 10. Код програми | | | | | | |
|  |  | *Решетнюк Д.Р.* |  |  | *ДУ «Житомирська політехніка».23.121.4.000 – Лр2* | *Арк.* |
|  |  | *Голенко М.Ю.* |  |  | *9* |
| *Змн.* | *Арк.* | *№ докум.* | *Підпис* | *Дата* |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Рис. 12. Результат виконання  **2.3 (сигмоїдальне ядро):**    Рис. 13. Код програми | | | | | | |
|  |  | *Решетнюк Д.Р.* |  |  | *ДУ «Житомирська політехніка».23.121.4.000 – Лр2* | *Арк.* |
|  |  | *Голенко М.Ю.* |  |  | *10* |
| *Змн.* | *Арк.* | *№ докум.* | *Підпис* | *Дата* |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Рис. 14. Код програми | | | | | | |
|  |  | *Решетнюк Д.Р.* |  |  | *ДУ «Житомирська політехніка».23.121.4.000 – Лр2* | *Арк.* |
|  |  | *Голенко М.Ю.* |  |  | *11* |
| *Змн.* | *Арк.* | *№ докум.* | *Підпис* | *Дата* |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Рис. 16. Результат виконання  Лінійне ядро має високі показники точності та recall, але найгіршу F1 score, що може вказувати на менш точне розділення класів.  У поліномінального ядра усі показники доволі високі, але менші за показники гаусового ядра.  Найкращі показники має гаусове ядро, хоча йому потрібно більше часу на навчання та налаштування параметрів.  Сигмоїдальне ядро має помірні показники, але вони нижчі за показники поліномінального та гаусового ядра.  В контексті даної задачі рекомендовано використовувати гаусове ядро, оскільки воно, хоча і вимагає більше часу, дає найкращі результати та ефективно справляється з класифікацією. | | | | | | |
|  |  | *Решетнюк Д.Р.* |  |  | *ДУ «Житомирська політехніка».23.121.4.000 – Лр2* | *Арк.* |
|  |  | *Голенко М.Ю.* |  |  | *12* |
| *Змн.* | *Арк.* | *№ докум.* | *Підпис* | *Дата* |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Завдання 2.3.**  **Крок 1:**    Рис. 17. Код програми (вивчення даних) | | | | | | |
|  |  | *Решетнюк Д.Р.* |  |  | *ДУ «Житомирська політехніка».23.121.4.000 – Лр2* | *Арк.* |
|  |  | *Голенко М.Ю.* |  |  | *13* |
| *Змн.* | *Арк.* | *№ докум.* | *Підпис* | *Дата* |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Рис. 18. Результат виконання (вивчення даних) | | | | | | |
|  |  | *Решетнюк Д.Р..* |  |  | *ДУ «Житомирська політехніка».23.121.4.000 – Лр2* | *Арк.* |
|  |  | *Голенко М.Ю.* |  |  | *14* |
| *Змн.* | *Арк.* | *№ докум.* | *Підпис* | *Дата* |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Крок 2:**    Рис. 19. Код програми | | | | | | |
|  |  | *Решетнюк Д.Р.* |  |  | *ДУ «Житомирська політехніка».23.121.4.000 – Лр2* | *Арк.* |
|  |  | *Голенко М.Ю.* |  |  | *15* |
| *Змн.* | *Арк.* | *№ докум.* | *Підпис* | *Дата* |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Рис. 20. Результат виконання    Рис. 21. Діаграма розмаху для ознак | | | | | | |
|  |  | *Решетнюк Д.Р.* |  |  | *ДУ «Житомирська політехніка».23.121.4.000 – Лр2* | *Арк.* |
|  |  | *Голенко М.Ю.* |  |  | *16* |
| *Змн.* | *Арк.* | *№ докум.* | *Підпис* | *Дата* |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Рис. 22. Гістограма вхідних даних кожної ознаки    Рис. 23. Матриця діаграм розсіювання | | | | | | |
|  |  | *Решетнюк Д.Р.* |  |  | *ДУ «Житомирська політехніка».23.121.4.000 – Лр2* | *Арк.* |
|  |  | *Голенко М.Ю.* |  |  | *17* |
| *Змн.* | *Арк.* | *№ докум.* | *Підпис* | *Дата* |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Крок 3 і далі:**    Рис. 23. Код програми | | | | | | |
|  |  | *Решетнюк Д.Р.* |  |  | *ДУ «Житомирська політехніка».23.121.4.000 – Лр2* | *Арк.* |
|  |  | *Голенко М.Ю.* |  |  | *18* |
| *Змн.* | *Арк.* | *№ докум.* | *Підпис* | *Дата* |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Рис. 24. Код програми    Рис. 25. Результат виконання | | | | | | |
|  |  | *Решетнюк Д.Р..* |  |  | *ДУ «Житомирська політехніка».23.121.4.000 – Лр2* | *Арк.* |
|  |  | *Голенко М.Ю.* |  |  | *19* |
| *Змн.* | *Арк.* | *№ докум.* | *Підпис* | *Дата* |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Рис. 26. Порівняння алгоритмів  Під час виконання завдання вдалося досягти точності приблизно 96.67%. Це означає, що 96.67% даних контрольної вибірки було правильно класифіковано.  Щодо матриці помилок, виявлено, що в рядку 2, стовпчику 3 був приклад, який було неправильно класифіковано. Він мав належати до 2-го класу, але був призначений до 3-го.  Алгоритм SVM найкраще впорався із завданням, маючи точність 0.9833 та стандартне відхилення 0.0333. Це свідчить про високу ефективність моделі. Квітка належить до класу Iris-setosa. | | | | | | |
|  |  | *Решетнюк Д.Р.* |  |  | *ДУ «Житомирська політехніка».23.121.4.000 – Лр2* | *Арк.* |
|  |  | *Голенко М.Ю.* |  |  | *20* |
| *Змн.* | *Арк.* | *№ докум.* | *Підпис* | *Дата* |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Завдання 2.4.**    Рис. 26. Код програми | | | | | | |
|  |  | *Решетнюк Д.Р.* |  |  | *ДУ «Житомирська політехніка».23.121.4.000 – Лр2* | *Арк.* |
|  |  | *Голенко М.Ю.* |  |  | *21* |
| *Змн.* | *Арк.* | *№ докум.* | *Підпис* | *Дата* |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Рис. 27. Код програми | | | | | | |
|  |  | *Решетнюк Д.Р.* |  |  | *ДУ «Житомирська політехніка».23.121.4.000 – Лр2* | *Арк.* |
|  |  | *Голенко М.Ю.* |  |  | *22* |
| *Змн.* | *Арк.* | *№ докум.* | *Підпис* | *Дата* |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Рис. 28. Код програми | | | | | | |
|  |  | *Решетнюк Д.Р.* |  |  | *ДУ «Житомирська політехніка».23.121.4.000 – Лр2* | *Арк.* |
|  |  | *Голенко М.Ю.* |  |  | *23* |
| *Змн.* | *Арк.* | *№ докум.* | *Підпис* | *Дата* |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Рис. 29. Результат виконання    Рис. 30. Порівняння accuracy | | | | | | |
|  |  | *Решетнюк Д.Р.* |  |  | *ДУ «Житомирська політехніка».23.121.4.000 – Лр2* | *Арк.* |
|  |  | *Голенко М.Ю.* |  |  | *24* |
| *Змн.* | *Арк.* | *№ докум.* | *Підпис* | *Дата* |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Рис. 31. Порівняння f1 score    Рис. 32. Порівняння recall | | | | | | |
|  |  | *Решетнюк Д.Р.* |  |  | *ДУ «Житомирська політехніка».23.121.4.000 – Лр2* | *Арк.* |
|  |  | *Голенко М.Ю.* |  |  | *25* |
| *Змн.* | *Арк.* | *№ докум.* | *Підпис* | *Дата* |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Рис. 33. Порівняння precision  Моделі LDA та CART показали найкращі результати для даного завдання. Для вибору оптимальної моделі важливо проаналізувати отримані показники якості.  На перший погляд, CART виглядає кращою, оскільки має вищі значення F1 score та recall. Проте важливо врахувати, що CART більше спрямований на класифікацію позитивних випадків. У даному завданні важлива рівновага між класами.  LDA має вищі показники accuracy та precision і робить менше помилок у класифікації позитивних випадків, ніж CART. Вона добре збалансована між recall та accuracy. Тому для даної задачі LDA є кращим вибором, оскільки забезпечує потрібну рівновагу у класифікації | | | | | | |
|  |  | *Решетнюк Д.Р.* |  |  | *ДУ «Житомирська політехніка».23.121.4.000 – Лр2* | *Арк.* |
|  |  | *Голенко М.Ю.* |  |  | *26* |
| *Змн.* | *Арк.* | *№ докум.* | *Підпис* | *Дата* |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Завдання 2.5.**    Рис. 34. Код програми | | | | | | |
|  |  | *Решетнюк Д.Р.* |  |  | *ДУ «Житомирська політехніка».23.121.4.000 – Лр2* | *Арк.* |
|  |  | *Голенко М.Ю.* |  |  | *27* |
| *Змн.* | *Арк.* | *№ докум.* | *Підпис* | *Дата* |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Рис. 35. Результат виконання  Модель Ridge з параметрами tol=1e-2 та solver="sag" досягає 75.56% точності зі збалансованими значеннями precision та recall (відповідно 83.33% та 75.56%). F1 score складає 75.03%, а Cohen Kappa score - 0.6431, що свідчить про помірну згоду між прогнозами моделі та реальними даними. | | | | | | |
|  |  | *Решетнюк Д.Р.* |  |  | *ДУ «Житомирська політехніка».23.121.4.000 – Лр2* | *Арк.* |
|  |  | *Голенко М.Ю.* |  |  | *28* |
| *Змн.* | *Арк.* | *№ докум.* | *Підпис* | *Дата* |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Рис. 36. Confusion.jpg  3. Ця матриця відображає кількість правильно та неправильно класифікованих екземплярів для кожного класу. Зі значеннями на діагоналі відображається кількість правильних класифікацій, допомагаючи в оцінці ефективності моделі для конкретних класів.  https://github.com/cpacemon/ai\_Lab\_reshetnyuk | | | | | | |
|  |  | *Решетнюк Д.Р..* |  |  | *ДУ «Житомирська політехніка».23.121.4.000 – Лр2* | *Арк.* |
|  |  | *Голенко М.Ю.* |  |  | *29* |
| *Змн.* | *Арк.* | *№ докум.* | *Підпис* | *Дата* |