Визначення кардіоваскулярних захворювань методами машинного навчання

Студент:
Гірянський Богдан
Дип. керівник :
Харченко К.В.

Мета

Дослідити сучасні методи машинного навчання та визначити, які з них найбільше підходять для діагностування кардіоваскулярних захворювання.

Актуальність

ССЗ - це перша причина смерті в усьому світі, щорічно від ССЗ помирає більше людей, ніж від будь-якої іншої причини.

За оцінками, 17,3 мільйона людей померли від ССЗ у 2008 році, що становить 30% усіх смертей у світі. З цих смертей 7,3 мільйонів були наслідком ішемічної хвороби серця, 6,2 мільйона - через інсульт.

Країни з низьким і середнім рівнем доходу зазначають непропорційний вплив: понад 80% смертей від ССЗ, мають місце в країнах з низьким та середнім рівнем доходу і майже однаковою мірою зустрічаються серед чоловіків та жінок.

Етапи які були пройдені при виконанні роботи

1. Аналіз області визначення кардіоваскулярних захворювань за допомогою методів машинного навчання.
2. Аналіз методів: система опорних векторів, наївного Байаса, логістичної регресії, дерева рішень, нейронної мережі.
3. Створені моделі методами машинного навчання визначеними для даної роботи.
4. Визначено методи машинного навчання, які дали найкращі результати у діагностуванні кардіоваскулярних захворювань.

Використані технології

- Python 3.7 - мова програмування, яка легка для розуміння і містить велику кількість реалізованих методів.
- Google Colab - безкоштовний хмарний сервіс на основі Jupyter Notebook. Google Colab надає все необхідне для машинного навчання прямо в браузері, дає безкоштовний доступ до неймовірно швидким GPU і TPU.
- SKlearn - зібрані ефективні інструменти для прогнозування даних. Вони доступні для всіх і вільно можуть використання в різних задачах.

Дані

Для вирішення задачі визначення кардіоваскулярних захворювань було вибрано три набори даних, які були у вільному доступі, у яких містилися характеристики параметрів людини та мітка наявності захворювання.

Перший набір даних містить наступні характеристики: вік, зріст, стать, верхній артеріальний тиск, нижній артеріальний тиск, рівень холестерину, рівень глюкози, паління, вживання алкоголю, фізична активність, відсутність або наявність захворювання.

Дані

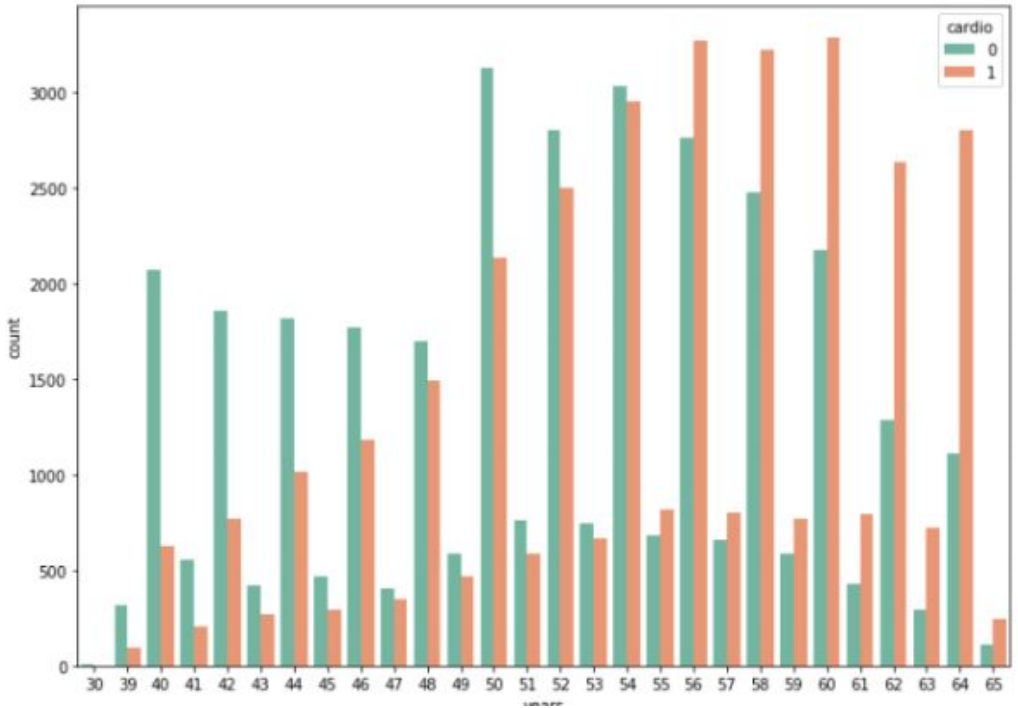
У другому та третьому наборі даних характеристики співпадають: вік, стать, рівень болю в грудях, кров’яний тиск у стані спокою, рівень холестерину, результат електрокардіограми у стані спокою, максимальна частота ударів серця, чи потрібно виконувати вправи через ангіну, результати стрес-тесту,нахил найвищого сегменту стрес-тесту, кількість основних судин (0-3), забарвлених флоросопією.

Дані

Номер набору даних	Кількість характеристик	Кількість записів
1	12 + статус захворювання	70000
2	13 + статус захворювання	270
3	13 + статус захворювання	303

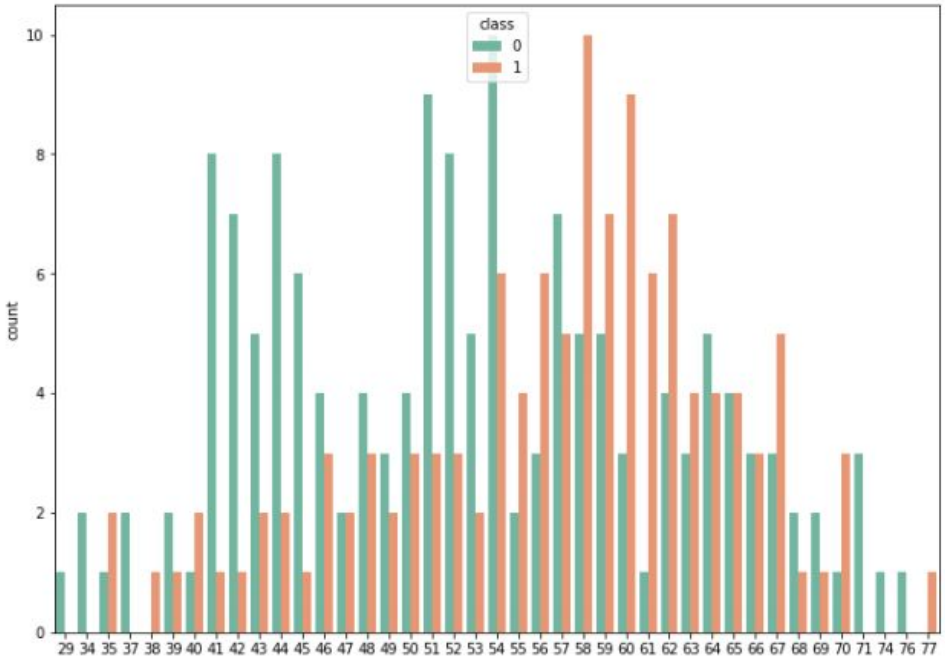
У першому наборі даних залишилося після предобробки 68983 записи, у другому наборі 260 та у третьому 292.

Розподіл захворювання по віку



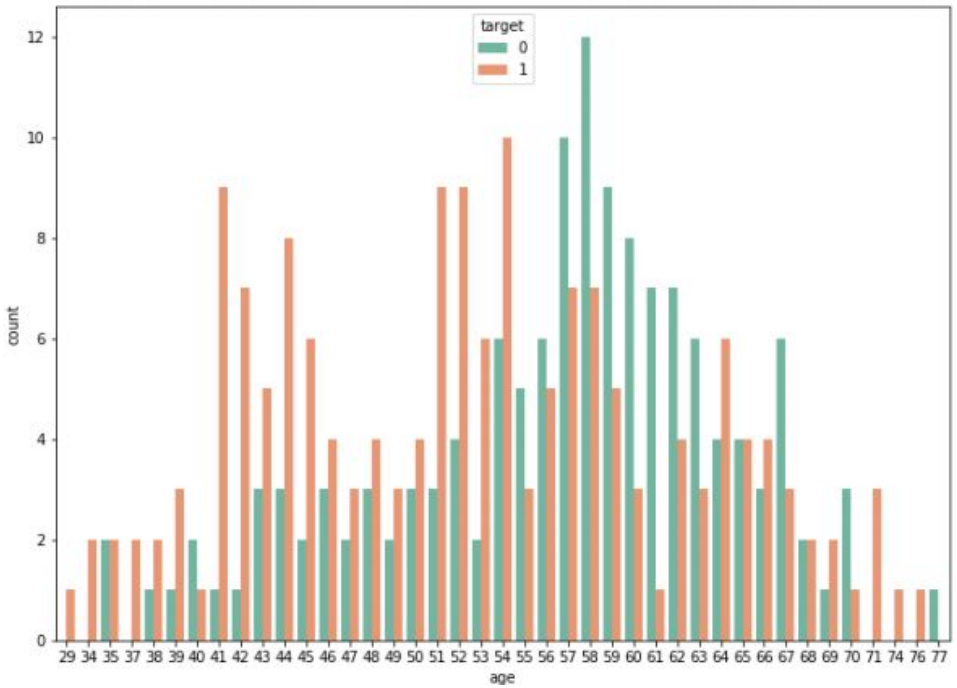
Набір даних №1

Розподіл захворювання по віку



Набір даних №2

Розподіл захворювання по віку



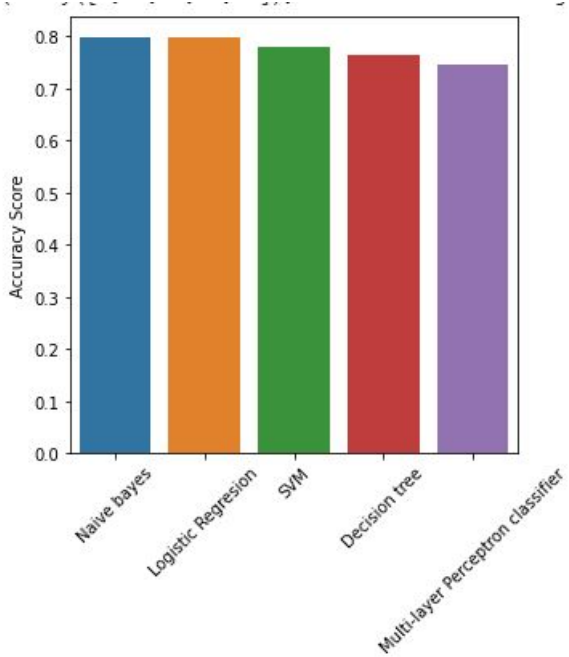
Набір даних №3

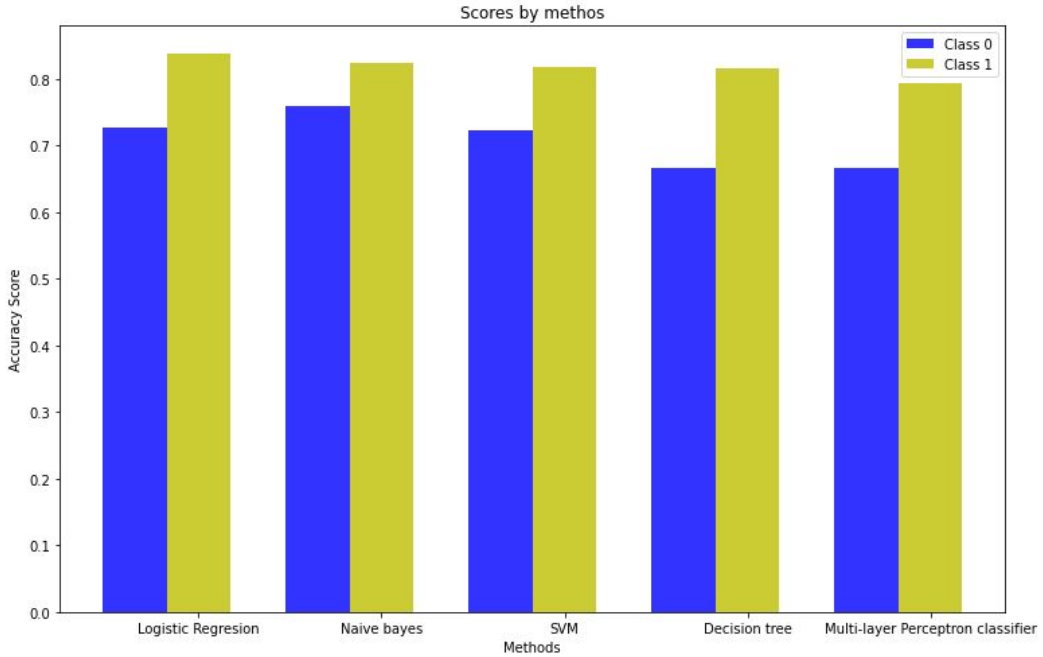
Результати роботи методів машинного навчання

	Naive bayes	Logistic Regresion	SVM	Decision tree	Multi-layer Perceptron classifier
Accuracy Score	0.7966	0.7966	0.7797	0.7627	0.7458
Class 0 Precision	0.7600	0.8421	0.7727	0.8235	0.7500
Class 1 Precision	0.8235	0.7750	0.7838	0.7381	0.7436
Class 0 F1-score	0.7600	0.7273	0.7234	0.6667	0.6667
Class 1 F1-score	0.8235	0.8378	0.8169	0.8158	0.7945
Class 0 Recall	0.7600	0.6400	0.6800	0.5600	0.6000
Class 1 Recall	0.8235	0.9118	0.8529	0.9118	0.8529
Class 0 Support	25.0000	25.0000	25.0000	25.0000	25.0000
Class 1 Support	34.0000	34.0000	34.0000	34.0000	34.0000
Studying time	0.0005	0.0073	0.0032	0.0008	1.8374
Testing time	0.0004	0.0005	0.0006	0.0004	0.0009
Time sum	0.0009	0.0077	0.0038	0.0012	1.8383

Результати навчених моделей, набір даних №3

Результати роботи методів машинного навчання





Діаграма результати роботи моделей за точністю, набір даних №3

Результати роботи методів машинного навчання

Виключаний GPU прискорювач

	Studying time	Testing time	Time sum
Multi-layer Perceptron classifier	1.837383	0.000889	1.838272
Logistic Regresion	0.007272	0.000474	0.007746
SVM	0.003206	0.000595	0.003800
Decision tree	0.000753	0.000411	0.001165
Naive bayes	0.000505	0.000411	0.000915

Включаний GPU прискорювач

	Studying time	Testing time	Time sum
Multi-layer Perceptron classifier	1.527838	0.000777	1.528616
Logistic Regresion	0.005477	0.000314	0.005792
SVM	0.002258	0.000467	0.002725
Decision tree	0.000633	0.000292	0.000925
Naive bayes	0.000408	0.000350	0.000758

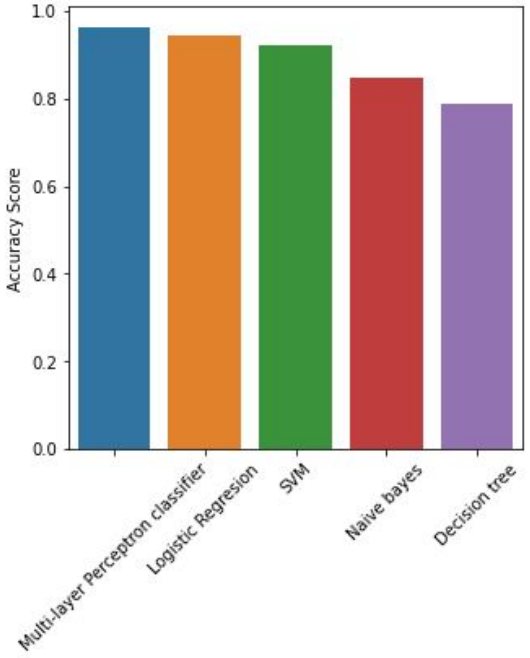
Діаграма результати роботи моделей за часом навчання, набір даних №3

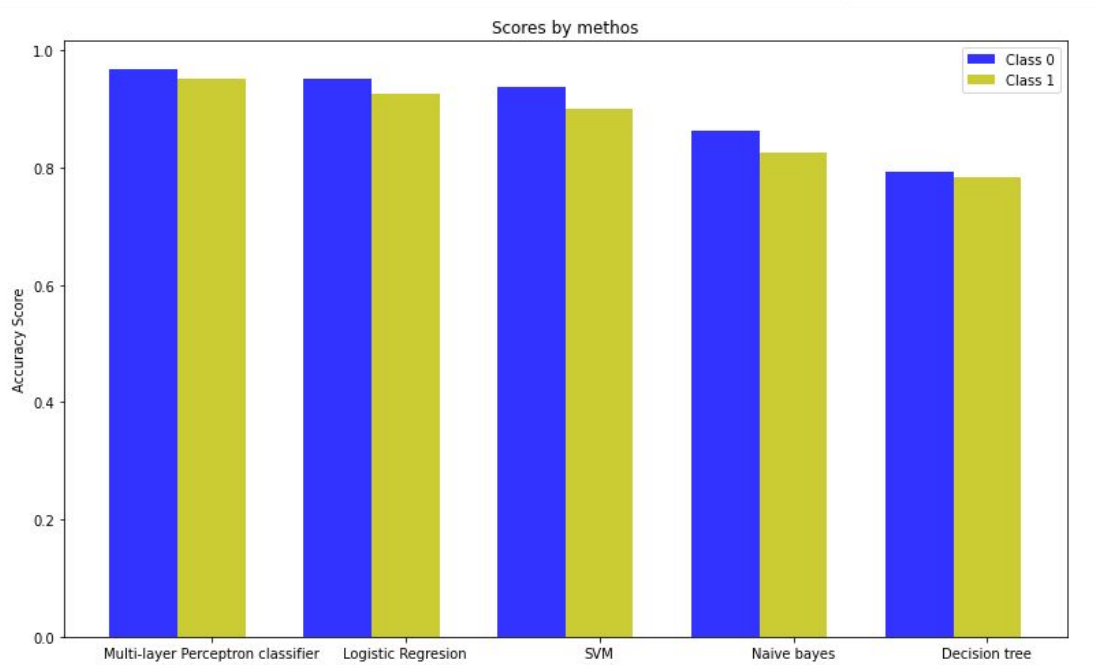
Результати роботи методів машинного навчання

	Multi-layer Perceptron classifier	Logistic Regresion	SVM	Naive bayes	Decision tree
Accuracy Score	0.9615	0.9423	0.9231	0.8462	0.7885
Class 0 Precision	0.9677	0.9375	0.9091	0.9259	0.9545
Class 1 Precision	0.9524	0.9500	0.9474	0.7600	0.6667
Class 0 F1-score	0.9677	0.9524	0.9375	0.8621	0.7925
Class 1 F1-score	0.9524	0.9268	0.9000	0.8261	0.7843
Class 0 Recall	0.9677	0.9677	0.9677	0.8065	0.6774
Class 1 Recall	0.9524	0.9048	0.8571	0.9048	0.9524
Class 0 Support	31.0000	31.0000	31.0000	31.0000	31.0000
Class 1 Support	21.0000	21.0000	21.0000	21.0000	21.0000
Studying time	0.1172	0.0926	0.0105	0.0005	0.0013
Testing time	0.0007	0.0009	0.0007	0.0004	0.0003
Time sum	0.1179	0.0935	0.0112	0.0009	0.0017

Результати навчених моделей, набір даних №2

Результати роботи методів машинного навчання





Діаграма результати роботи моделей за точністю, набір даних №2

Результати роботи методів машинного навчання

Виключаний GPU прискорювач

	Studying time	Testing time	Time sum
Multi-layer Perceptron classifier	0.161555	0.000860	0.162416
Logistic Regresion	0.054259	0.000906	0.055164
SVM	0.010366	0.000619	0.010985
Decision tree	0.001343	0.000447	0.001790
Naive bayes	0.000643	0.000591	0.001234

Включаний GPU прискорювач

	Studying time	Testing time	Time sum
Multi-layer Perceptron classifier	0.125666	0.000675	0.126342
Logistic Regresion	0.083000	0.001064	0.084064
SVM	0.010203	0.000832	0.011035
Decision tree	0.001293	0.000421	0.001714
Naive bayes	0.000468	0.000404	0.000872

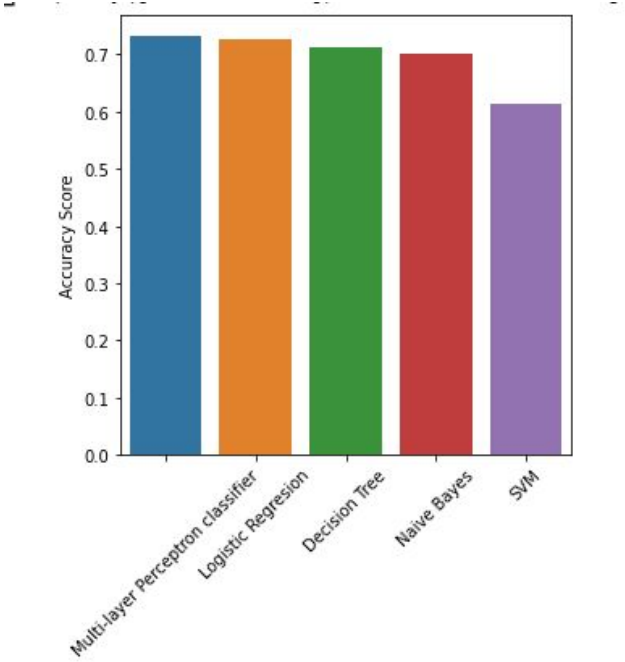
Діаграма результати роботи моделей за часом навчання, набір даних №2

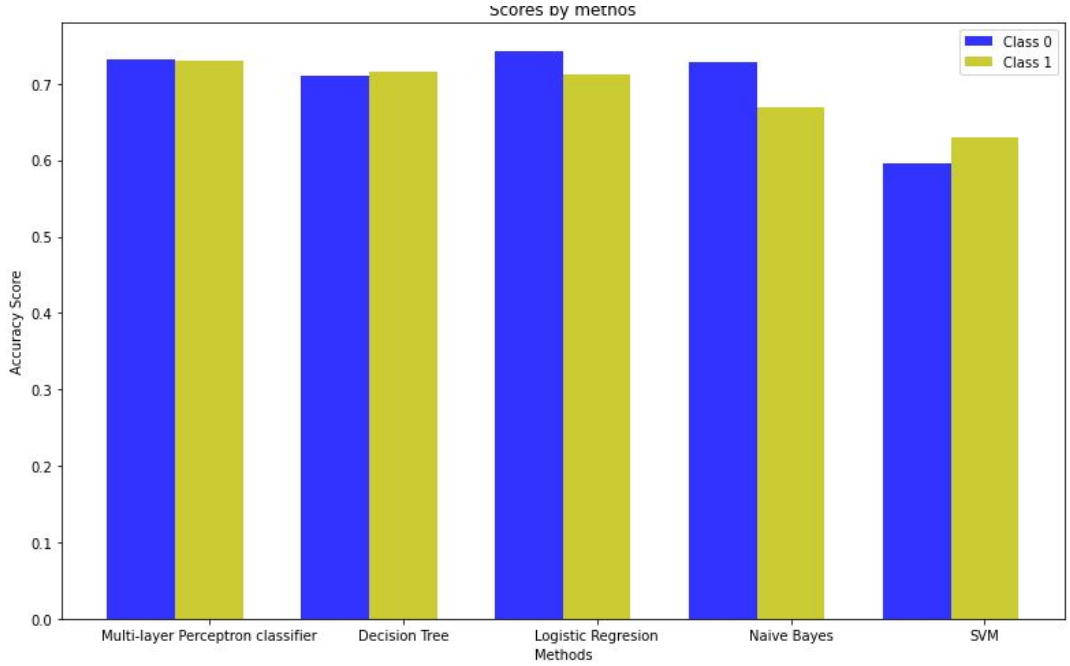
Результати роботи методів машинного навчання

	Multi-layer Perceptron classifier	Logistic Regresion	Decision Tree	Naive Bayes	SVM
Accuracy Score	0.7318	0.7281	0.7128	0.7025	0.6130
Class 0 Precision	0.7388	0.7129	0.7253	0.6760	0.6320
Class 1 Precision	0.7248	0.7467	0.7011	0.7409	0.5973
Class 0 F1-score	0.7327	0.7424	0.7103	0.7290	0.5951
Class 1 F1-score	0.7308	0.7121	0.7153	0.6701	0.6293
Class 0 Recall	0.7266	0.7745	0.6960	0.7911	0.5623
Class 1 Recall	0.7370	0.6807	0.7301	0.6117	0.6648
Class 0 Support	6980.0000	6980.0000	6980.0000	6980.0000	6980.0000
Class 1 Support	6817.0000	6817.0000	6817.0000	6817.0000	6817.0000
Studying time	89.6086	1.7547	0.0833	0.0156	146.4230
Testing time	0.0467	0.0028	0.0027	0.0046	8.3821
Time sum	89.6554	1.7576	0.0860	0.0203	154.8051

Результати навчених моделей на набір даних №1

Результати роботи методів машинного навчання





Результати навчених моделей на набір даних №1

Результати роботи методів машинного навчання

Виключаний GPU прискорювач

	Studying time	Testing time	Time sum
SVM	146.423021	8.382121	154.805142
Multi-layer Perceptron classifier	89.608643	0.046730	89.655374
Logistic Regresion	1.754727	0.002841	1.757568
Decision Tree	0.083340	0.002668	0.086008
Naive Bayes	0.015631	0.004623	0.020254

Включаний GPU прискорювач

	Studying time	Testing time	Time sum
SVM	143.149073	8.764758	151.913831
Multi-layer Perceptron classifier	74.332118	0.047669	74.379787
Logistic Regresion	0.930808	0.002733	0.933540
Decision Tree	0.086016	0.003266	0.089282
Naive Bayes	0.016044	0.004637	0.020681

Діаграма результати роботи моделей за часом навчання, набір даних №1

Висновки

Визначення кардіоваскулярних захворювань є актуальною на сьогоднішній день. Дане захворювання тільки прогресуватиме у найближчому майбутньому через погіршення стану екології та якості харчування. Тому стоїть питання швидкої їх діагностики та покращення здоров’я.

На проведених мною дослідженнях було визначено, що модель побудована методом нейронної мережі була кращою чим метод логістичної регресії, наївного Байаса, дерева рішень, векторів підтримки.

Дякую за увагу)