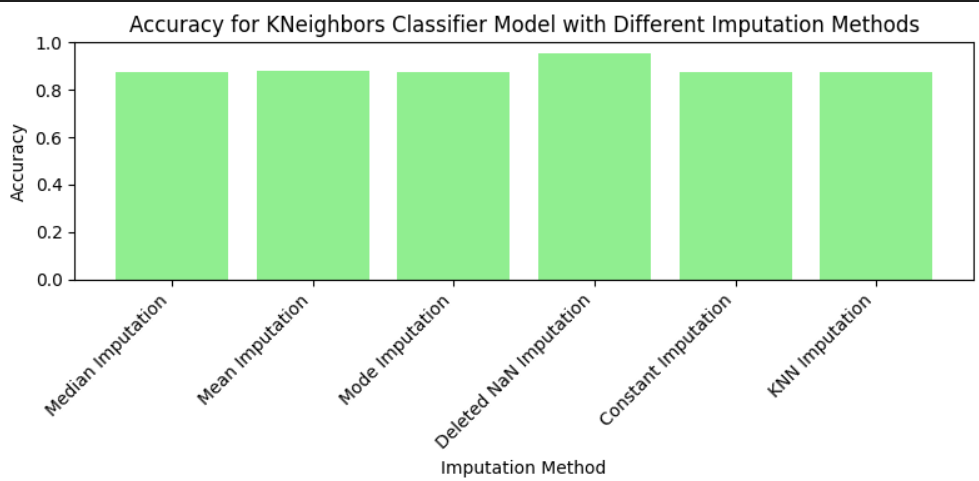
**4 ОБГОВОРЕННЯ РЕЗУЛЬТАТІВ ЕКСПЕРИМЕНТІВ**

У цьому розділі ми детально розглянемо результати експериментів, проведених з різними методами обробки пропущених значень та застосуванням різних алгоритмів машинного навчання на різних наборах даних. Ми оцінюватимемо, які комбінації методів обробки даних та алгоритмів машинного навчання дають найкращі результати для 3 наборів даних.

**4.1 Опис і порівняння результатів використовуючи Heart Attack Dataset**

Порівняння усіх методів імпутації для певного методу машинного навчання.

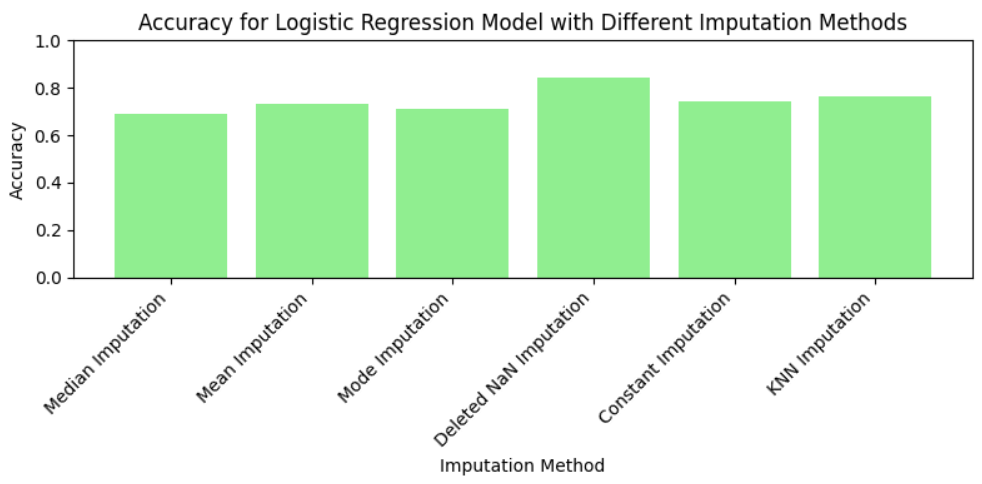


**Рисунок 1** Діаграма точності для методу КНН

|  |  |
| --- | --- |
| Модель машинного навчання КНН | |
| Види імпутації | Точність тренування |
| Медіана | 87.65% |
| Середнє значення | 88.27% |
| Мода | 87.65% |
| Видалення NaN значень | 95.48% |
| Заміна на кностанту | 87.65% |
| КНН | 87.65% |

**Таблиця 2** Результати точності методу КНН з 6-тьма методами імпутації.

Рисунок 1 показує діаграму, на якій є порівняння результатів використовуючи метод машинного навчання КНН з 6 методами обробки пропущених значень. Таблиця 2 показує точність кожного методу ообреблення пропущених значень. Рисунок 1 і 2 показують найефективніший метод обробки пропущених значень для методу КНН у цьому датасеті є видалення пропущених значень яка досягає точності 95,48%. Найменш ефективним методом обробки пропущених значень є методи: медіана, мода, константа і КНН – з точністю 87.6%.

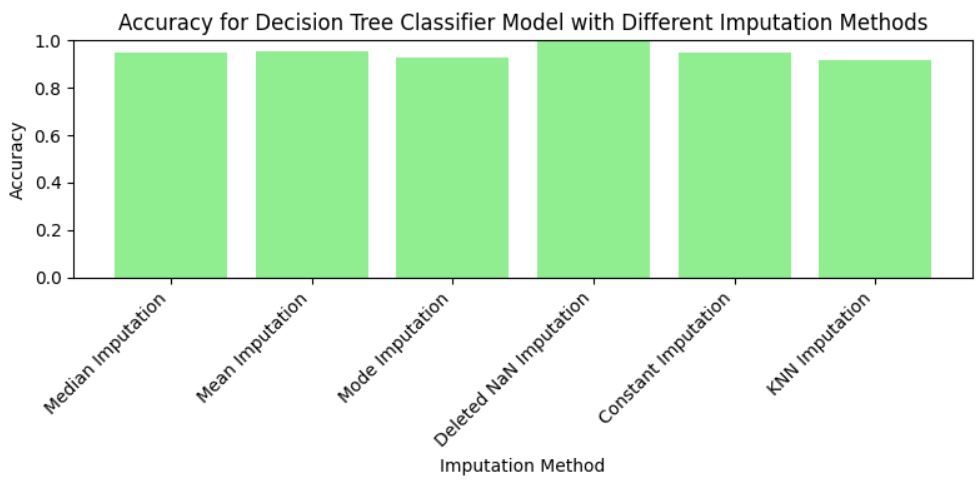
****

**Рисунок 3** Діаграма точності логістичної регресії

|  |  |
| --- | --- |
| Модель машинного навчання Логістична регресія | |
| Види імпутації | Точність тренування |
| Медіана | 69.13% |
| Середнє значення | 73.45% |
| Мода | 70.98% |
| Видалення NaN значень | 84.21% |
| Заміна на кностанту | 74.07% |
| КНН | 76.65% |

**Таблиця 4** Результати точності логістичної регресії з 6-тьма методами імпутації

Рисунок 3 є діаграмою, яка містить порівняння усіх методів оброблення пропущених значення використовуючи логістичну регресію. Рисунок 3 і Таблиця 4 демонструють, що найкраща точність досягається при видалені пропущених значень, а саме 84.21% точності, найгірша точність спостерігається у медіани і моди – 69.13% і 70.98% відповідно.

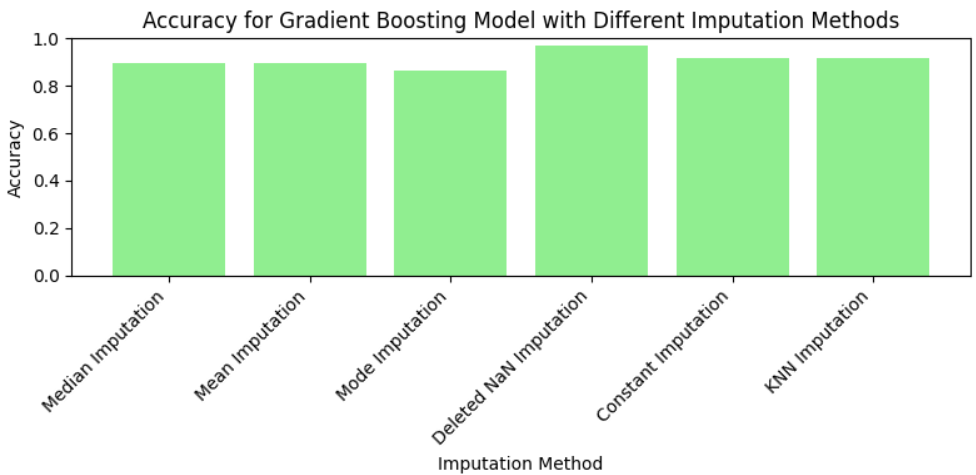
****

**Рисунок 5** Діаграма точності для дерева рішень

|  |  |
| --- | --- |
| Модель машинного навчання Дерево рішень | |
| Види імпутації | Точність тренування |
| Медіана | 95.06% |
| Середнє значення | 95.67% |
| Мода | 92.59% |
| Видалення NaN значень | 100.00% |
| Заміна на кностанту | 95.06% |
| КНН | 91.97% |

**Таблиця 6** Результати точності дерева рішень зі шістьма методами імпутації

У Рисунку 5 показано діаграму порівнянь усіх пропущених значень з методом машинного навчання: дерево рішень. Таблиця 6 показує точність кожного методу обробки пропущених значень для методу дерева ріщень. Проаналізувавши рисунки 5 і 6, вияснено, що найкраща точність досягається використовуючи видалення пропущених значень – 100% точності. Методи обробки пропущених значень такі як: медіана, середнє значення, заміна на константу – мають точність +- 95.3%. Найменш ефективними є мода і КНН – їхня точність 92.59% і 91.97%.

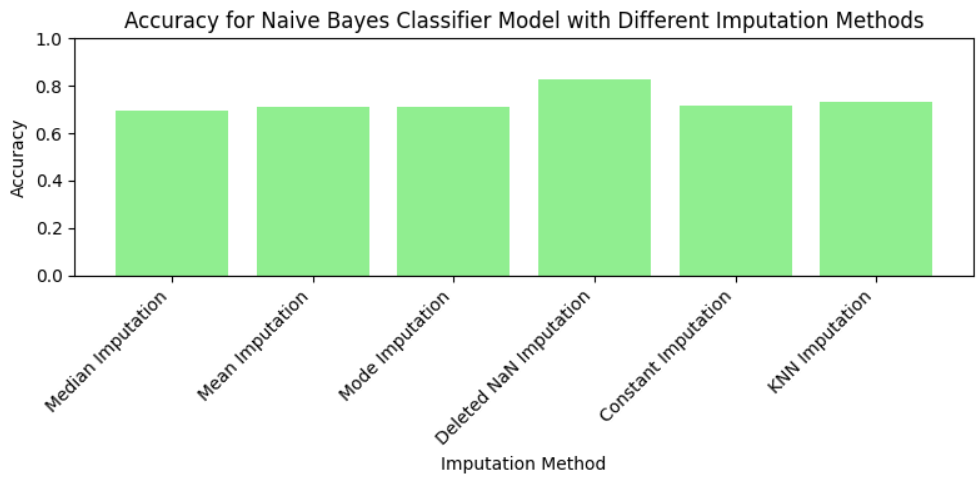
****

**Рисунок 7** Діаграма точності для градієнтного прискорення

|  |  |
| --- | --- |
| Модель машинного навчання Градієнтне прискорення | |
| Види імпутації | Точність тренування |
| Медіана | 89.50% |
| Середнє значення | 89.50% |
| Мода | 86.41% |
| Видалення NaN значень | 96.99% |
| Заміна на кностанту | 91.97% |
| КНН | 91.97% |

**Таблиця 8** Результати моделі машинного навчання – градієнтного прискорення 6 методами імпутації.

Рисунок 7 є діаграмою, яка відображає різні методи імпутації і їхню точність при виконанні з градієнтним прискоренням. Таблиця 8 надає результати при виконані алгоритму градієнтного прискорення. Рисунки 7 і Таблиця 8 надають порівняти найоптимальніші методи оброблення пропущених значень. Найоптимальніший метод є видаленням пропущених значень, який досягає точності в 96.99%, а найменш ефективний метод є мода, який має точність 86.41%.



**Рисунок 9** Діаграма точності для Наївного Баєса

|  |  |
| --- | --- |
| Модель машинного навчання Наївний Баєс | |
| Види імпутації | Точність тренування |
| Медіана | 69.75% |
| Середнє значення | 70.98% |
| Мода | 70.98% |
| Видалення NaN значень | 82.70% |
| Заміна на кностанту | 71.60% |
| КНН | 73.34% |

**Таблиця 10** Результати Наївного Баєса з різними методами імпутації.

Рисунок 9 являється діаграмою, яка містить різні методи обробки пропущених значень і їхню точність з використанням методу машинного навчання Наївного Баєса. Таблиця 10 Даний метод має найкращі показники з поєднанням методу видалення пропущених значень, і найгіршу точність з методом медіани 69.75%.

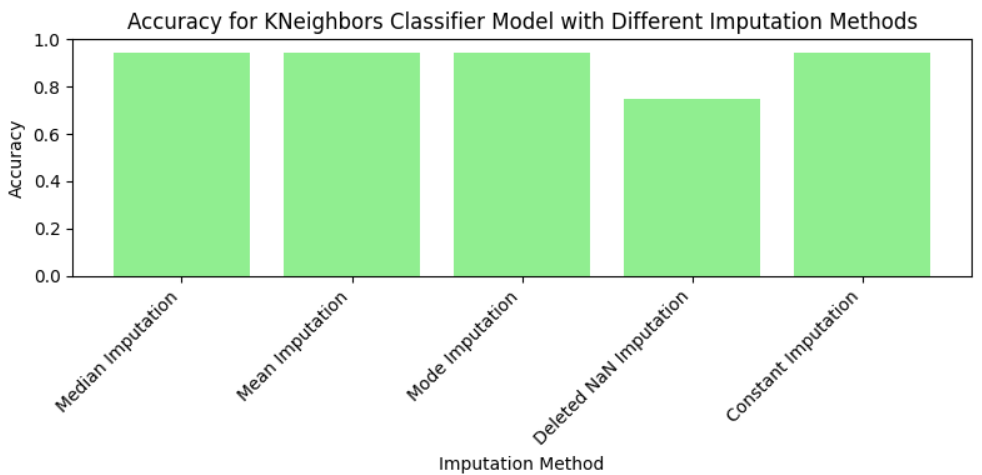
На Рисунках 1 – 10 порівняно 6 методів оброблення пропущених значень з 5 методами машинного навчання. Найкраща модель машинного навчання для цього датасету є дерево рішень, у цьому наборі найефективніший результат має точність 100%. Найгірший метод машинного навчання для даного датасету є Наївний Баєс, який має найкращу точність 82% з видаленням пропущених даних, а найгіршу 69% з медіаною.

Найкраще поєднання методу і машинного навчання є дерево рішень з видаленням пропущених значень, точність комбінації досягає 100%. Найгірша комбінація машинного навчання і обробки пропущених значень є логістична регресія з медіаною і Наївний Баєс з медіаною, їхня точність 69%.

Метод КНН є ефективним у цьому датасеті, оскільки даний метод швидко виконувався, через те що у наборі даних містилося 1622 рядків.

**4.2 Опис і порівняння результатів використовуючи Covid-19 Dataset**

Порівняння усіх методів імпутації для певного методу машинного навчання.

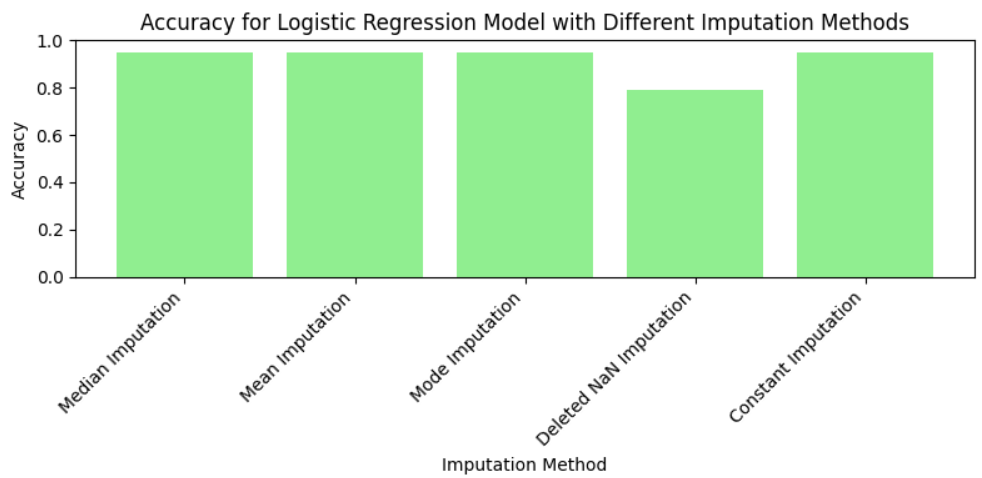


**Рисунок 11** Результати КНН з різними методами імпутації.

|  |  |
| --- | --- |
| Модель машинного навчання КНН | |
| Види імпутації | Точність тренування |
| Медіана | 94.25% |
| Середнє значення | 94.21% |
| Мода | 94.25% |
| Видалення NaN значень | 74.79% |
| Заміна на кностанту | 94.25% |

**Таблиця 12** Результати КНН з 5 методами імпутації.

Рисунок 11 є діаграмою, яка показує точність методу КНН з 5 методами імпутації. Рисунок 11 і Таблиця 12 показують точність даного методу при обробці різними методами імпутації. Метод КНН має доволі точні результати, а саме 4/5 методів імпутації мають точність 95%. Видалення пропущених значень мають точність 75%, що є найменшою точністю у цьому методі. Також даний метод обчислювався найповільніше зі всіх методів обробки, оскільки він створений для малої кількості даних. Повільніше виконувався в 40 разів порівнянні з іншими методами машинного навчання.

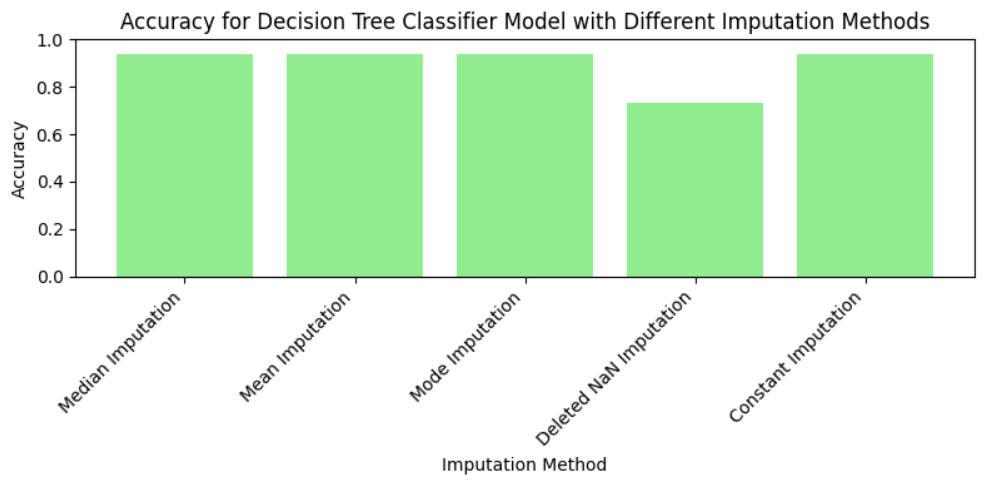


**Рисунок 13** Результати логістичної регресії з різними методами імпутації.

|  |  |
| --- | --- |
| Модель машинного навчання Логістична регресія | |
| Види імпутації | Точність тренування |
| Медіана | 94.82% |
| Середнє значення | 94.84% |
| Мода | 94.83% |
| Видалення NaN значень | 78.81% |
| Заміна на кностанту | 94.82% |

**Таблиця 14** Результати логістична регресія з 5 методами імпутації

Рисунок 13 є діаграмою, яка показує точність результатів логістичної регресії з різними видами обробки пропущених значень. Таблиця 14 показує детальніше інформацію про результати обробки. Найкращі результати є у заміни на константу, медіана, середнє значення, мода – 94% точності. Найгірші показники має метод видалення пропущених значень – 78.8%.

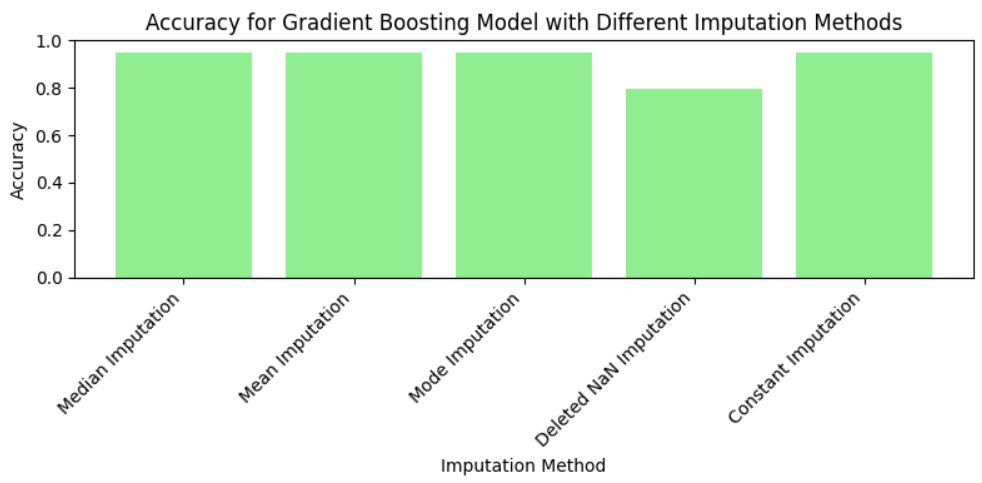


**Рисунок 15** Результати дерева рішень з різними методами імпутації.

|  |  |
| --- | --- |
| Модель машинного навчання Дерево рішень | |
| Види імпутації | Точність тренування |
| Медіана | 93.83% |
| Середнє значення | 93.92% |
| Мода | 93.85% |
| Видалення NaN значень | 73.48% |
| Заміна на кностанту | 93.84% |

**Таблиця 16** Результати дерева рішень з 5 методами імпутації.

Рисунок 15 показує точність кожного методу обробки пропущених значень використовуючи дерево рішень. Таблиця 16 надає точну інформацію щодо результатів точності методу дерева рішень. Найефективнішим методом обробки у даному методі є: медіана, середнє значення, мода, заміна на константу – їх точність досягає 93%. Метод середнього значення є на 0.1% є кращим інші методи. Найгірші показники є у методі видалені NaN значень – 73.48%, що є доволі низьким результатом.

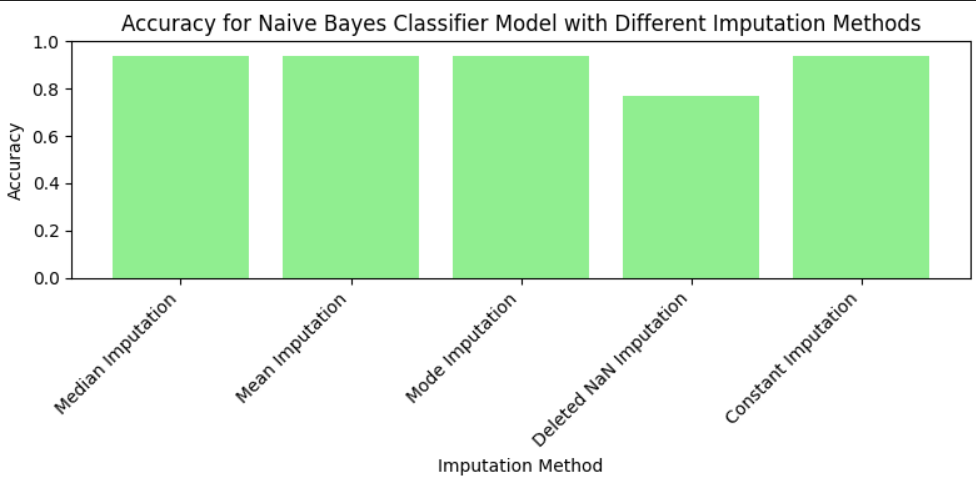


**Рисунок 17** Результати градієнтне прискорення з різними методами імпутації.

|  |  |
| --- | --- |
| Модель машинного навчання Градієнтне прискорення | |
| Види імпутації | Точність тренування |
| Медіана | 95.10% |
| Середнє значення | 95.15% |
| Мода | 95.09% |
| Видалення NaN значень | 79.76% |
| Заміна на кностанту | 95.08% |

**Таблиця 18** Результати градієнтне прискорення з 5 методами імпутації.

Рисунок 17 і Таблиця 18 надають інформацію щодо точності алгоритму машинного навчання, а саме градієнтного прискорення і його точність в залежності від різних методів обробки пропущених значень. Найгірший результат є виконаний за допомогою видалення пропущених значень, його точність досягає 79.76%. Усі інші методи мають точність 95%.



**Рисунок 19** Результати Наївного Баєса з різними методами імпутації.

|  |  |
| --- | --- |
| Модель машинного навчання КНН | |
| Види імпутації | Точність тренування |
| Медіана | 93.84% |
| Середнє значення | 93.88% |
| Мода | 93.84% |
| Видалення NaN значень | 76.89% |
| Заміна на кностанту | 93.84% |

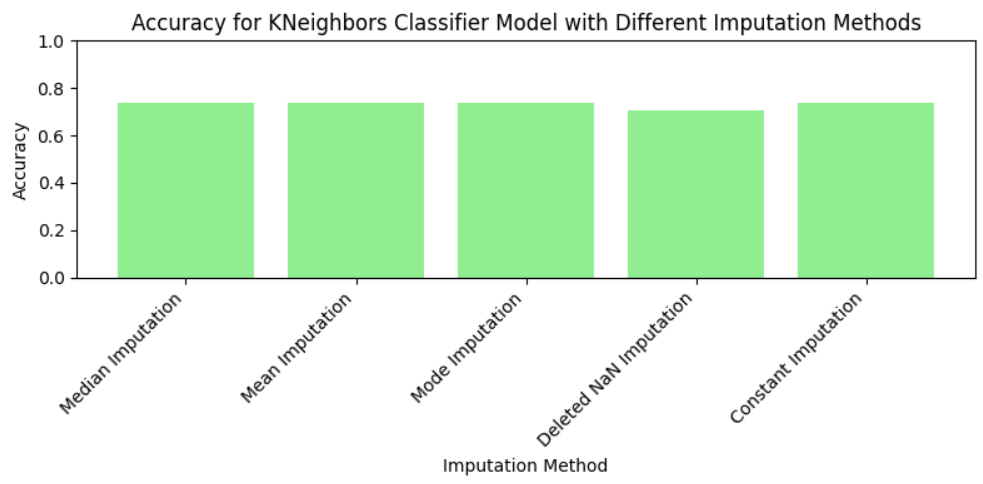
**Таблиця 20** Результати Наївного Баєса з 5 методами імпутації.

Рисунок 19 і Таблиця 20 показують точність е методів імпутації використовуючи метод Наївного Баєса. Найгірші результати також спостерігаються у метода видалення пропущених значень – 76.89%. Інші моделі мають точність 93%. Найкращий метод оброблення пропущених значень у Наївному Баєсі є середнє значення, цей метод кращий на 0.05% від другої з моделі, має точність 93.88%.

У цьому датасеті були проаналізовані Рисунки 11 – 20. У цьому датасеті всі моделі машинного навчання мають приблизно одинакову точність 93-95%. Найкраща модель машинного навчання для датасету Covid-19 є градієнтне прискорення, його середня точність 95%. Найгірша модель машинного навчання у даного датасету є Наївний Баєс і дерево рішень, які мають точність 93.5%. Найгірший метод оброблення пропущених значень у всіх 5 методів машинного навчання є видалення пропущених значень. Його точність досягає у всіх методах машинного навчання – 73-79%.

**4.3 Опис і порівняння результатів використовуючи Diabetes Health Indicators Dataset**

У цьому датасеті кількість пропущених даних досягає 10%. Порівняння усіх методів імпутації для певного методу машинного навчання.

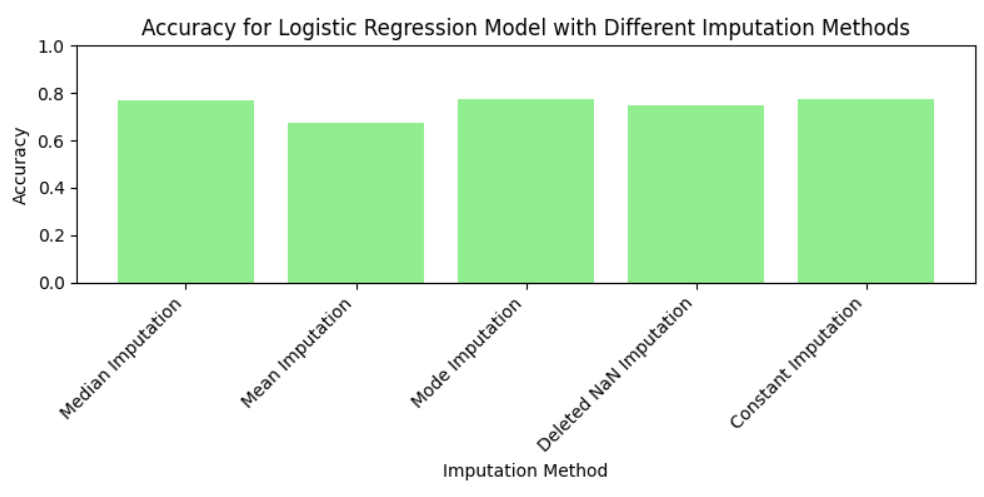


**Рисунок 21** Діаграма точності для методу КНН

|  |  |
| --- | --- |
| Модель машинного навчання КНН | |
| Види імпутації | Точність тренування |
| Медіана | 74.39% |
| Середнє значення | 74.39% |
| Мода | 74.95% |
| Видалення NaN значень | 70.71% |
| Заміна на кностанту | 74.02% |

**Таблиця 22** Результати точності методу КНН з 5-тьма методами імпутації

Рисунок 1 показує діаграму, на якій здійснюється порівняння результатів методів машинного навчання КНН з 6 методами обробки пропущених значень. Таблиця 22 показує точність кожного методу оброблення пропущених значень. Рисунок 1 і 2 показують найефективніший метод обробки пропущених значень для методу КНН у цьому датасеті є мода: 74.95%. Найменш ефективним методом обробки пропущених значень є метод видаленням пропущених значень – 70.71%.

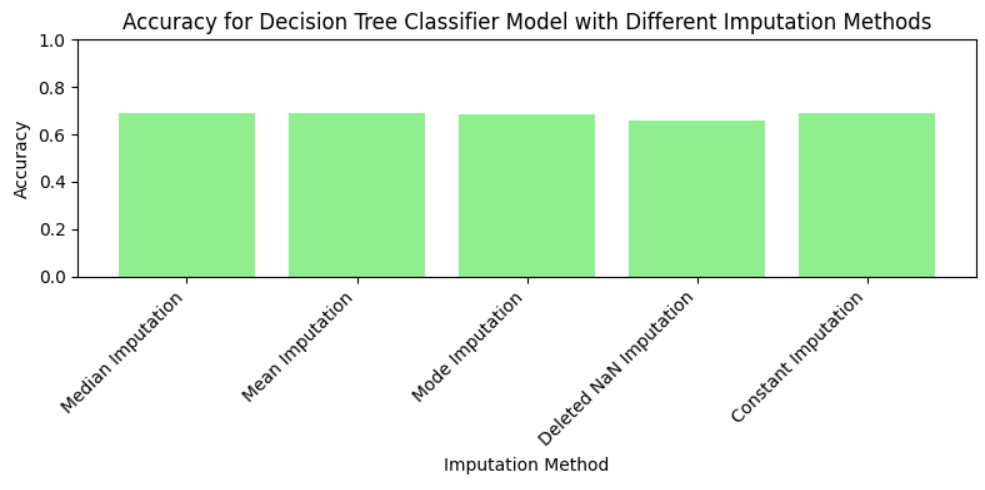
****

**Рисунок 23** Діаграма точності логістичної регресії

|  |  |
| --- | --- |
| Модель машинного навчання Логістична регресія | |
| Види імпутації | Точність тренування |
| Медіана | 78.96% |
| Середнє значення | 67.43% |
| Мода | 78.34% |
| Видалення NaN значень | 74.86% |
| Заміна на кностанту | 77.58% |

**Таблиця 24** Результати точності логістичної регресії з 5-тьма методами імпутації

Рисунок 23 є діаграмою, яка містить порівняння усіх методів оброблення пропущених значення використовуючи логістичну регресію. Рисунок 23 і Таблиця 24 показують, що найкраща точність досягається при методі імпутації: медіана, а саме 78.96% точності, найгірша точність спостерігається у середнього значення – 67.43%.

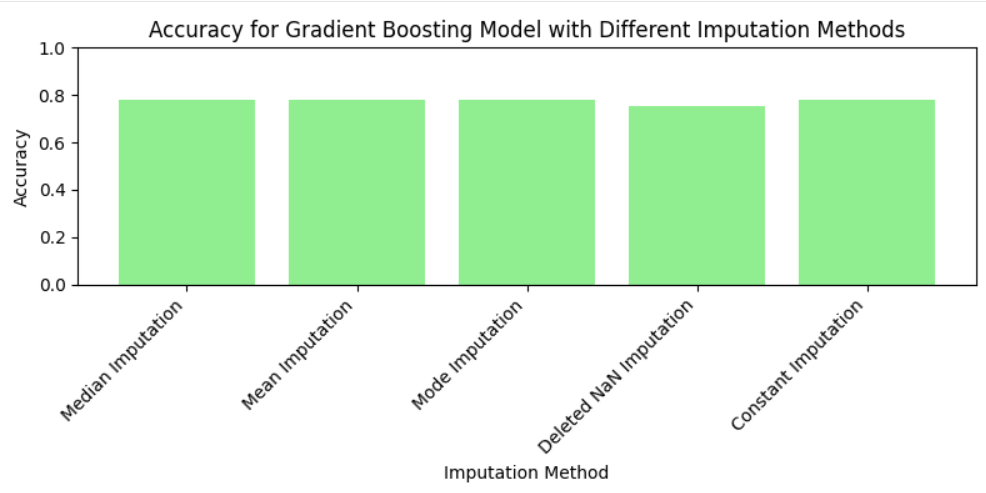
****

**Рисунок 25** Діаграма точності для дерева рішень

|  |  |
| --- | --- |
| Модель машинного навчання Дерево рішень | |
| Види імпутації | Точність тренування |
| Медіана | 69.03% |
| Середнє значення | 69.25% |
| Мода | 68.77% |
| Видалення NaN значень | 65.97% |
| Заміна на кностанту | 68.68% |

**Таблиця 26** Результати точності дерева рішень зі п’ятьма методами імпутації

У рисунку 25 показано діаграми для метода машинного навчання – дерево рішень, яке оброблене 5 видами імпутації. Таблиця 26 демонструє точність результатів оброблення датасету різними методами імпутації. Через те що види імпутації є різними, то набір даних має різні точності, найкраща з яких середнє значення – 69.25%, а найменш ефективним – видалення пропущених значень – 65.97%.

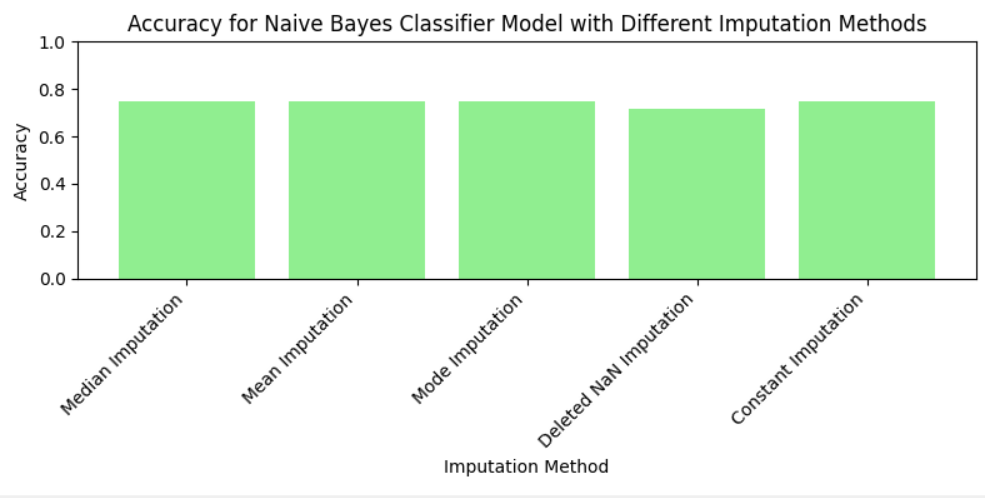
****

**Рисунок 27** Діаграма точності для градієнтного прискорення

|  |  |
| --- | --- |
| Модель машинного навчання Градієнтне прискорення | |
| Види імпутації | Точність тренування |
| Медіана | 77.96% |
| Середнє значення | 78.18% |
| Мода | 77.97% |
| Видалення NaN значень | 75.13% |
| Заміна на кностанту | 78.02% |

**Таблиця 28** Результати моделі машинного навчання – градієнтного прискорення 5 методами імпутації.

Рисунок 27 і Таблиця 28 показують точність моделі градієнтного прискорення з 5 методами обробки пропущених значень. Таблиця містить детальніші результати. Найкращий метод є медіана, його точність є 78.18%, а найгіршим методом являється видаленням пропущених значень – 75.13%.



**Рисунок 29** Діаграма точності для Наївного Баєса

|  |  |
| --- | --- |
| Модель машинного навчання Наївний Баєс | |
| Види імпутації | Точність тренування |
| Медіана | 74.94% |
| Середнє значення | 74.94% |
| Мода | 74.94% |
| Видалення NaN значень | 74.90% |
| Заміна на кностанту | 71.94% |

**Таблиця 30** Результати Наївного Баєса з різними методами імпутації.

Рисунок 29 і Таблиця 30 показують точність методу машинного навчання, а саме Наївного Баєса з різними методами імпутації. Найгірші показники має метод заміни а константу – 71.94%. Інші методи мають одинакову точність – приблизно 74.94%.

На рисунках 1 – 30 було порівняно різні методи імпутації з 5 методами машинного навчання. 3 набори даних містили різну кількість даних, їхня кількість суттєво відрізняється. Найгіршим методом для обробки значень у великих наборах даних виявився КНН, цей метод є часо-затратним і зі збільшенням кількості даних, його обчислення сповільнювалося. Цей метод підійшов тільки для першого датасету. Для останнього датасету він не міг обчислитися протягом 10 годин, через довгу класифікацію.

Метод видалення пропущених значень добре показав себе у першому датасеті при наборі даних 1622. Його найкраща точність досягнула 100% з комбінацією дерева рішень. У більших наборах даних, другий і ретій датасет, видалення пропущених значень мало найгірші результати при різних моделях машинного навчання.

При великій обробці даних найкраще себе показав метод обробки мода і медіана. Використовуючи 5 методів машинного навчання, результати найкращої ефективності мінялися, мода – медіана.

Нажаль немає найкращого методу оброблення пропущених значень у великих обсягах даних. Все залежить від набору даних, чи він є збалансованим, методів машинного навчання. Але з найкращого методу перетворюється в найгірший метод оброблення пропущених значень – видалення пропущених значень. Цей метод надає гірші результати зі збільшенням кількості даних у датасеті.