Міністерство освіти і науки України

Національний університет «Львівська політехніка»

Інститут комп’ютерних наук та інформаційних технологій

Кафедра автоматизованих систем управління



РОЗРАХУНКОВО-ГРАФІЧНА РОБОТА

з дисципліни

“Основи смарттехнологій і систем”

**Виконав**

студент групи *ОІ-31*

*Мацишин Михайло*

**Прийняла:**

*Нарушинська О. О.*

Львів – 2024

**ЗМІСТ**

[Вибір датасету та опис бізнес-проблеми 3](#_Toc182411610)

[Етап розуміння даних 5](#_Toc182411611)

[Етап підготовки даних (Feature engineering) 15](#_Toc182411612)

[Етап моделювання та вибір найкращої моделі, її оптимізації 21](#_Toc182411613)

[Етап розгортання моделі 29](#_Toc182411614)

[ВИСНОВОК 41](#_Toc182411615)

# **Вибір датасету та опис бізнес-проблеми**

Я обрав датасет [Asteroid Dataset](https://www.kaggle.com/datasets/sakhawat18/asteroid-dataset/data) з Kaggle, він містить детальну інформацію про астероїди, включаючи їхні фізичні характеристики та орбітальні параметри. Цей набір даних зацікавив мене своєю різноманітністю параметрів, що дозволяє детально аналізувати кожен астероїд та визначати потенційні загрози для Землі. Окрім того, набір даних містить важливі показники, які можуть бути використані для прогнозування, що робить його ідеальним для регресійних моделей.

Переглянувши даний датасет, я вирішив обрати цільовою змінною діаметр астероїда (**diameter**), оскільки саме ця характеристика дає змогу оцінити потенційну небезпеку об'єкта для Землі. Цей показник залежить від ряду інших змінних, таких як орбітальні параметри та абсолютна зоряна величина, що робить його актуальним для задачі регресії. Прогнозування діаметра є ключовим, оскільки розмір астероїда прямо пов’язаний з рівнем потенційної небезпеки для Землі. Модель, що передбачає розмір астероїда, допоможе краще оцінювати його ризик і планувати заходи для забезпечення безпеки.

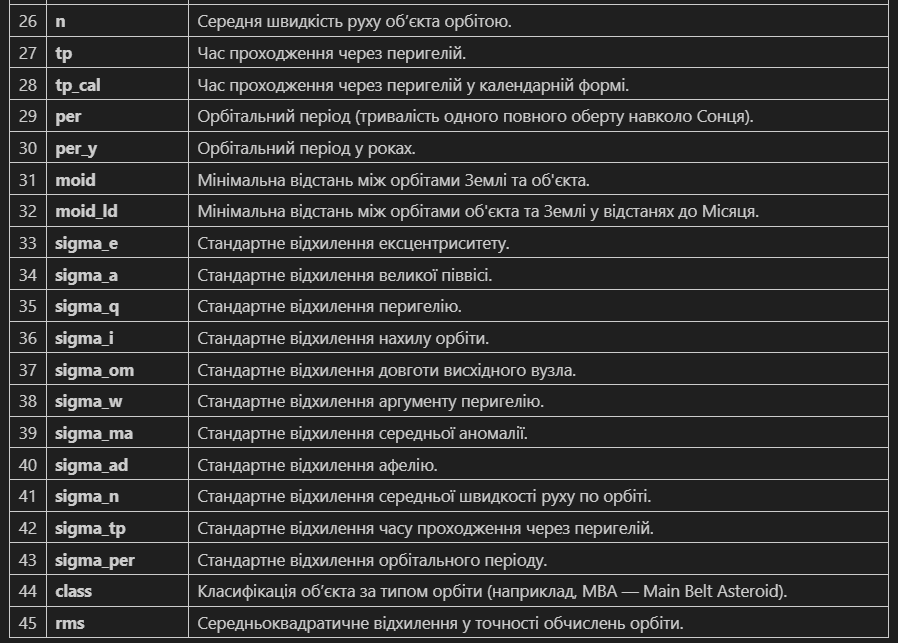
**Бізнес-цілі:**

1. **Оцінка ризиків зіткнення з астероїдами:**
   * **Ціль моделювання**: Розробка моделі для прогнозування розмірів астероїдів, які можуть представляти загрозу при зближенні із Землею. Модель дозволить виділити об’єкти, які варто моніторити в першу чергу.
   * **Критерій успіху**: Підвищення точності прогнозів щодо потенційно небезпечних об’єктів та зменшення витрат на моніторинг завдяки фокусуванню на найбільш ризикових астероїдах.
2. **Оптимізація програм моніторингу космічних об'єктів:**
   * **Ціль моделювання**: Створення моделі, яка дозволить прогнозувати великі астероїди, що можуть впливати на орбіти супутників або інші космічні об’єкти. Це допоможе оптимізувати програми спостережень.
   * **Критерій успіху**: Підвищення ефективності моніторингу та зменшення кількості збоїв, пов’язаних з потенційним впливом астероїдів на космічну техніку.
3. **Планування заходів для зниження загрози астероїдів:**
   * **Ціль моделювання**: Використання прогнозних моделей для визначення найнебезпечніших астероїдів та розробка стратегій для їх нейтралізації, включаючи можливе відхилення траєкторії.
   * **Критерій успіху**: Підвищення рівня безпеки планети шляхом завчасного виявлення небезпечних об'єктів та реалізації планів дій для запобігання можливим зіткненням.

Ці бізнес-цілі відображають стратегічний підхід до моніторингу та управління потенційними загрозами з боку астероїдів. Використання даних про фізичні та орбітальні характеристики дозволяє не лише виявляти найбільш небезпечні об'єкти, але й планувати ефективні заходи для запобігання можливим зіткненням. Таким чином, розробка регресійної моделі для передбачення розміру астероїдів допоможе оптимізувати ресурси, підвищити рівень безпеки планети та зменшити ризики для космічної інфраструктури.

# **Етап розуміння даних**

**Опис ознак:**

****

У цій частині я зчитав датасет, який містив 958,524 рядків та 45 колонок. Оскільки для подальшого аналізу я обрав менший обсяг даних, я зменшив розмір датасету до 2,000 записів.

Для цього я виконував наступні кроки:

1. **Вибір рядків без жодного пропуску**, крім стовпця 'prefix', який не використовувався в аналізі. Я видалив ці стовпці та вибрав перші 1,700 повних рядків.
2. **Вибір рядків з невеликою кількістю пропусків** (до 5 пропусків у кожному рядку). З цих рядків я вибрав додатково 300 записів, щоб мати 2,000 рядків загалом.
3. **Об'єднання двох вибірок**: одну з повними записами (1,700 рядків) та іншу з неповними (300 рядків).

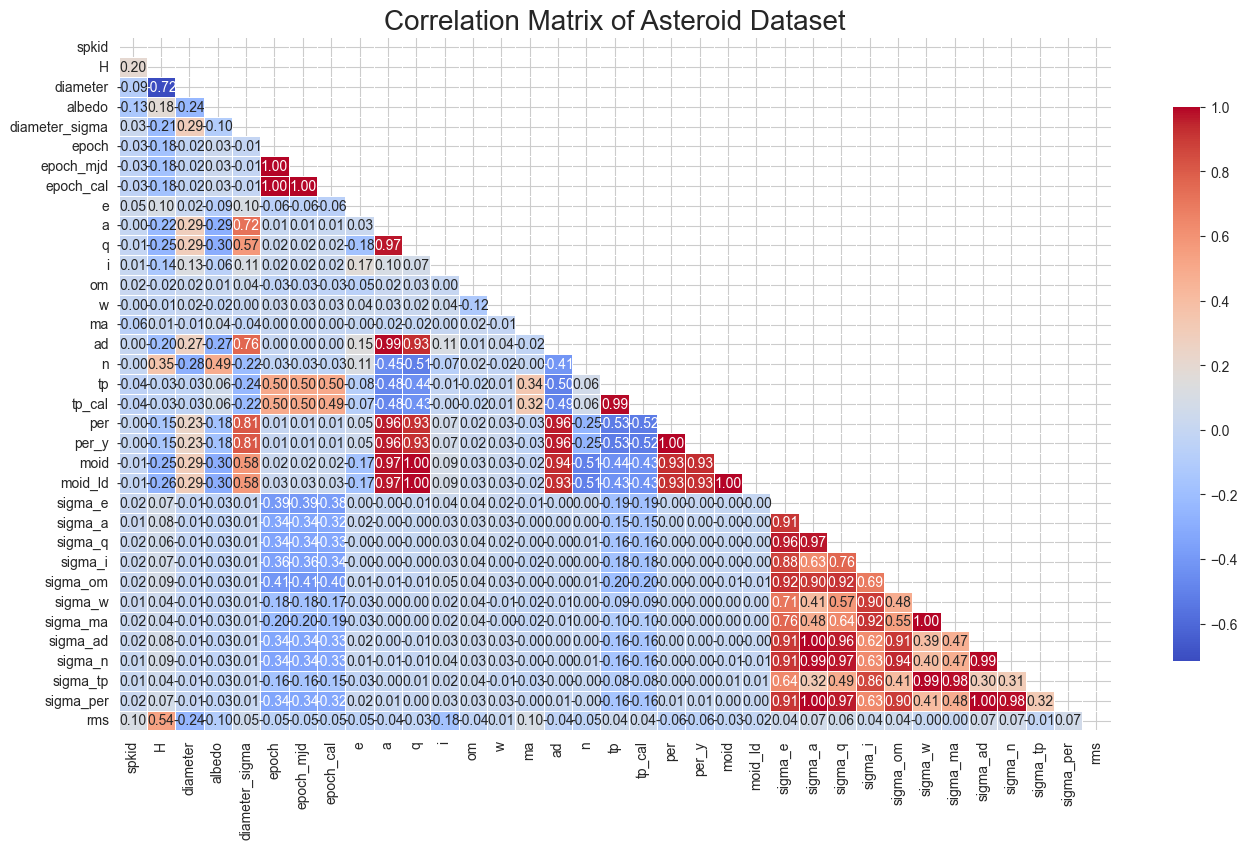
Нарешті, я зберіг цей зменшений датасет у файл filtered\_dataset.csv для подальшого використання. Розмір отриманого датасету складав 2,000 рядків та 45 колонок.

Далі виконано перевірку наявності пропущених значень у тренувальному наборі даних. Як показано в результатах, деякі стовпці мають пропущені значення, зокрема:

* **name**: 290 пропущених значень
* **prefix**: 2000 пропущених значень
* **pha**: 6 пропущених значень
* **H**: 2 пропущених значення
* **diameter**: 246 пропущених значень
* **albedo**: 247 пропущених значень
* **diameter\_sigma**: 246 пропущених значень
* Інші стовпці, такі як **moid**, **sigma\_e**, **sigma\_a** та інші, також містять пропущені значення в різних кількостях (6 пропущених значень у кожному з цих стовпців).

Ці пропущені значення потребують обробки перед подальшим використанням даних для побудови моделі. Залежно від контексту можна застосувати різні методи обробки, наприклад, імпутацію середнім значенням, заповнення значеннями за замовчуванням чи видалення неповних рядків.

**Кореляційна матриця:**



Кореляційна матриця дозволяє визначити силу взаємозв'язку між числовими змінними в датасеті. Чим ближче значення кореляції до +1 або -1, тим сильніший позитивний чи негативний зв'язок між змінними. Ось основні спостереження:

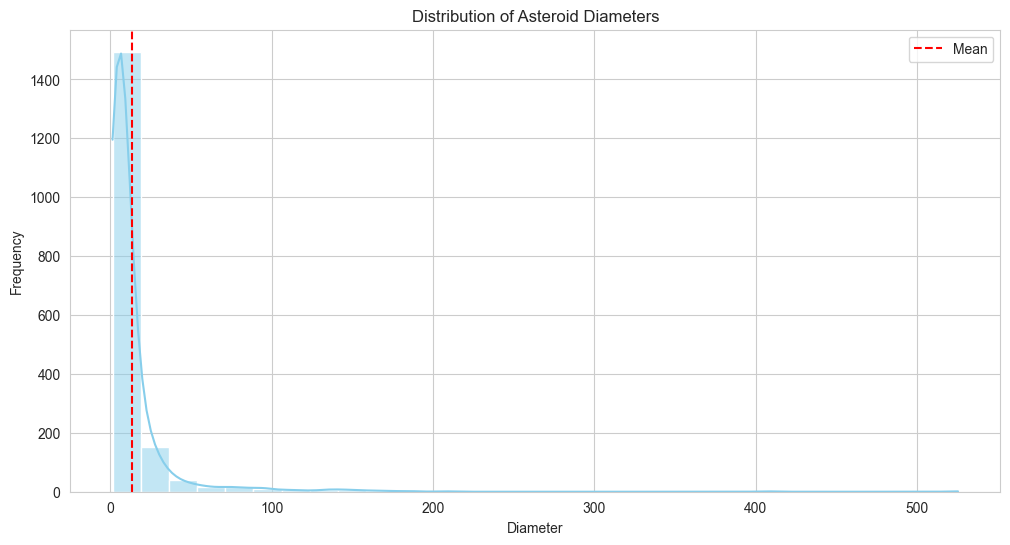
**Основні спостереження**

1. **Висока кореляція між орбітальними параметрами**:
   * Спостерігається дуже висока позитивна кореляція між орбітальними параметрами астероїда, зокрема між параметрами a (велика піввісь) і ad (апогей). Це логічно, оскільки ці параметри фізично взаємопов'язані.
   * Також параметр per (період обертання) має високу кореляцію з параметрами a та ad, що вказує на зв'язок між великими орбітальними характеристиками і періодом обертання астероїда.
2. **Сильна кореляція серед sigma-параметрів**:
   * Параметри невизначеності вимірювань (sigma\_\*), наприклад, sigma\_e (невизначеність ексцентриситету) і sigma\_q (невизначеність перигею), мають високу кореляцію. Це свідчить про те, що невизначеності вимірювань можуть бути пов'язані з певними характеристиками інструментів або методами вимірювань.
3. **Кореляція діаметра з іншими змінними**:
   * Кореляція між діаметром астероїда та іншими змінними є відносно низькою. Діаметр має слабкий позитивний зв'язок з H (абсолютна зоряна величина), але цей зв'язок не є достатньо сильним для точного прогнозування.
   * Відсутність сильних кореляцій з діаметром вказує на те, що його важко точно передбачити лише на основі орбітальних параметрів, і для цього можуть знадобитися додаткові характеристики, наприклад, відбивна здатність або склад астероїда.
4. **Залежність між параметрами epoch\_mjd, epoch\_cal, tp і tp\_cal**:
   * Параметри epoch\_mjd, epoch\_cal, tp, і tp\_cal мають високу кореляцію між собою. Це пояснюється тим, що ці параметри можуть представляти ту саму подію або дуже схожі події, виміряні в різних одиницях або форматах.

**Висновок**

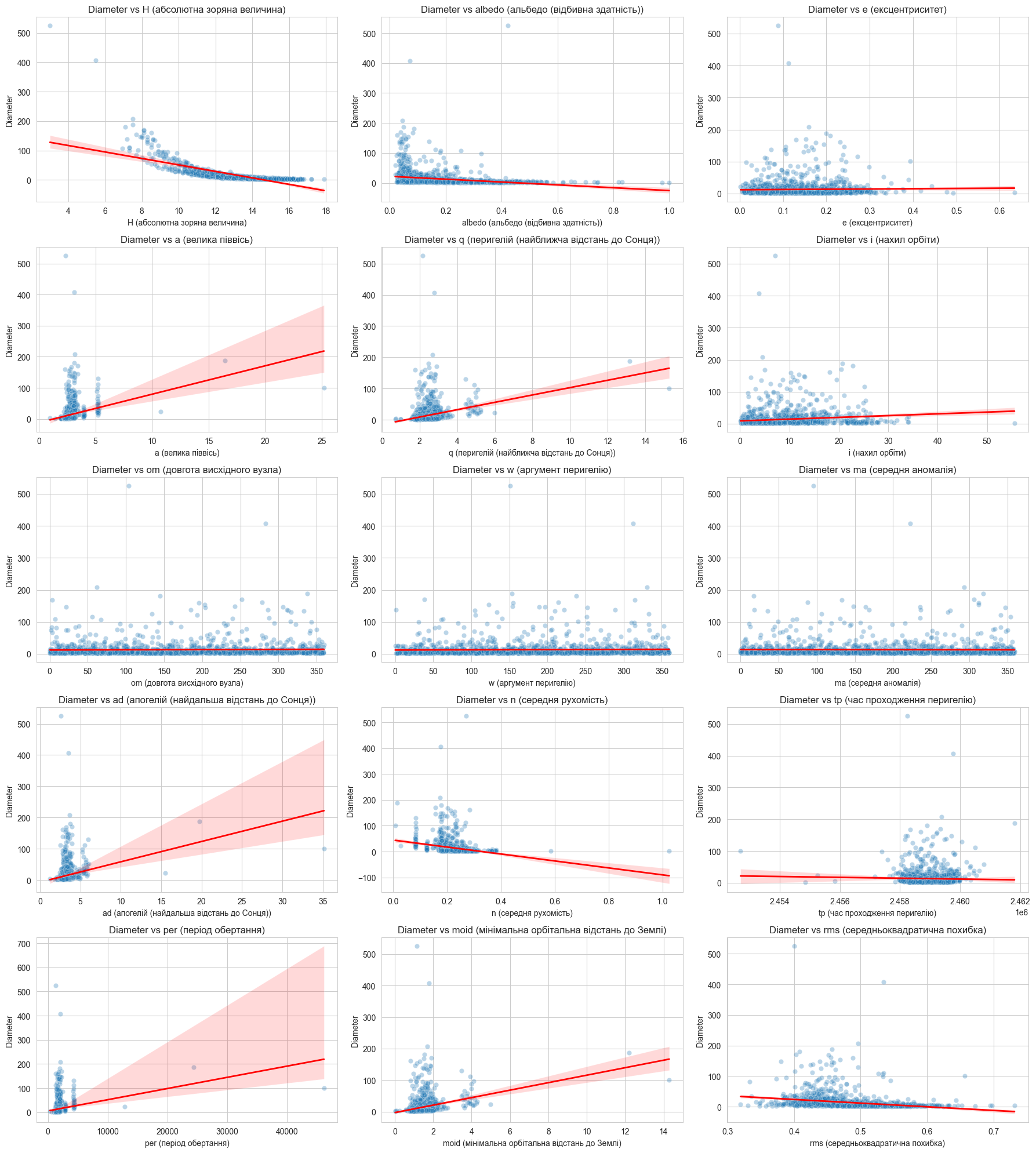
Кореляційна матриця показує, що орбітальні параметри і їхні невизначеності мають високий ступінь взаємозв'язку, що є характерним для фізичних характеристик астероїдів. Однак діаметр астероїда не має сильних кореляцій з іншими числовими змінними, що свідчить про обмежені можливості точного передбачення діаметра лише на основі орбітальних даних. Це може вимагати розширення моделі, включивши додаткові змінні або інші джерела даних.

**Гістограма розподілу діаметра астероїдів**



Цей графік показує розподіл діаметрів астероїдів в наборі даних. Спостерігається, що більшість астероїдів мають малі діаметри, і лише кілька мають великі діаметри. Графік показує експоненційний спад, де велика частина астероїдів зосереджена в межах малих діаметрів, з рідкісними більшими об'єктами.

Червона пунктирна лінія на графіку позначає середнє значення діаметра астероїдів у датасеті. Цей графік дає чітке уявлення про те, що розподіл діаметрів астероїдів сильно асиметричний, з переважною більшістю об'єктів із малими розмірами, що підкреслює важливість правильної обробки таких даних для прогнозування та моделювання.



На цих графіках показано, як діаметр астероїда змінюється залежно від різних фізичних і орбітальних параметрів. Ось деякі ключові спостереження:

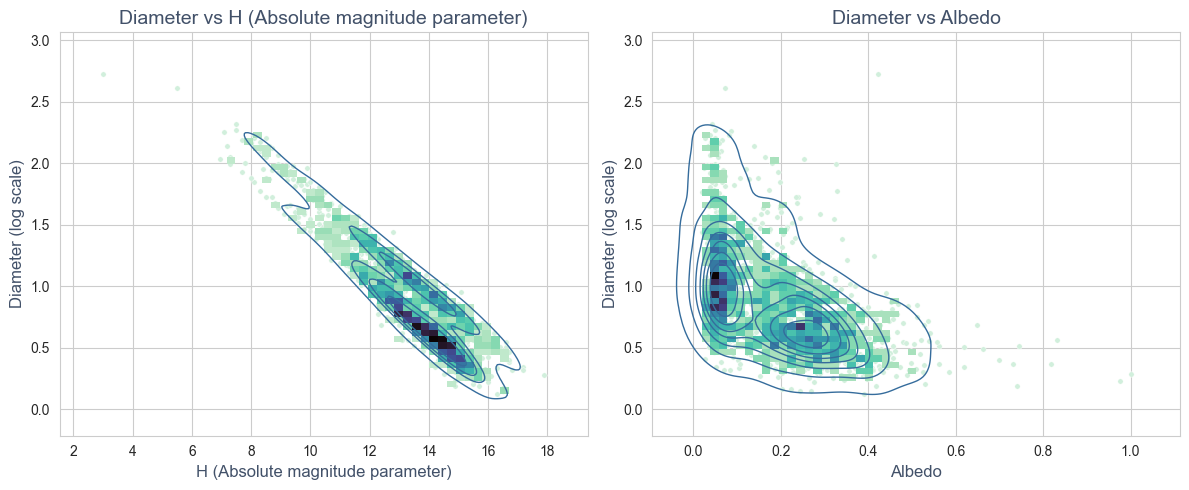
1. **Діаметр vs H (Абсолютна зоряна величина)**:
   * Спостерігається слабка позитивна кореляція між діаметром астероїда і його абсолютною зоряною величиною. Це означає, що астероїди з великими діаметрами мають, як правило, менші значення зоряної величини.
2. **Діаметр vs Albedo (відбивна здатність)**:
   * Між діаметром і відбивною здатністю (albedo) не спостерігається сильного зв'язку. Відбивна здатність астероїдів не є значним предиктором для їхнього діаметра.
3. **Діаметр vs Eccentricity (ексцентриситет)**:
   * Також спостерігається відсутність сильної кореляції між діаметром і ексцентриситетом орбіти астероїда. Ці параметри, схоже, не мають прямого зв'язку.
4. **Діаметр vs Orbital parameters (параметри орбіти: a, q, om, tp, tp\_cal, період обертання, moid)**:
   * Кілька параметрів орбіти, таких як велика піввісь (a), перигей (q), параметри ексцентриситету (om), не мають сильного кореляційного зв'язку з діаметром астероїда.
   * Ці графіки показують, що орбітальні параметри на кшталт величини періоду обертання (per) і мінімальної відстані до Землі (moid) можуть дещо корелювати з діаметром, хоча цей зв'язок слабкий.
5. **Діаметр vs sigma та інші невизначеності**:
   * Графіки, що зображують зв'язок діаметра з параметрами невизначеності (sigma\_\*), вказують на те, що невизначеності в даних, наприклад, для ексцентриситету або орбітальних параметрів, не суттєво впливають на оцінку діаметра астероїдів.



Графік показує залежність альбедо (відбивної здатності) від абсолютної зоряної величини (H) для астероїдів різного діаметру. Ось деякі ключові спостереження:

1. **Зменшення альбедо з ростом H**:
   * На графіку видно, що зі збільшенням абсолютної зоряної величини (H) значення альбедо загалом зменшується. Це може свідчити про те, що об'єкти з високим H мають меншу відбивну здатність.
2. **Вплив діаметра**:
   * Розміри точок на графіку відповідають діаметру астероїда. Помітно, що астероїди більшого діаметру мають меншу абсолютну зоряну величину (нижче значення H) і часто мають більші значення альбедо.
3. **Взаємозв'язок між діаметром, H та альбедо**:
   * Взаємозв'язок між діаметром, абсолютною зоряною величиною та альбедо є цікавим аспектом для подальшого дослідження. Більші астероїди можуть мати тенденцію до більшого альбедо, що вказує на їхні особливості поверхні або склад.

Отримані дані можуть бути використані для кращого розуміння фізичних властивостей астероїдів, а також для оцінки їхньої відбивної здатності залежно від розміру та яскравості.



Графік показує контурні діаграми для залежності діаметра астероїда від двох параметрів: абсолютної зоряної величини (H) та альбедо.

1. **Діаметр vs H (абсолютна зоряна величина)**:
   * З графіка видно, що між діаметром астероїда і його абсолютною зоряною величиною існує виражена залежність. Як правило, зменшення значення H (тобто, зменшення яскравості астероїда) супроводжується збільшенням діаметра астероїда.
   * Контури на графіку показують, що астероїди з більшими значеннями H (менша яскравість) мають більш варіативний діаметр. Це вказує на наявність астероїдів різних розмірів в одному діапазоні H.
2. **Діаметр vs Альбедо**:
   * У цьому графіку видно, що між діаметром та альбедо також існує залежність. Вищі значення альбедо (більш відбивна поверхня) частіше асоціюються з меншими астероїдами.
   * Контури відображають, що астероїди з високим альбедо мають схильність до менших діаметрів, тоді як астероїди з низьким альбедо можуть бути як маленькими, так і великими.

Ці графіки допомагають краще зрозуміти, як фізичні параметри астероїдів взаємопов'язані між собою, і можуть бути корисні для подальшого аналізу їхніх характеристик і прогнозування розмірів.

**Підсумок**

Загалом, я дійшов висновку, що діаметр астероїда має тісний зв'язок з його абсолютною зоряною величиною та альбедо. Ці характеристики можна використовувати для оцінки розміру астероїда за допомогою астрономічних спостережень. Дослідження також показало, що орбітальні параметри можуть бути корисними для класифікації та прогнозування властивостей нових астероїдів.

# **Етап підготовки даних (Feature engineering)**

**Обробка пропущених значень**

На цьому етапі я здійснив обробку пропущених значень у датасеті. Ось кроки, які я виконав:

1. **Обробка колонки 'name'**:
   * У цій колонці відсутні 290 значень. Я замінив ці пропущені значення на 'Unknown', оскільки вони не мають значення для подальшої обробки.
2. **Обробка колонки 'prefix'**:
   * Колонка 'prefix' містить 2000 пропущених значень. Оскільки ця колонка не несе корисної інформації для моделювання, я вирішив видалити її.
3. **Обробка колонки 'pha'**:
   * У колонці 'pha' 6 пропущених значень, що позначають потенційну небезпеку (значення 'Y' або 'N'). Я заповнив пропущені значення як 'N' (Not hazardous).
4. **Обробка колонки 'H' (абсолютна зоряна величина)**:
   * Для цієї колонки було заповнено пропущені значення медіанним значенням, оскільки це є ефективним способом обробки пропусків для числових даних.
5. **Обробка колонки 'albedo' (відбивна здатність)**:
   * Для колонки 'albedo' з 247 пропущеними значеннями я обчислив середнє значення для астероїдів, що мають схожі характеристики у колонці 'H', і використав це значення для заповнення пропусків.
6. **Обробка колонки 'diameter' (діаметр астероїда)**:
   * Для цієї ключової характеристики я застосував формулу для обчислення діаметра на основі значень 'H' та 'albedo'. Якщо значення 'H' або 'albedo' відсутнє, діаметр залишався пропущеним (NaN).
7. **Обробка колонки 'diameter\_sigma'**:
   * У цій колонці міститься 246 пропущених значень, які я залишив як NaN, оскільки вона містить похибки діаметра, і для передбачень не є критично важливою.
8. **Обробка колонки 'sigma\_\*' (похибки орбітальних параметрів)**:
   * Параметри, які починаються з 'sigma\_', мають похибки орбітальних вимірів. Пропущені значення були залишені як NaN, оскільки вони не можуть бути обчислені через відсутність точних формул.
9. **Обробка колонки 'moid' (мінімальна відстань до Землі)**:
   * Для цієї колонки з 6 пропущеними значеннями я обчислив середнє значення в групах, утворених на основі орбітальних параметрів. Заповнив пропуски середнім значенням по цих групах.

Ці кроки дозволили успішно заповнити пропущені значення та підготувати дані до подальшого аналізу та моделювання.

**Обробка пропущених значень**

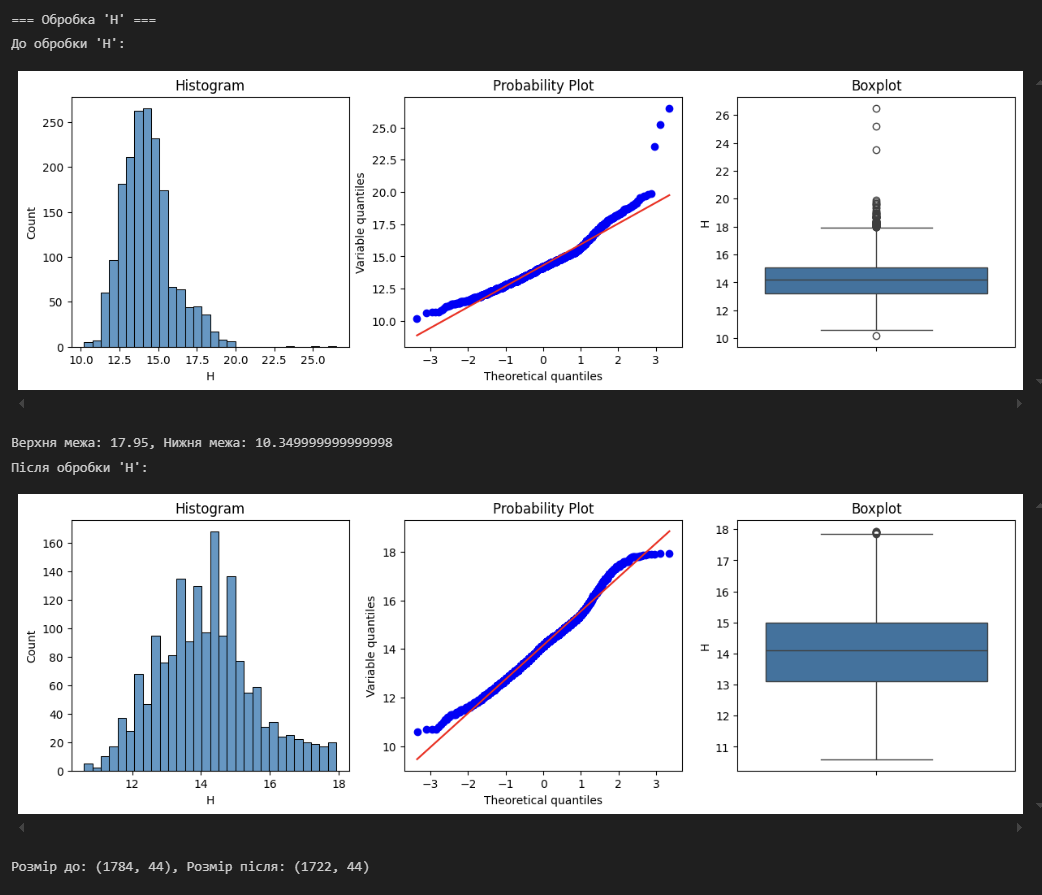
На етапі підготовки даних я виконав енкодинг категоріальних змінних для деяких колонок, таких як neo, pha, class та equinox, використовуючи методи **Label Encoding** для двійкових колонок та **One-Hot Encoding** для категоріальних змінних. Я використовував LabelEncoder для колонок з двійковими значеннями, такими як pha (позначення потенційної небезпеки для Землі), де значення були перетворені на числові.

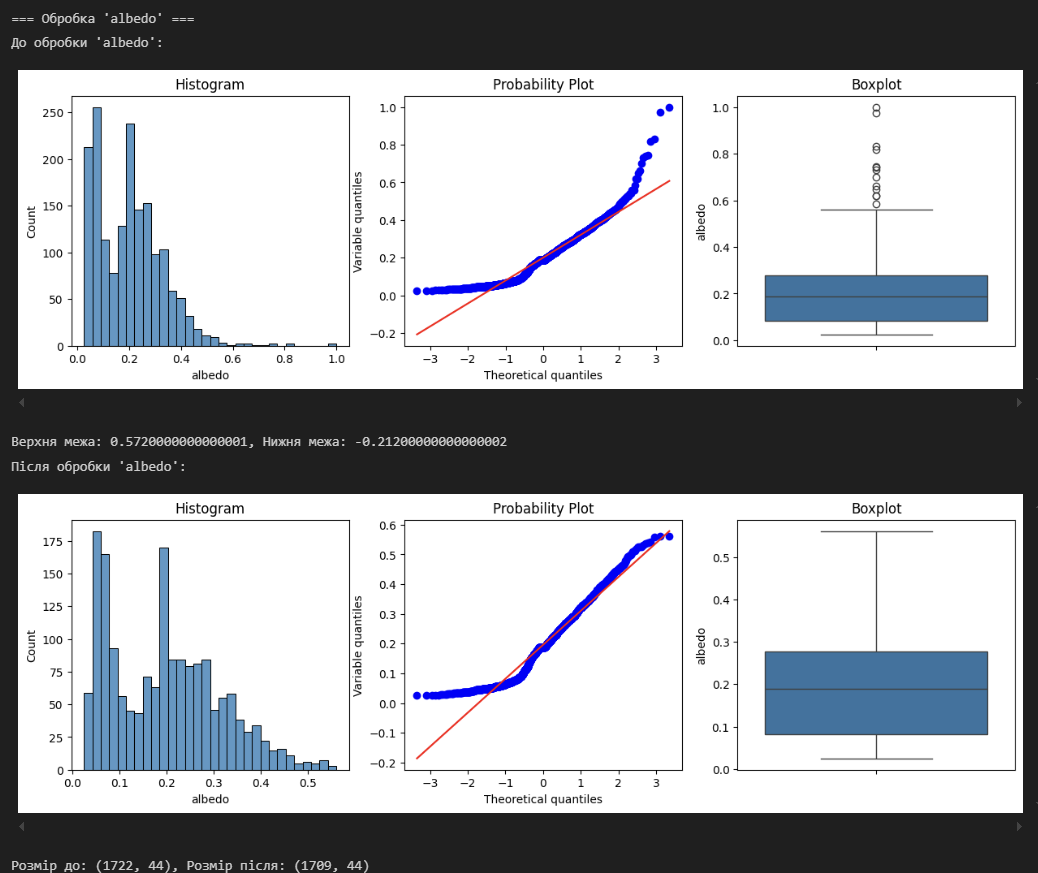
Згодом я також використав **One-Hot Encoding** для категоріальних колонок, таких як neo (позначення потенційних загроз), class (категорії астероїдів) і equinox, щоб перетворити їх у бінарні стовпці.

Проте, на подальших етапах роботи я з'ясував, що ці категоріальні ознаки не мали прямого впливу на модель для регресії, оскільки я працював з числовими ознаками, які впливали на цільову змінну (діаметр астероїда). Тому ці етапи кодування не знадобилися для моєї регресійної моделі.

**Обробка аномалій**







На етапі обробки аномалій я виконав кілька ключових кроків для виявлення та обробки аномальних значень у даних. Ось що я зробив:

1. **Визначення меж для аномалій**: Я застосував метод для визначення меж аномалій для числових змінних, таких як діаметр (diameter), абсолютна зоряна величина (H), альбедо (albedo), ексцентриситет (e), великі півосі (a), мінімальна відстань до Землі (moid) тощо. Для кожної змінної я обчислив межі, використовуючи метод міжквартильного розмаху (IQR), що дозволяє виявити аномальні значення, які знаходяться поза межами цих квантильних значень.
2. **Візуалізація до та після обробки**: Я створив графіки для візуалізації даних до і після обробки аномалій. Це включає гістограми, графіки ймовірностей і боксплоти для кожної з змінних. Графіки показують, як виглядають дані до обробки (з аномаліями) і після видалення аномальних значень.
3. **Видалення аномалій**: Після того, як були визначені межі для аномальних значень, я видалив рядки, які містили ці аномалії, застосувавши маскування на основі виявлених меж. Це дозволило очистити набір даних від аномальних значень і покращити якість даних для подальшого аналізу.
4. **Результати**: Після видалення аномалій розмір набору даних був зменшений, і я перевірив, що більше не залишилось значень поза межами допустимого діапазону.

Ці кроки дозволили покращити якість даних і підготувати їх до подальшого аналізу та побудови моделей.

**Масштабування ознак**

На етапі масштабування ознак я застосував два методи:

1. **Масштабування за допомогою StandardScaler**: Це нормалізація ознак для приведення їх до стандартного розподілу з нульовим середнім та одиничним стандартним відхиленням.
2. **Масштабування за допомогою MinMaxScaler**: Це перетворення значень ознак в діапазон [0, 1], що корисно для деяких моделей.

Я створив гістограми для кожної з ознак до та після масштабування, щоб перевірити, як змінились розподіли. Проте, після вибору найкращої моделі XGBoost для задачі регресії, я вирішив, що масштабування для цієї моделі не є необхідним, і тому не включав його в фінальну версію проекту.

# **Етап моделювання та вибір найкращої моделі, її оптимізації**

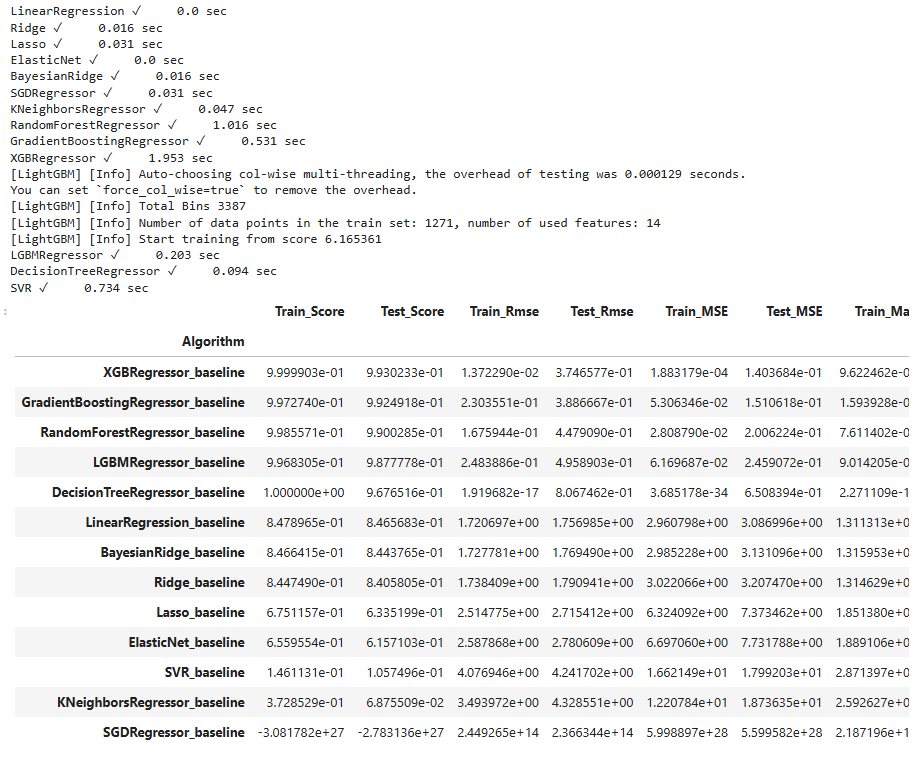
**Визначення цільової змінної, ознак та поділ датасету**

На етапі **визначення цільової змінної та ознак** я спочатку вибрав цільову змінну, якою є **діаметр астероїда** (diameter), а ознаки для моделювання були обрані з різних параметрів астероїдів, таких як H, albedo, e, a, q, i, om, w, ma, ad, n, per, per\_y, moid. Після цього я сформував матрицю ознак X та вектор цільових значень y, що відображає залежність діаметра від зазначених характеристик.

Після цього я здійснив **поділ даних на навчальний та тестовий набори**. Використовував стандартну функцію train\_test\_split, де 80% даних було використано для тренування, а решта 20% для тестування моделі. Це дозволяє перевірити модель на нових даних та оцінити її загальну ефективність.

В результаті, розмір тренувального набору становив 1271 приклади, а тестового — 318 прикладів.

**Вибір найкращої моделі**

****

На етапі **моделювання та вибору найкращої моделі**, для прогнозування діаметра астероїдів я використав різні регресійні алгоритми, такі як:

1. **Лінійні моделі**:
   * Лінійна регресія (Linear Regression)
   * Розширена лінійна регресія з L2-регуляризацією (Ridge Regression)
   * Лінійна регресія з L1-регуляризацією (Lasso Regression)
   * ElasticNet (комбінація Ridge та Lasso)
   * Байєсівська регресія (Bayesian Ridge)
2. **Деревоподібні моделі**:
   * Регресор на основі дерева (Decision Tree Regressor)
   * Ліс регресорів (Random Forest Regressor)
   * Градієнтний бустинг (Gradient Boosting Regressor)
   * Оптимізована версія градієнтного бустингу (XGBoost)
3. **Моделі на основі сусідів**:
   * Регресор на основі найближчих сусідів (KNeighbors Regressor)
4. **Підтримуючі векторні регресори**:
   * Підтримуючий векторний регресор (Support Vector Regressor)

Після побудови моделей, було проведено оцінку їхньої ефективності за допомогою метрик, таких як **Test\_Rmse**, **Test\_Mae**, **R²**, що дозволило оцінити точність і помилки кожної моделі.

Найкращою моделлю виявився **XGBRegressor\_baseline**, оскільки він продемонстрував найкращі показники:

* **Test\_Rmse**: 0.3746577
* **Test\_Mae**: 0.2132664
* **Test\_Score (R²)**: 0.9930323

Цей алгоритм був обраний через високу точність, низькі помилки та ефективність (швидкий час виконання).

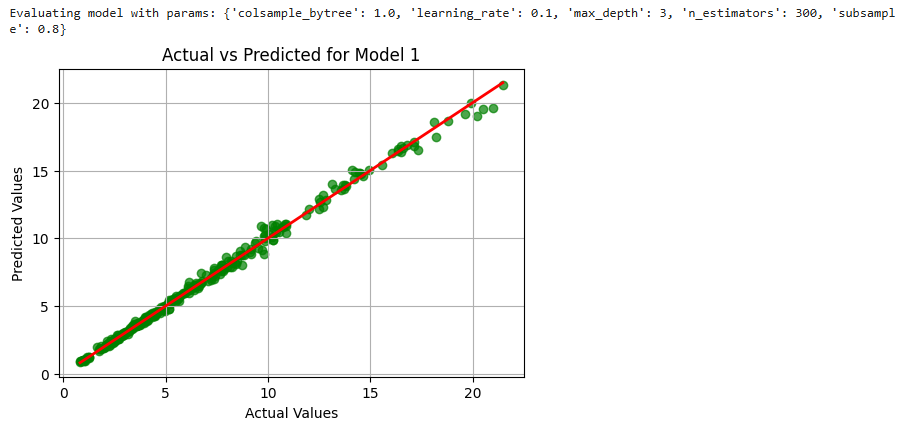
Інші моделі, такі як **GradientBoostingRegressor\_baseline** і **RandomForestRegressor\_baseline**, також показали хороші результати, але не мали такої ж високої точності та ефективності, як XGBoost.

Таким чином, для подальшого аналізу та прогнозування було обрано **XGBRegressor\_baseline**, оскільки він показав найкращі результати за всіма основними метриками.

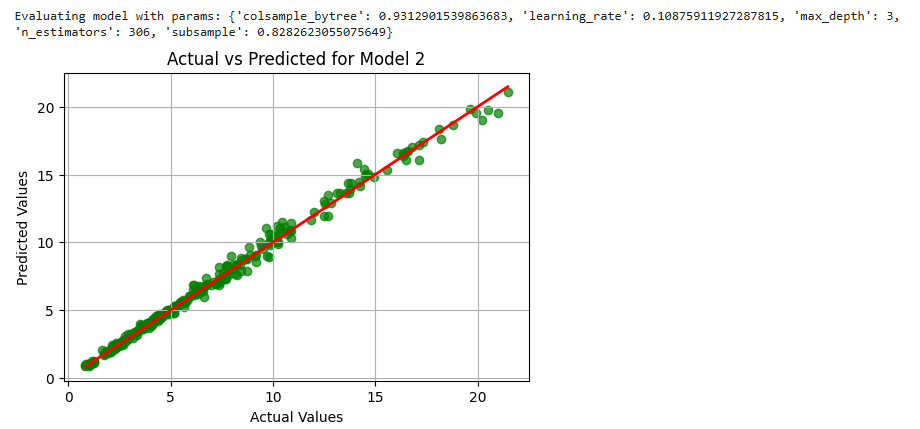
**Вибір найкращих гіперпараметрів**

Ось як виглядає етап вибору гіперпараметрів для моделі XGBoost:

1. **Опис гіперпараметрів**:
   * **n\_estimators**: Визначає кількість дерев у моделі. Більше дерев може покращити здатність моделі до узагальнення, але це також може призвести до перенавчання.
   * **max\_depth**: Максимальна глибина кожного дерева. Контролює, наскільки складним може бути дерево. Глибші дерева можуть моделювати складніші залежності, але можуть перенавчатися.
   * **learning\_rate**: Коефіцієнт навчання, який контролює, наскільки сильно кожне дерево змінює результати моделі. Низьке значення потребує більше ітерацій для досягнення стабільних результатів, але допомагає уникнути перенавчання.
   * **subsample**: Частка даних, яка використовується для навчання кожного дерева. Використання меншої частки даних може допомогти уникнути перенавчання і зробити модель більш стійкою до шуму.
   * **colsample\_bytree**: Частка ознак, яка використовується для навчання кожного дерева. Це дозволяє моделі більш ефективно працювати з різними підмножинами ознак.
   * **gamma**: Мінімальна необхідна різниця для поділу вузлів дерева. Більші значення змушують модель бути більш консервативною, що може знизити складність моделі.
   * **reg\_alpha** (L1-регуляризація): Контролює L1-регуляризацію на ваги, допомагаючи боротися з перенавчанням і змушуючи модель обирати лише найбільш важливі ознаки.
   * **reg\_lambda** (L2-регуляризація): Контролює L2-регуляризацію на ваги, допомагаючи зменшити значення коефіцієнтів і уникнути перенавчання.
   * **min\_child\_weight**: Мінімальна сума ваг у піддереві. Вищі значення зменшують складність моделі, змушуючи її ігнорувати вузли з невеликою кількістю ваг.
2. **Оптимізація гіперпараметрів**:
   * Використано **Grid Search** для визначення оптимальних гіперпараметрів, що дозволяє перебирати комбінації значень і знаходити найкращі.



* + **Random Search** також використаний для випадкового вибору гіперпараметрів у заданому діапазоні, що дає змогу швидше знайти хорошу комбінацію.

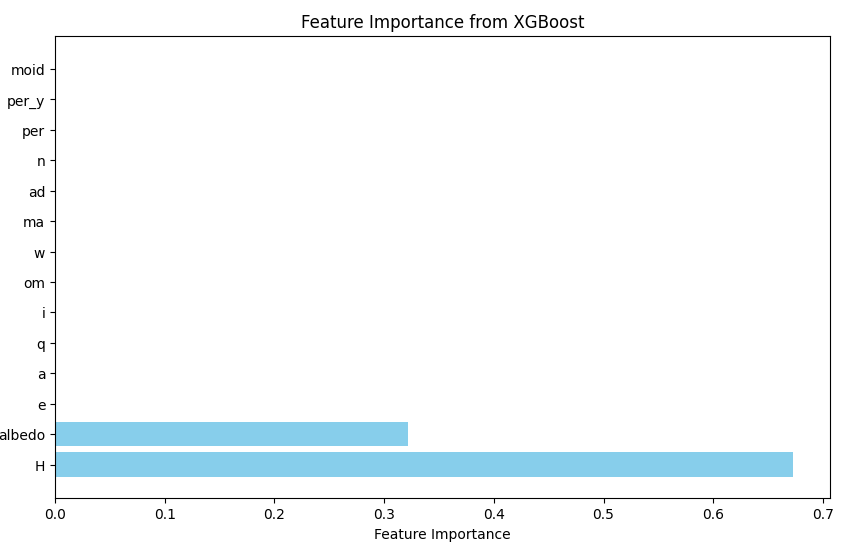


* + Для більш гнучкої оптимізації було застосовано **Bayesian Optimization**, який ефективно шукає оптимальні параметри, зменшуючи кількість перевірок, порівняно з іншими методами.



1. **Результати оптимізації**:
   * Найкращі параметри були знайдені через **Grid Search** з найменшими значеннями MSE, MAE, та RMSE, що вказує на високу точність.
   * Інші методи, як **Random Search** та **Bayesian Optimization**, також продемонстрували хороші результати, але з більшими часами виконання.

**Аналіз важливості ознак**

****

****

1. **Важливість ознак за моделлю XGBoost**:
   * Основними ознаками, які мають найбільший вплив на передбачення, є зоряна величина (H) і альбедо (albedo).
   * Значення H та albedo значно перевищують інші характеристики за важливістю, що свідчить про їхній ключовий внесок у передбачення діаметра астероїда.
2. **Permutation Importance**:
   * Техніка підтверджує, що H та albedo є найбільш важливими ознаками, оскільки їхнє виключення суттєво знижує точність моделі.
   * Серед менш важливих характеристик виокремлюються такі, як ексцентриситет (e) та мінімальна відстань до орбіти Землі (moid).
3. **Аналіз важливості для окремих прикладів**:
   * Результати Lime показують, що залежно від конкретного прикладу інші ознаки, як-от moid чи per\_y, можуть впливати на передбачення, але їхній вплив менший виражений у загальному аналізі.
   * Зоряна величина (H) нахиляється до домінуючої ознаки з негативним внеском, тобто зі зростанням її значення діаметр зменшується.
4. **Залежність між H, albedo і передбаченим діаметром**:
   * Візуалізація залежності показує, що із збільшенням H діаметр зменшується, що відповідає фізичній інтерпретації даних.
   * Альбедо (albedo) має складніший вплив, але зазвичай астероїди з більшим значенням альбедо мають менший діаметр.

**Висновки**:

* Найважливішими ознаками для передбачення діаметра астероїда є зоряна величина (H) та альбедо (albedo), які повинні бути обов'язково включені у модель для досягнення високої точності.
* Інші ознаки, такі як мінімальна відстань до орбіти Землі (moid) та ексцентриситет (e), мають менший, але помітний вплив на окремі передбачення.
* Використання кількох методів оцінки важливості ознак дозволяє підтвердити отримані результати та визначити ключові характеристики, що впливають на модель.

# **Етап розгортання моделі**

**Дерево проєкту**

RGR

├─ .git

├─ .gitignore

│

├─ configs

│  ├─ columns.py

│  ├─ model\_best\_hyperparameters.py

│  └─ \_\_pycache\_\_

│     ├─ columns.cpython-311.pyc

│     └─ model\_best\_hyperparameters.cpython-311.pyc

│

├─ data

│  ├─ comparison\_table.csv

│  ├─ new\_data.csv

│  ├─ new\_data\_with\_actual.csv

│  ├─ prediction\_results.csv

│  ├─ prime\_dataset.csv

│  └─ train\_data.csv

│

├─ docs

│  └─ ProjectStructure.md

│

├─ models

│  ├─ finalized\_model.sav

│  ├─ logs

│  │  ├─ prediction\_result

│  │  │  ├─ deviation\_scatter\_plot.png

│  │  │  └─ prediction\_result.png

│  │  └─ train\_model

│  │     ├─ cooks\_distance.png

│  │     ├─ error\_distribution.png

│  │     ├─ evaluation\_metrics.txt

│  │     ├─ feature\_importances.csv

│  │     ├─ feature\_importances.png

│  │     ├─ learning\_curve.png

│  │     ├─ residual\_plot.png

│  │     └─ y\_test\_vs\_y\_pred.png

│  └─ param\_dict.pickle

│

├─ pipeline

│  ├─ predict.py

│  └─ train.py

│

├─ README.md

├─ requirements.txt

│

├─ src

│  ├─ 1\_DataAnalysis.ipynb

│  ├─ 2\_DataProcessing.ipynb

│  ├─ 3\_ModelTraining.ipynb

│  ├─ filtered\_dataset.csv

│  └─ processed\_dataset.csv

│

└─ utils

   ├─ analyze\_results.py.ipynb

   ├─ plotting\_utils.py

   ├─ utils for dataset split and save.ipynb

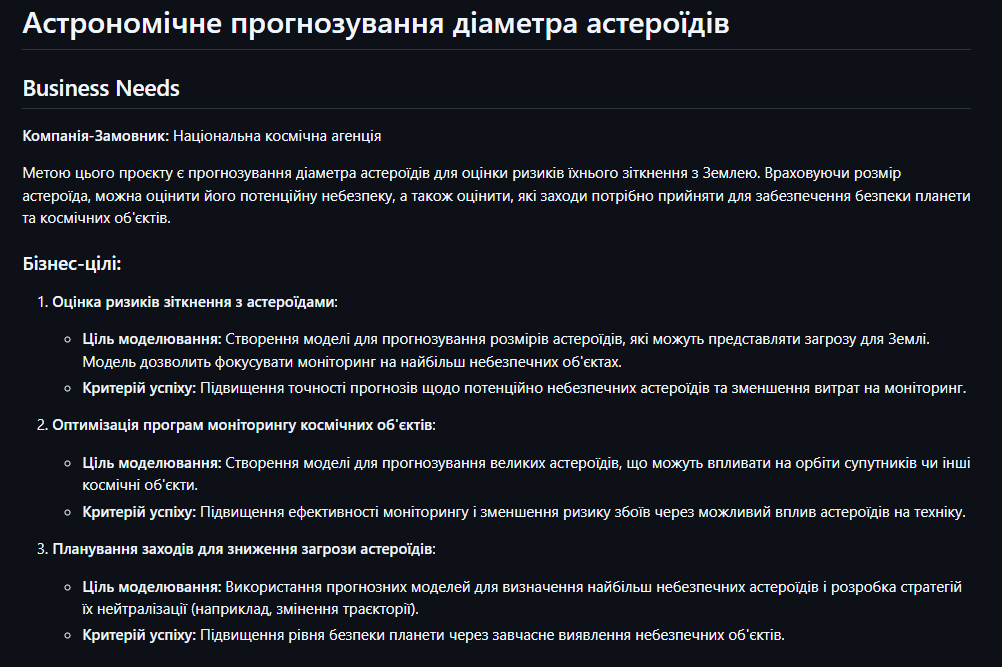
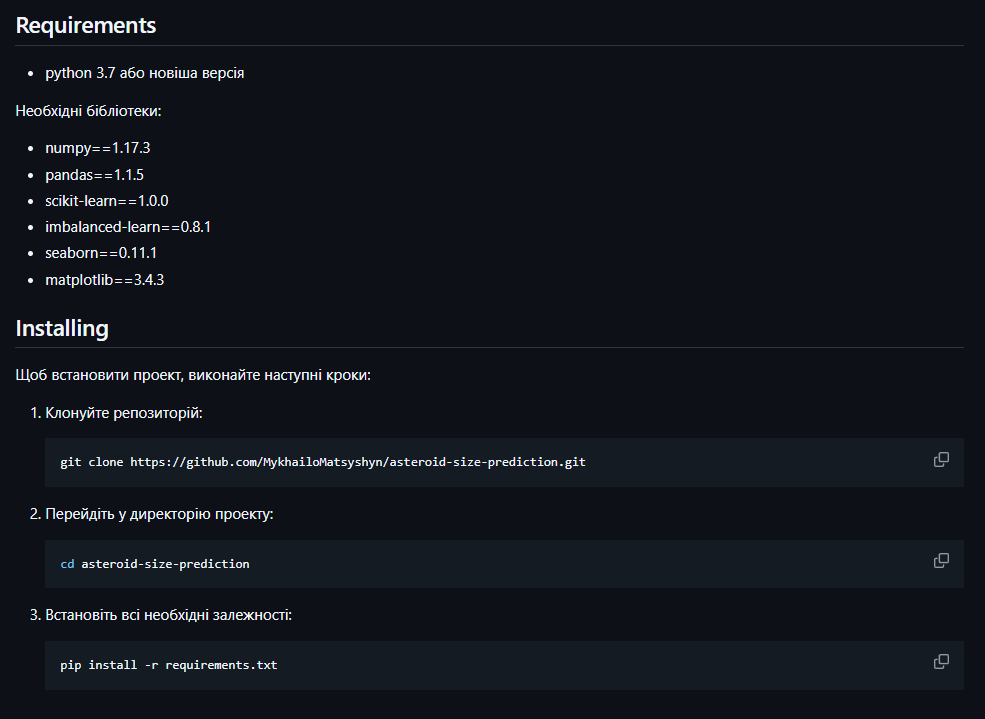
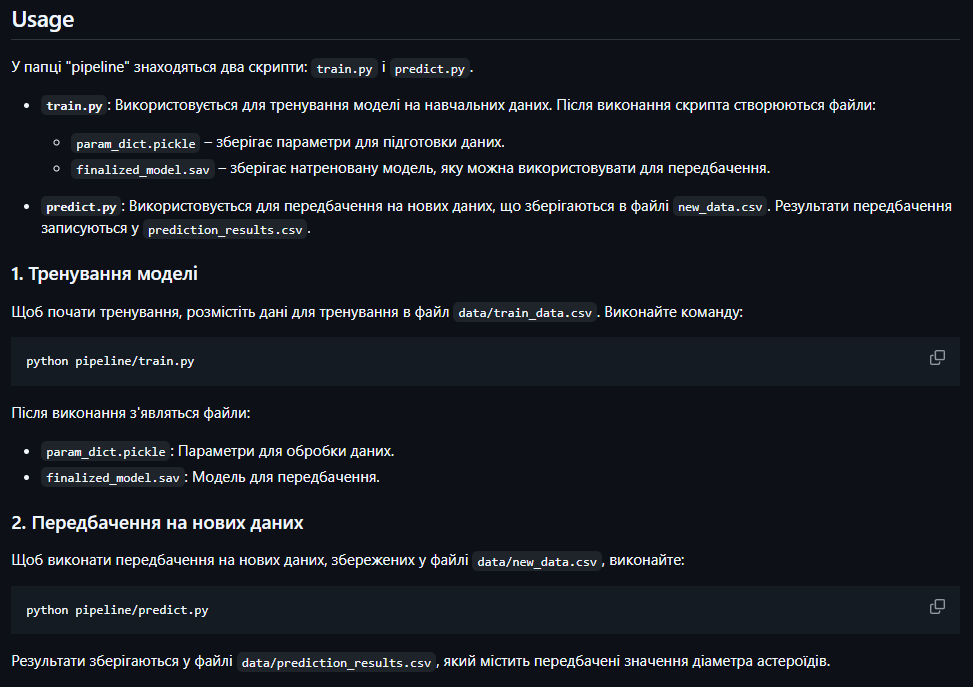
   └─ \_\_pycache\_\_

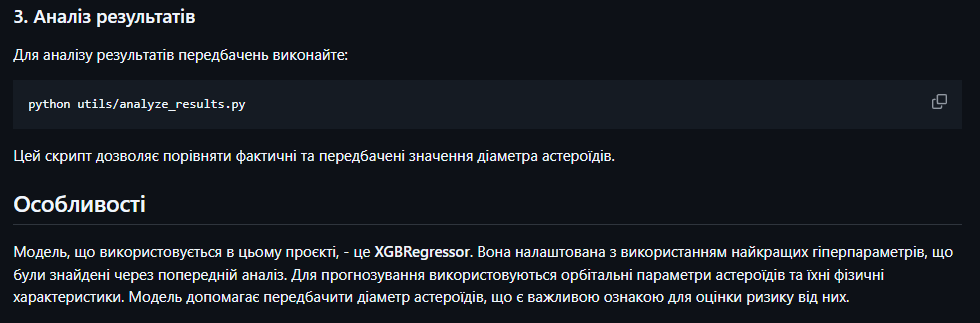
      └─ plotting\_utils.cpython-311.pyc

**Опис каталогів і файлів**

* **configs/**: Містить файли конфігурацій, що містять параметри для інженерії ознак і налаштування моделі.
  + columns.py: Містить інформацію про використані колонки у наборі даних.
  + model\_best\_hyperparameters.py: Гіперпараметри, використані для тренування моделі.
  + \_\_pycache/: Каталог з компільованими Python файлами.
* **data/**: Містить навчальні, тестові дані, а також файли з результатами.
  + comparison\_table.csv: Таблиця для порівняння фактичних і передбачених значень.
  + new\_data.csv: Нові дані для передбачення.
  + new\_data\_with\_actual.csv: Нові дані з доданими фактичними значеннями для порівняння.
  + prediction\_results.csv: Файл з результатами передбачення.
  + prime\_dataset.csv: Початковий набір даних з інформацією про астероїди.
  + train\_data.csv: Дані для тренування.
* **docs/**: Документація проекту.
  + ProjectStructure.md: Документація щодо структури проекту.
* **models/**: Містить треновану модель та результати.
  + logs/: Логи тренування моделі та візуалізації.
    - prediction\_result/: Логи для результатів передбачення.
      * deviation\_scatter\_plot.png: Графік відхилень між передбаченим і фактичним діаметром.
      * prediction\_result.png: Графік результатів передбачення.
    - train\_model/: Логи для тренування моделі.
      * evaluation\_metrics.txt: Метрики оцінки моделі.
      * feature\_importances.csv: CSV файл з важливістю ознак.
      * learning\_curve.png: Крива навчання.
      * residual\_plot.png: Графік залишків.
  + finalized\_model.sav: Файл з натренованою моделлю.
  + param\_dict.pickle: Параметри для обробки даних.
* **pipeline/**: Скрипти для тренування і передбачення моделі.
  + train.py: Скрипт для тренування моделі.
  + predict.py: Скрипт для передбачення на нових даних.
* **README.md**: Основна документація проекту з інформацією про використання, вимоги, і установку.
* **requirements.txt**: Файл із залежностями проекту для швидкої установки необхідних бібліотек.
* **src/**: Скрипти та файли для аналізу даних і тренування моделі.
  + 1\_DataAnalysis.ipynb: Jupyter Notebook для попереднього аналізу даних.
  + 2\_DataProcessing.ipynb: Jupyter Notebook для попередньої обробки даних.
  + 3\_ModelTraining.ipynb: Jupyter Notebook для тренування моделі.
  + filtered\_dataset.csv: Очищений набір даних.
  + processed\_dataset.csv: Оброблений набір даних.
* **utils/**: Утилітні файли та скрипти.
  + analyze\_results.py.ipynb: Jupyter Notebook для аналізу результатів передбачень.
  + plotting\_utils.py: Скрипт з функціями для візуалізації даних.
  + utils for dataset split and save.ipynb: Jupyter Notebook з функціями для поділу і збереження даних.
  + \_\_pycache/: Каталог з компільованими Python файлами.

**README.md:**

**** **** 

****

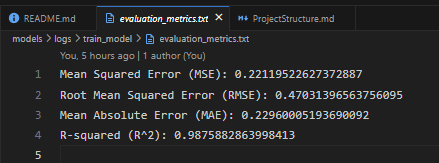
**Результати тренування моделі:**

1. **Оцінка моделі**

У файлі evaluation\_metrics.txt зафіксовано основні метрики, отримані після тренування моделі XGBoost:

* + **Mean Squared Error (MSE):** 0.22119522627372887
  + **Root Mean Squared Error (RMSE):** 0.47031956563756095
  + **Mean Absolute Error (MAE):** 0.22960005193609092
  + **R-squared (R²):** 0.9875882863998413

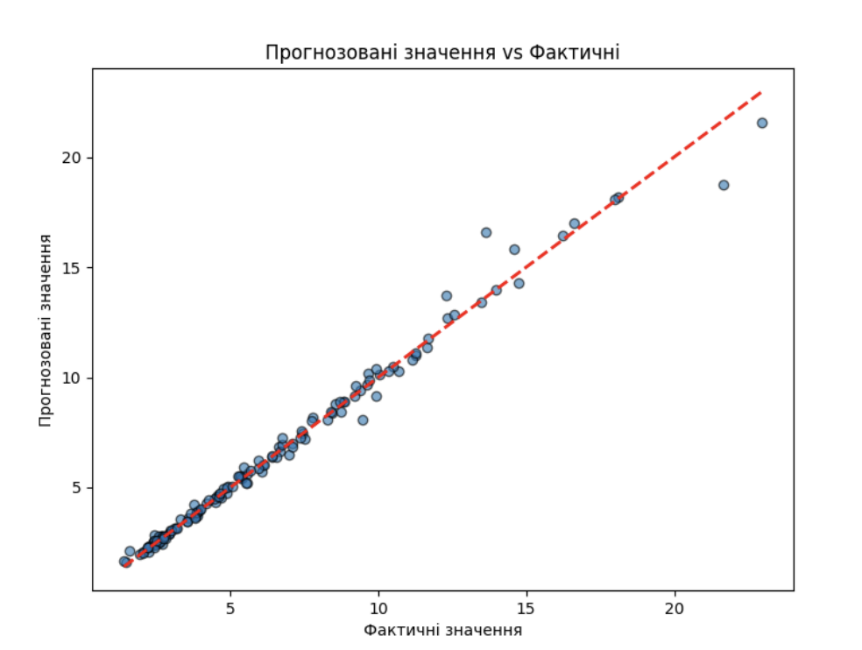
Ці значення підтверджують високу точність моделі, оскільки коефіцієнт детермінації R² наближається до 1, що вказує на хорошу здатність моделі пояснювати варіацію цільової змінної.



1. **Графік прогнозованих та фактичних значень**

На цьому графіку представлені прогнозовані значення проти фактичних для перевірки точності передбачень:

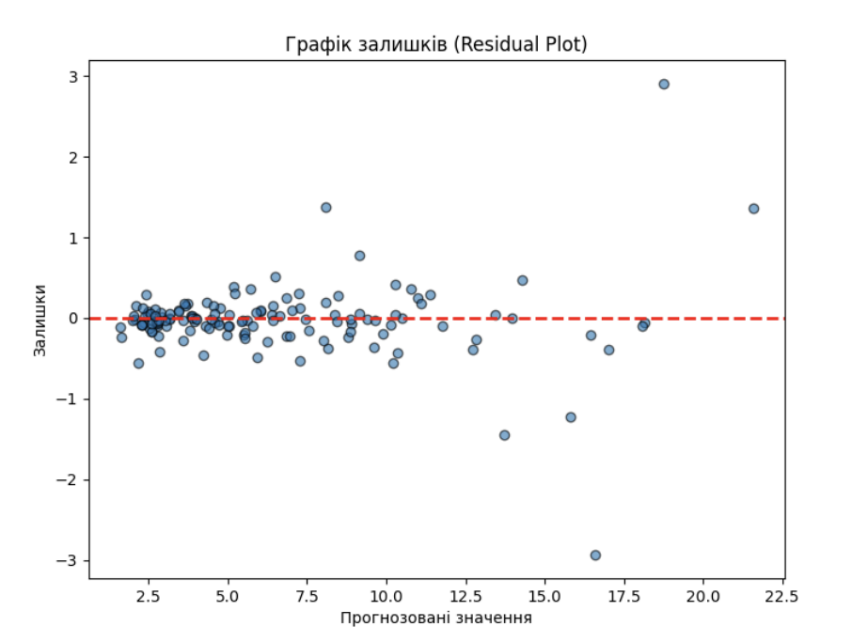
Графік демонструє, що модель добре відтворює фактичні значення, оскільки точки близькі до червоної лінії (лінія рівня точного прогнозу).



1. **Графік залишків**

Графік залишків показує різницю між фактичними та прогнозованими значеннями:

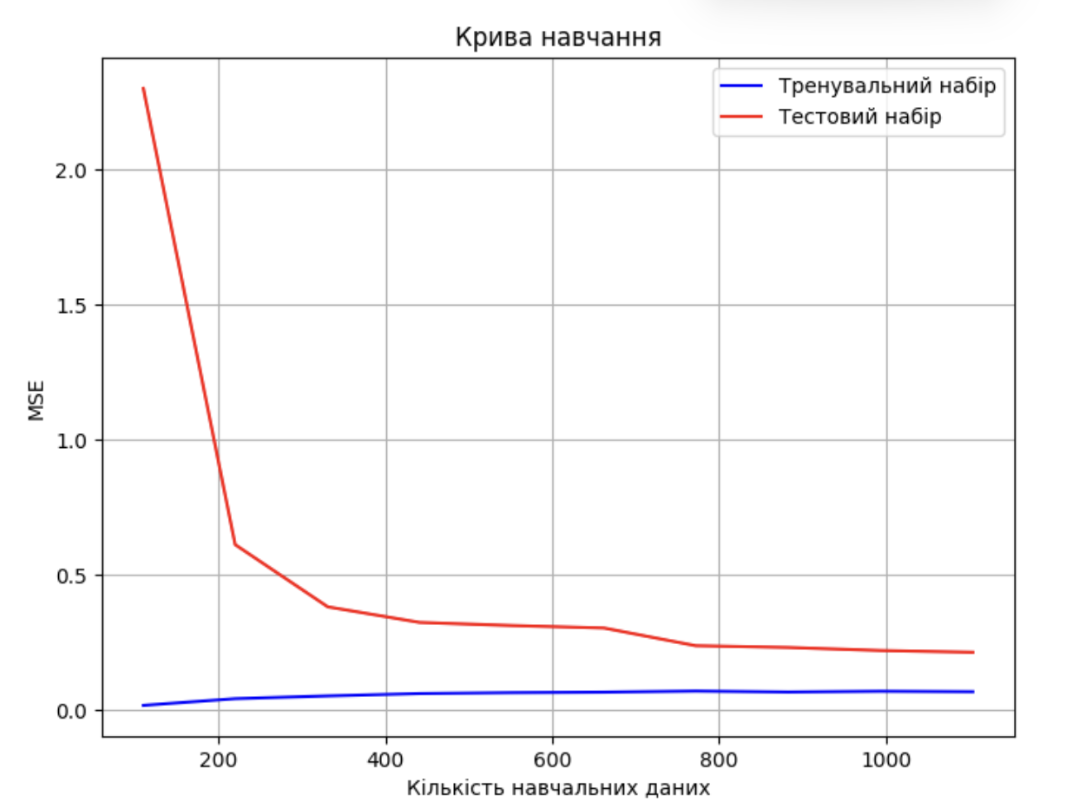
Залишки близькі до нуля без явних патернів, що свідчить про відсутність систематичних помилок і хорошу модель, яка не має перенавчання.



1. **Крива навчання**

Крива навчання показує, як змінюється значення MSE при збільшенні кількості навчальних даних:

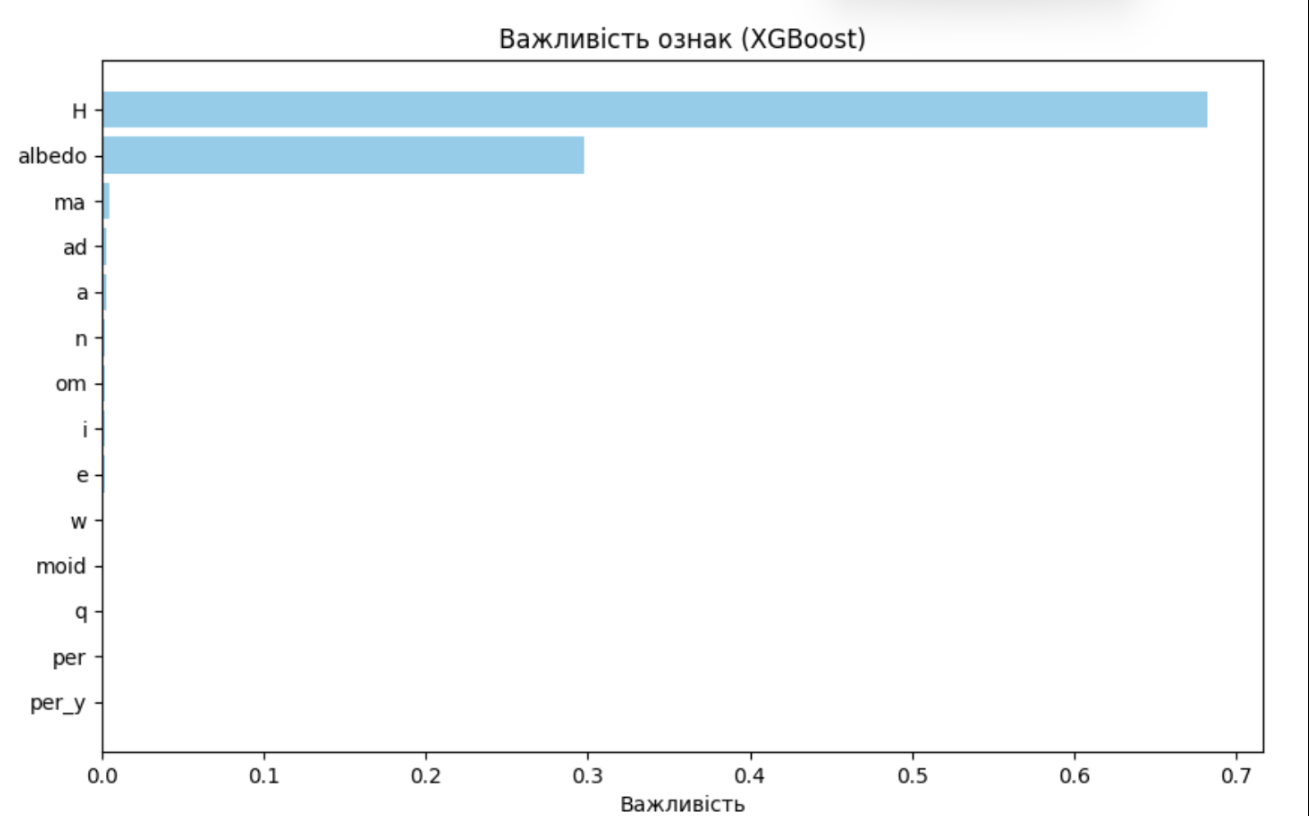
Це свідчить про зниження помилки в міру збільшення кількості даних, що підвищує загальну стабільність моделі.

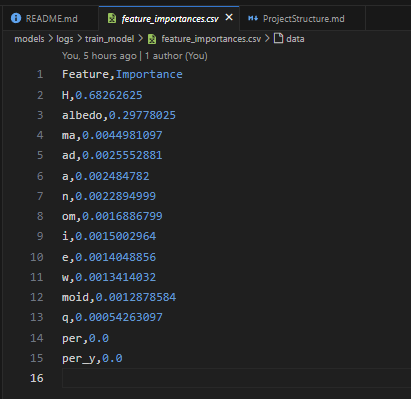


1. **Важливість ознак**

Графік важливості ознак показує, які змінні мають найбільший вплив на прогнозування діаметра астероїдів:

З графіку видно, що найбільш важливими ознаками є зоряна величина (H) та альбедо (albedo), що підтверджує їхній сильний вплив на прогнозування.

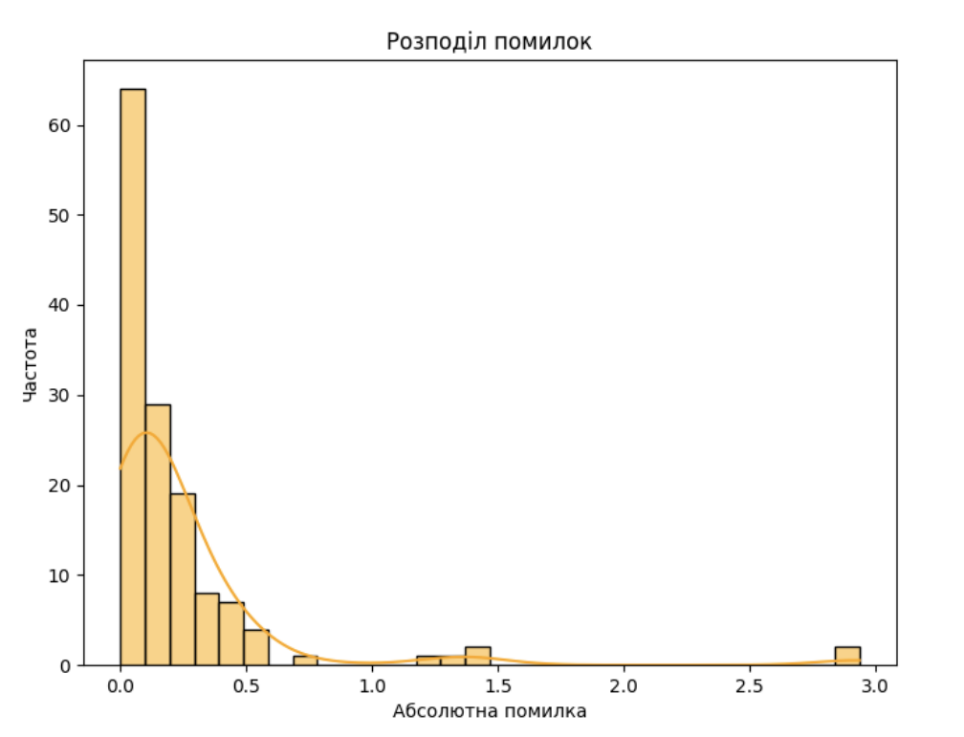




1. **Розподіл помилок**

Графік розподілу абсолютних помилок показує їх частоту:

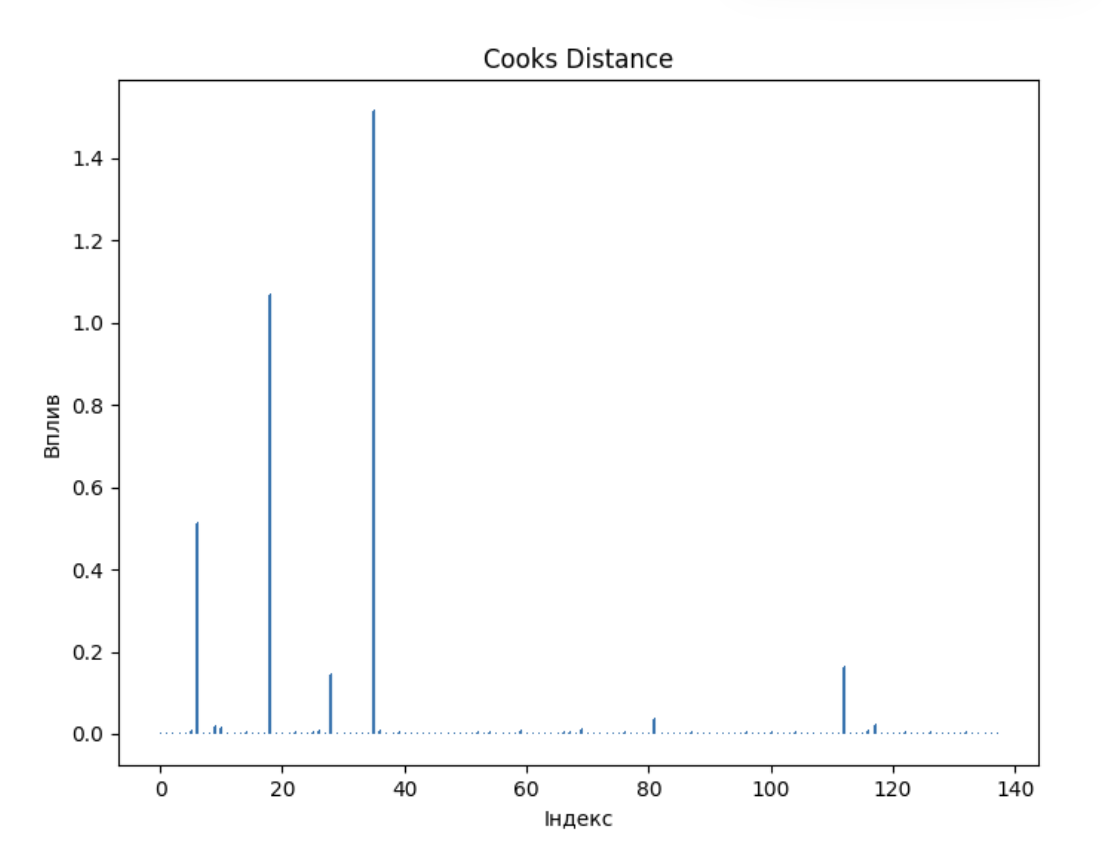
Більшість помилок знаходяться в межах малого значення, що вказує на високу точність моделі.



1. **Cooks Distance**

Графік Cook's Distance допомагає виявити впливові спостереження:

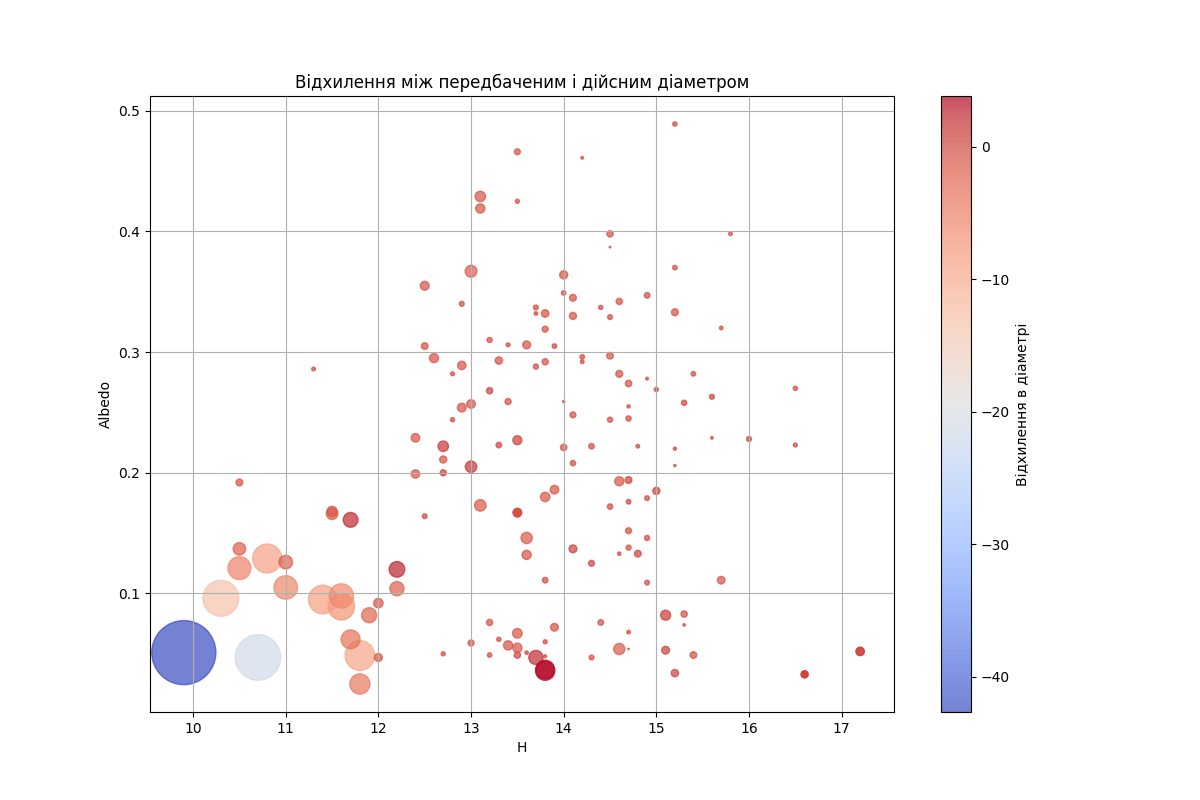
Показує, що декілька спостережень можуть мати вплив на модель, що варто врахувати при подальшій обробці даних.



**Висновки**

1. Модель XGBoost показала хороші результати з високою точністю (R² = 0.99) та низькими значеннями помилок (MSE, RMSE, MAE).
2. Найважливішими ознаками для передбачення діаметра астероїдів є зоряна величина (H) та альбедо (albedo).
3. Залишки і графік навчання підтверджують, що модель добре працює без перенавчання.
4. Подальша перевірка та обробка даних можуть бути необхідні для покращення результатів, враховуючи деякі потенційно впливові спостереження.

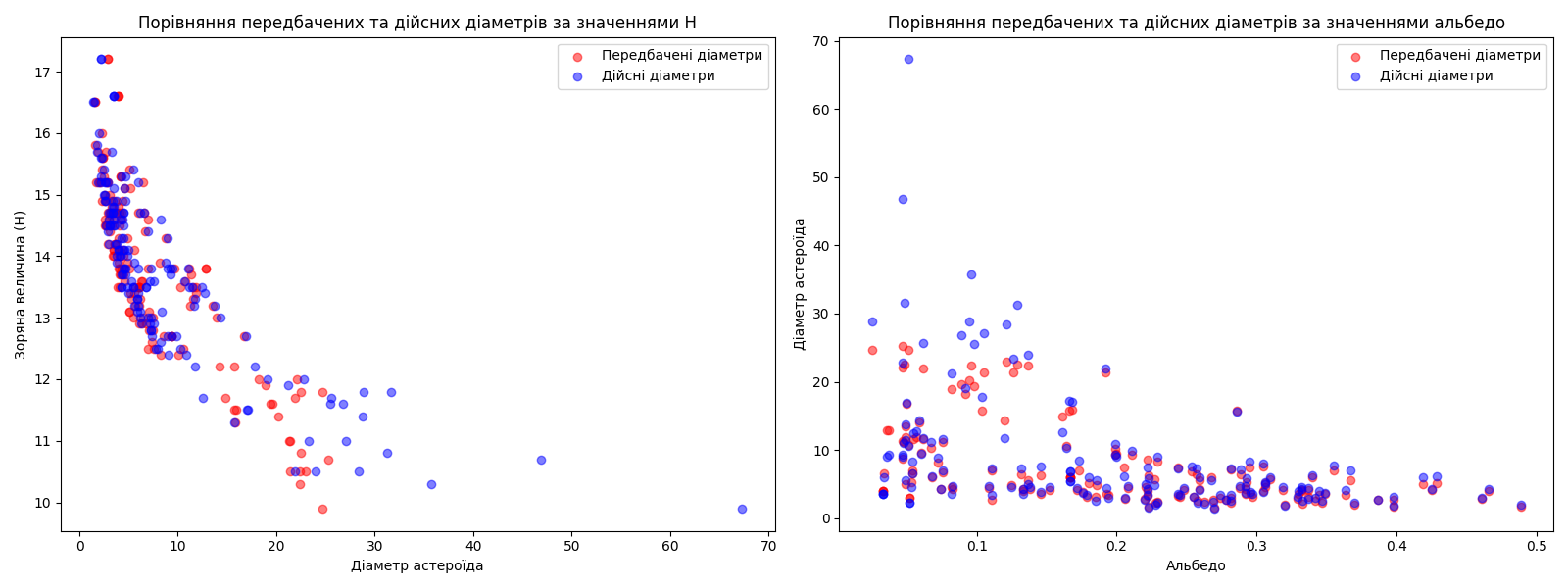
**Результати прогнозування моделі:**



Цей графік показує відхилення між передбаченим і дійсним діаметром астероїдів. Ось що важливо помітити на ньому:

1. **Ось X (H)** — це одна з ознак (можливо, зоряна величина), що використовується для передбачення діаметру астероїдів.
2. **Ось Y (Albedo)** — ще одна важлива ознака, що впливає на діаметр астероїда.
3. **Розмір точок** — відображає величину відхилення між передбаченим і дійсним діаметром. Великі точки означають більше відхилення.
4. **Колір точок** — відображає величину відхилення. Більш яскраві, червоні точки вказують на великі відхилення, а більш бліді сині точки — на менші відхилення.
5. **Колірна шкала** — праворуч від графіка є шкала, яка вказує, на скільки одиниць змінюється діаметр (в залежності від кольору точки). Червоні точки означають позитивні відхилення, а блакитні — негативні.

Таким чином, цей графік наочно показує, як різні значення H та albedo корелюють із відхиленнями у діаметрі. Великі відхилення з'являються там, де колір точок є найбільш яскравим (червоним або синім), що свідчить про значні розбіжності між передбаченими і дійсними діаметрами.



**Порівняння передбачених і дійсних діаметрів за значеннями H і Albedo**

Лівий графік показує порівняння передбачених і дійсних діаметрів за значеннями зоряної величини (H), а правий — за значеннями альбедо. На обох графіках видно, як передбачення моделі збігаються з фактичними значеннями, з деякими відхиленнями для значень з крайніх меж.

**Висновок:** Результати показують хорошу відповідність між передбаченими та дійсними діаметрами для більшості значень, зокрема для більшості астероїдів із середніми та високими значеннями H і альбедо. Для астероїдів з низькими значеннями цих ознак точність прогнозу дещо знижується.

# **ВИСНОВОК**

Для виконання цієї розрахункової роботи, спрямованої на створення регресійної моделі для прогнозування цільової змінної, я використав набір даних з Kaggle, що містить фізичні та орбітальні характеристики астероїдів. Задача полягала в прогнозуванні діаметра астероїдів, що є важливим показником для оцінки потенційних загроз для Землі. Модель була створена на основі числових ознак, таких як зоряна величина (H), альбедо (albedo) та орбітальні параметри астероїдів.

Процес розпочався з підготовки даних, включаючи очистку від пропущених значень та аномалій. Я застосував методи імпутації, щоб заповнити відсутні значення, та обробив аномальні дані за допомогою міжквартильного розмаху. Після цього я здійснив масштабування ознак, хоча для моделі XGBoost це не було необхідно. Вибір моделі базувався на кількох методах, і найкращою виявилась модель власне XGBoost через високу точність (R² = 0.99) та низькі значення помилок (MSE, RMSE, MAE). Оцінка важливості ознак показала, що зоряна величина та альбедо мають найбільший вплив на діаметр астероїдів.

Виконуючи цю роботу, я також навчився використовувати різні методи обробки даних, таких як заповнення пропусків, обробка аномалій і категоріальних змінних, а також вибір оптимальних гіперпараметрів для моделей машинного навчання. Крім того, я здобув досвід у використанні XGBoost для задач регресії, що дозволяє досягти високої точності прогнозів.

Складним етапом був вибір та налаштування гіперпараметрів для моделі XGBoost, оскільки необхідно було спробувати кілька методів оптимізації, таких як Grid Search і Random Search, щоб знайти найкращу комбінацію параметрів. Також завдання з обробки пропущених значень вимагало уважності, оскільки необхідно було обрати правильні методи для кожного стовпця, щоб не спотворити результати.

В цілому робота була цікавою та пізнавальною. Я отримав цінний досвід у застосуванні методів машинного навчання до реальних даних та розв'язанні складних задач прогнозування. Виконуючи роботу, я навчився ефективно підготовлювати дані для аналізу та моделювання, працювати з відсутніми значеннями та аномаліями, а також застосовувати алгоритми машинного навчання для вирішення практичних задач. Крім того, я здобув досвід у виборі найкращої моделі та оптимізації її гіперпараметрів для досягнення високих результатів.