Для всіх 4-х варіацій баз даних спостерігається схожа тенденція: при збільшені кількості вхідних параметрів – зменшується MSE, тобто збільшується точність прогнозування, так для $\boldsymbol{\varepsilon I_{sc}}$ **[База1]** маємо $\langle MSE \rangle = 0.0526, 0.0575, 0.0581$ (при трьох різних random_state), для випадку $\boldsymbol{\varepsilon I_{sc} + \varepsilon \eta}$ **[База2]** маємо $\langle MSE \rangle = 0.0307, 0.0317, 0.0338$, для випадку $\boldsymbol{\varepsilon I_{sc} + \varepsilon \eta + \varepsilon V_{oc}}$ **[База3]** маємо $\langle MSE \rangle = 0.0217, 0.0233, 0.0237$ і найточніші прогнози для випадку $\boldsymbol{\varepsilon I_{sc} + \varepsilon \eta + \varepsilon V_{oc} + \varepsilon FF}$ **[База4]**, а саме $\langle MSE \rangle = 0.0186, 0.0193, 0.0194$. Крім цього бачимо, що параметр num_candidate_attributes_ratio приймає в найкращих моделях 2 значення: 0.6 та 1.0. Якщо багато ФЕ параметрів – як правило 0.6, якщо мало – 1.0. Цей параметр може бути корисним для контролю перенавчання. Зменшення кількості ознак, що розглядаються під час кожного розбиття, може зробити модель менш чутливою до «шуму» і поліпшити її узагальнюючу здатність. Однак занадто низьке значення цього параметра може призвести до втрати інформації та погіршення продуктивності моделі. Треба підкреслити, що 0.6 та 1.0 це ліва та права границі інтервалу можливих гіперпараметрів, тобто ми розглядаємо [1.0, 0.8, 0.6], виходячи з цього, можна змінити простір цього гіперпараметра, наприклад, на [1.0, 0.9, 0.7, 0.6, 0.5] (дослідити окіл 1.0 та окіл 0.6). Цікаво що для найбільшої бази даних, не залежно від різних random_state, найкращі моделі завжди 26та, 31ша, 2га, 1ша та 6та. При зменшені кількості ФЕ параметрів такої тенденції не спостерігається. На мою думку, score vs step – важливий графік, який багато про що може нам розповісти. Враховуючи інтервали можливих параметрів:

tuner.choice("num_trees", [100, 150, 200, 250])

tuner.choice("max_depth", [10, 15, 20])

tuner.choice("min_examples", [5, 10, 15])

tuner.choice("num_candidate_attributes_ratio", [1.0, 0.8, 0.6])

Бачимо, що в **[База4]** найкращі параметри (150, 20, 5, 0.6) причому для всіх трьох random_state, в **[База3]** 2 рази зустрічається найкращий набір (150, 20, 5, 0.6) та один раз набір (150, 20, 5, 1.0), в **[База2]** вже відбувається «перетягування канату»: 2 рази найкращий набір (150, 20, 5, 1.0) та 1 раз (150, 20, 5, 0.6), що пов’язано зі зменшенням вибірки вхідних даних, в **[База1]** найкращими вибірками є (100, 15, 5, 0.6), (100, 20, 5, 1.0), (150, 20, 5, 0.6). Тобто тепер недостатньо використовувати замість 60% ознак 100% ознак, тепер сам вигляд дерев змінюється (num_trees та max_depths), при цьому min_examples всюди однакове – 5. Це наштовхує на думку, що потрібно змінити інтервал цього гіперпараметра, наприклад, на [3, 4, 5, 6, 7]. Також пропоную змінити інтервал для кількості дерев, наприклад, на [100, 130, 150, 170] та глибину дерев на [20, 25, 30]. При цьому можна сильно збільшити параметр max_trials до 100 або до 200.