**O datasetu:**

Dataset koji je koriscen sluzi za analizu energetske efikasnosti. Analiza se izvodi pomocu 12 razlicitih gradjevinskih oblika simuliranih u Ecotect-u. Zgrade se razlikuju u odnosu na oblast glazure, raspodelu glazure I orijentaciju, izmedju ostalih parametara. Simuliraju se razlicita podesavanja kao funkcije pomenutih karakteristika za dobijanje 768 gradjevinskih oblika.

Skup podataka sadrzi 768 uzoraka I 8 atributa, sa ciljem da se predvide dva odgovora (u ovom zadatku vrsena je analiza nad samo jednom vrednoscu).

Dataset se sastoji iz 8 atributa (oznacenih sa X1 do X2) I ima 2 vrednosti koje se predvidjaju (Y1 I Y2). Cilj je da uz pomoc tih 8 atributa predvidimo Y1.

Atributi su:

X1: Relativna kompaktnost

X2: Povrsinska oblast

X3: Zidna oblast

X4: Krovna oblast

X5: Ukupna visina

X6: Polozaj

X7: Oblast glazure

X8: Distribucija oblasti glazure

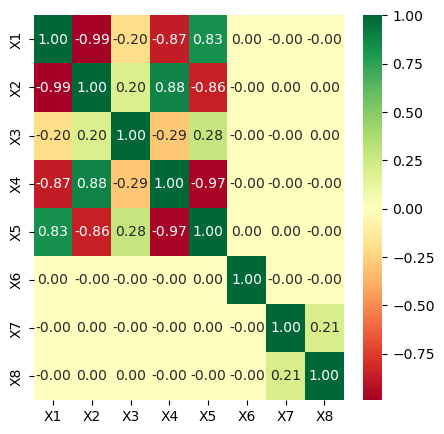
Y1: Opterecenje grejanja

Y2: Opterecenje hladjenja

**O radu:**

Model linearne regresije je pravljen u Jypiter Notebook-u, rucno. Vrsena je analiza nad dataset-om i utvrdjeno je da nema nedostajucih vrednosti, nema duplikata, sve vrednosti su kontinualne, tako da nije bilo potrebe za preciscavanjem podataka u samom pocetku. Kasnije je vrsena normalizacija atributa kako bi sve vredonsti atributa bile na istoj skali, buduci da je opseg vrednosti izmedju nekih atributa dosta razlicit (npr. minimalne I maksimalne vredosti za atribut X1 su 0.10577 I 0.98, dok su za npr X2 to 514.5 I 808.5). Time uvodimo stabilnost u treniranju modela.

Nakon normalizacije, izvrsena je redukcija dimenzionalnosti na osnovu medjusobne korelacije feature-a. Iz svakog para cija je apsolutna vrednost medjusobne korelacije veca od 0.95, izbacen je po jedan feature. U ovom slucaju, to su feature-i X2 i X5.



Ovde se zavrsava analiza i preprocesiranje podataka I pocinje treniranje modela. Algoritam Linearne regresije je implementiran uz pomoc metode najmanjih kvadrata.  Ovaj metod se koristi za određivanje pribliznog resenja nesaglasnog sistema linearnih jednacina

𝐴𝑥=𝑏

gde su 𝐴∈M𝑚×𝑛 matrica sistema punog ranga kolona, 𝑥∈ℝ𝑛 vektor nepoznatih i 𝑏∈ℝ𝑚 vektor slobodnih clanova (u ovom radu, b predstavlja tacne vrednosti zavisne varijable Y1 koju treba predvideti, A predstavlja feature-e, dok x predstavlja koeficijente regresije kojima se predvidja zavisna varijabla Y1). U slucaju kada je rang(𝐴)=𝑛<𝑚 manji od  rang([𝐴|𝑏])=𝑛+1≤𝑚  sistem 𝐴𝑥=𝑏 nema resenja. U takvoj situaciji upotrebljava se metod najmanjih kvadrata za pronalazenje pribliznog resenja 𝑥̂ koje ce minimizirati odstupanje ‖𝑏−𝐴𝑥‖2. To odstupanje ce biti najmanje kada je rezidualni vektor 𝑟=𝑏−𝐴𝑥 ortogonalan na prostor kolona matrice 𝐴, tj. kada je 𝐴𝑥 ortogonalna projekcija vektora 𝑏 na prostor kolona od A.

𝐴T𝑟=𝐴𝑇(𝑏−𝐴𝑥)=𝜃

𝐴𝑇𝐴𝑥=𝐴𝑇𝑏

𝑥ˆ=(𝐴𝑇𝐴)−1𝐴𝑇𝑏

𝑥ˆminimizira sumu greska u koordinatama.

Nakon treniranja modela I njegove primene, dobijeni rezultati su:

* Root mean squared error rucne implementacije: 0.82295
* Root mean squared error ScikitLearn implementacije: 0.89941