

# Predicting BMI Using Machine Learning



ECUTBILDNING

Quang Tri Tran

EC Utbildning

Projekt\_kurs

## Abstract

The goal of this project is to build a ML model that is capable of predicting Body Mass Index, or BMI for short, using key features such as: gender, weight and height. BMI serves as a important metric for categorizing individuals into groups of: underweight, normal weight, overweight, and obesity. in this project we conducted exploratory data analysis(EDA) to gain deeper insights into the dataset and preprocess the data by handling missing values and encoding categorical variables into numerals to insert into the ML model. The models used are four regression models: Linear regression, Decision tree regression, Random forest regression and XGBoost regression. Evaluation metric used to determine the performance of each model is Root mean squared error(RMSE). Among the models tested, the random forest model achieved lowest RMSE and was selected for further optimization.

This study demonstrates the use of machine learning for BMI prediction with high accuracy. Future enhancement would be to explore additional features such as age and physical activity levels to enhance our models prediction accuracy.

## Innehållsförteckning

|                                    |          |
|------------------------------------|----------|
| <b>Abstract.....</b>               | <b>2</b> |
| 1 Inledning.....                   | 1        |
| 2 Teori.....                       | 2        |
| 2.1 BMI.....                       | 2        |
| 2.2 Linear regression.....         | 2        |
| 2.3 Decision tree regression.....  | 2        |
| 2.4 Random forest regression.....  | 2        |
| 2.5 Xgboost regression.....        | 2        |
| 3 Metod.....                       | 3        |
| 3.1 Datainsamling.....             | 3        |
| 3.2 Utforskning av Data & EDA..... | 3        |
| 3.3 Modell.....                    | 3        |
| 3.4 Agil arbetsmetodik.....        | 3        |
| 4 Resultat.....                    | 5        |
| 5 Slutsats.....                    | 5        |
| 6 Självutvärdering.....            | 6        |
| Appendix A.....                    | 7        |
| Källförteckning.....               | 11       |

# 1 Inledning

I detta projekt, med hjälp av maskininlärnings modeller kommer vi att förutsäga BMI på individer baserad på information från individens:

- Kön
- Längd
- Vikt

BMI har olika upplägg av skala men i detta projekt kommer vi använda följande skala: Undervikt, normalvikt, övervikt, fetma klass 1, fetma klass 2 och fetma klass 3. Regressionsmodeller kommer att användas och tävla mot varandra för att ta reda på vilken modell som passar bäst för just denna projekt.

Syftet med denna rapport är att bygga en prediktions modell som kan förutse en individs BMI med god träffsäkerhet. För att uppfylla syftet kommer följande frågeställning(ar) att besvaras:

1. Undersökning av vilka faktorer som har påverkan på modellens resultat.
2. Hur presterar olika regressionsmodeller i förhållande till varandra vid BMI prediktionen?

## 2 Teori

### 2.1 BMI

BMI är en indikator för kroppens hälsa och visar förhållandet mellan längd och vikt, detta resultera sedan till en siffra som indikera var i BMI skalan personen befinner sig, skalan kan man ser nedan under BMI-Intervall.

$$BMI = Kilogram \div (Meter)^2$$

$$BMI = vikt \div (längd)^2$$

| BMI-Intervall     | Kategori        |
|-------------------|-----------------|
| BMI <18.5         | Undervikt       |
| 18.5 < BMI < 24.9 | Normalvikt      |
| 25.0 ≤ BMI < 29.9 | Övervikt        |
| 30.0 ≤ BMI < 34.9 | Fetma klass I   |
| 35.0 ≤ BMI < 39.9 | Fetma klass II  |
| BMI ≥ 40.0        | Fetma klass III |

### 2.2 Linear regression

Modellen beskriver sambandet mellan en beroende variabel (i vårt fall Index) och en eller flera oberoende variabler( som vikt och längd) genom en linjär relation:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \epsilon$$

### 2.3 Decision tree regression

En beslutsträd modell som segmenterar datamängden i mindre grupper genom uppdelningar baserade på olika villkor för att generera sina förutsägelser.

### 2.4 Random forest regression

Ensemblemodell som sammanför flera beslutsträd, där varje träd tränas individuellt och den slutliga prediktionen beräknas som medelvärdet av samtliga trädens förutsägelser.

### 2.5 Xgboost regression

Ensemblemodell som konstrueras av flera små beslutsträd. Modellen förbättras genom att lära sig från sina tidigare fel av varje nytt träd . Prediktionen sker genom att successivt addera flera svaga prediktorer, detta gör att modellen får en bättre träffsäkerhet och är mer robust.

## 3 Metod

### 3.1 Datainsamling

Dataseten är offentlig och är ursprungligen insamlad från Kaggles hemsida.

I denna dataseten har vi kolumnerna, kön, längd, vikt, BMI-index.

Det finns en ny kolumn "bmi" inlagd, den skapades genom att använda bmi formeln.

För att kunna mata in informationen senare i maskininlärnings modeller så behövdes kolumnen "kön" omvandlas till numerisk värde där: 0 representera män och 1 representera kvinnor.

### 3.2 Utforskning av Data & EDA

Datan utforskades och följande visualisering av datan togs fram:

1. Könsfördelning
2. BMI-fördelning
3. Förhållanden mellan variablerna

För att se visualiseringen, kolla in Appendix A (längst ned).

### 3.3 Modell

Modeller som används i arbetet:

- Linear regression
- Decision tree regression
- Random forest regression
- XGBoost regression

Efter jämförelse så valdes Random forest ut som bästa modellen för projektet då den fick lägsta RMSE värde.

Datasetet uppdelades i träning och test. Där 70% är träningsdatan och 30% är testdata.

Vår data är förberedd med StandardScaler och kategoriska data är omvandlad till numeriska.


RMSE och korsvalidering användes som bedömningsmatriser till varje modell.

Ett försök att hitta bästa inställningen på valda modellen utfördes.











Modellen är sparad och återanvändningsbar.

### 3.4 Agil arbetsmetodik


Vi följde en agil metodik där vi inledningsvis undersökte relevanta dataset och formulerade problemställningar innan vi påbörjade utvecklingen och testandet av modellen. Kontinuerliga förbättringar infördes stegvis under projektets gång. Om vi stötte på hinder eller osäkerheter, markerades dessa som ärenden och diskuterades gemensamt för att hitta lösningar till eventuella problem.





 **EC\_DS23\_Projekt\_kurs**  
Programvaruprojekt

PLANERING

-  Sammanfattning
-  Tidslinje
-  Backlog
-  **Tavla**
-  Kalender
-  Lista
-  Formulär
-  Mål
-  Ärenden
-  + Lägg till vy




UTVECKLING

-  Kod

-  Projektsidor
-  Lägg till genväg
-  Projektinställningar
-  Arkiverade ärenden **NY**

Projekt / EC\_DS23\_Projekt\_kurs

## KAN-Task board

   Epic ▾

TO DO


+ Skapa ärende

IN PROGRESS

REVIEW


DONE 13 ✓

Save and testing model


☒ KAN-23 ✓ 

Hitta projekt


PROJEKT\_KURS

☒ KAN-3 ✓ 


Formulera problem & mål

☒ KAN-12 ✓ 


EDA

☒ KAN-13 ✓ 


Preprocess data

☒ KAN-14 ✓ 


Model building - Linear regressor

☒ KAN-15 ✓ 

Model building - Decision tree regressor

☒ KAN-16 ✓ 

Model building - Random Forest regressor

☒ KAN-17 ✓ 

## 4 Resultat

| Regressionsmodeller     | RMSE  |
|-------------------------|-------|
| Linear Regression       | 0.557 |
| Decision Tree Regressor | 0.436 |
| Random Forest Regressor | 0.332 |
| XGBoost Regressor       | 0.377 |

| RMSE med optimala parametrar (Random forest) |       |
|--|-------|
| Träning                                      | 0.343 |
| Korsvalidering                               | 0.425 |

| RMSE för test data      |       |
|-------------------------|-------|
| Random Forest Regressor | 0.330 |

| Predikerat BMI-index |       |
|----------------------|-------|
| Man, 180cm, 80kg     | 2.129 |

## 5 Slutsats

Från Resultatet kan vi se att fyra olika modeller utvärderades, och Random forest modellen valdes eftersom den uppvisade bäst prestanda med ett RMSE-värde på 0.332, bäst både på träningsdata och testdata. vilket gör den till det mest pålitliga alternativet för BMI-prediktionen i detta projektet. För att ytterligare förbättra modellens noggrannhet genomfördes hyperparameter justering och resultatet gav:

- Träning: 0.343
- korsvalidering: 0.425
- Test: 0.330

Resultatet ovan tyder att modellen ger bra generalisering då det inte är en stor skillnad mellan resultaten, men det finns en viss variation mellan tränings- och korsvalideringsresultaten som indikerar att modellen har potential att ytterligare förbättras, vilket kan förbättra generaliseringen och minska risken för överanpassning.

Projektet visar en god implementering av maskininlärning för BMI-prediktion och erbjuder ett effektivt verktyg för snabb bedömning av BMI.



Analysen av modellens prediktorer visade att vikt och längd hade störst påverkan på resultatet, medan kön visade ingen påverkan.

Förbättringar i framtiden skulle vara att utforska mer variabler som ålder och fysisk aktivitet för att förbättra modellens prediktions prestanda. bl.a samla mer data och fortsätta hypertuning för att hitta bättre optimala parametrarna.

## 6 Självutvärdering

1. Utmaningar du haft under arbetet samt hur du hanterat dem.  
Största utmaningen var att komma på projekt att arbeta med. Det gäller att ha tålamod och fortsätta diskutera och leta fram ett gemensamt projekt.
2. Vilket betyg du anser att du skall ha och varför.  
en stark G men helt klart vill man ha VG, däremot är jag osäker om jag har träffat alla punkterna för VG. Punkterna som jag är osäkra på är:

### **2.Motivera val av valda metoder och tekniker i det praktiska projektet.**

Lite osäker hur denna ska tolkas men jag förstår detta som varför just modellen vi har valdes istället för dem andra och det är pga av RMSE resultatet. "tekniker i det praktiska projektet" Jag gissar ni menar då kors-valideringen och hypertuning av hyperparametrar som användes för att förbättra modellens prestanda, tänker mig det anses som typ av teknik?

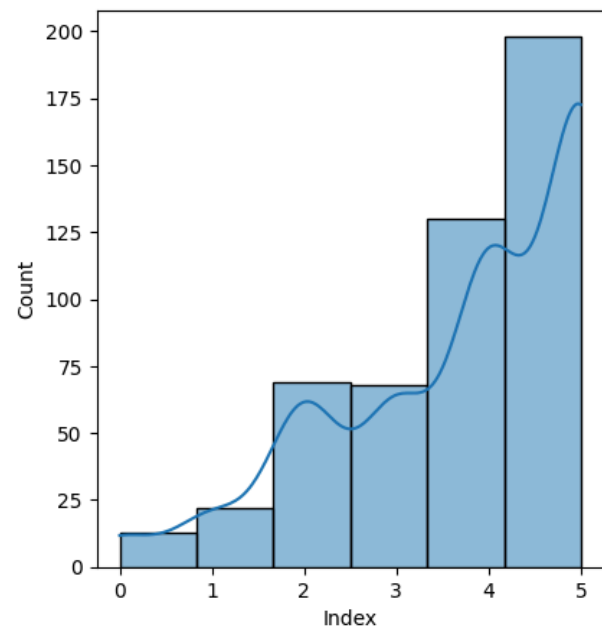
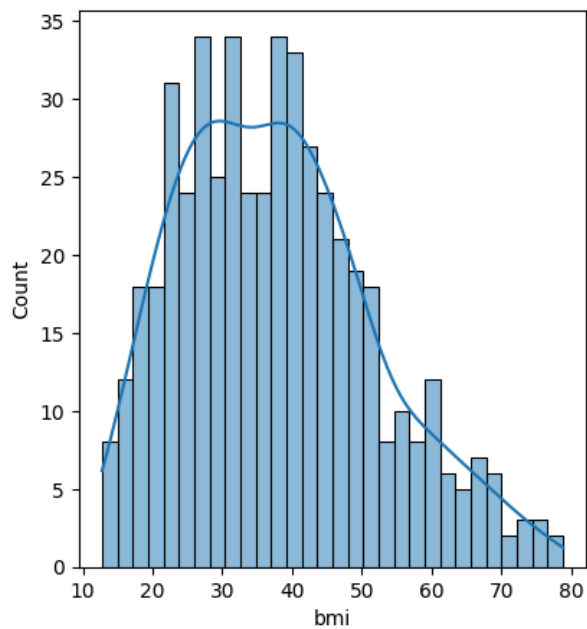
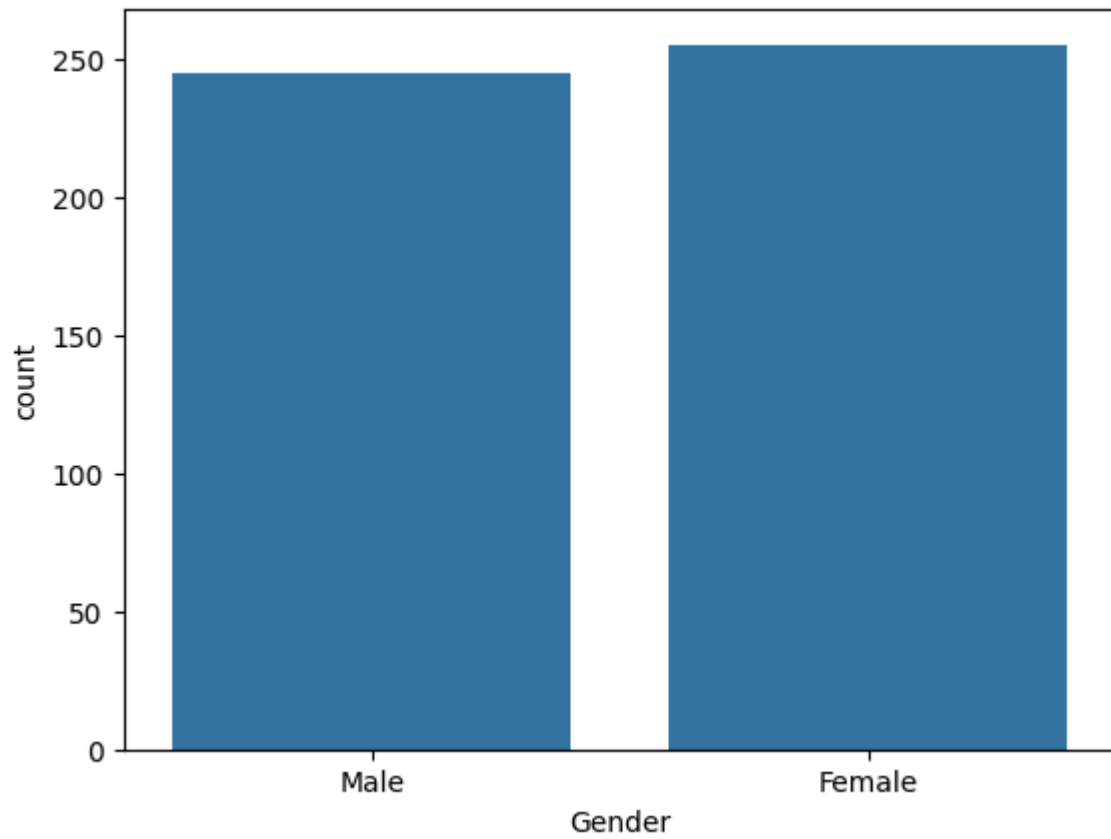
### **6. Skapa olika modeller för ett use case, utvärdera dessa och argumentera för en lämplig modell.**

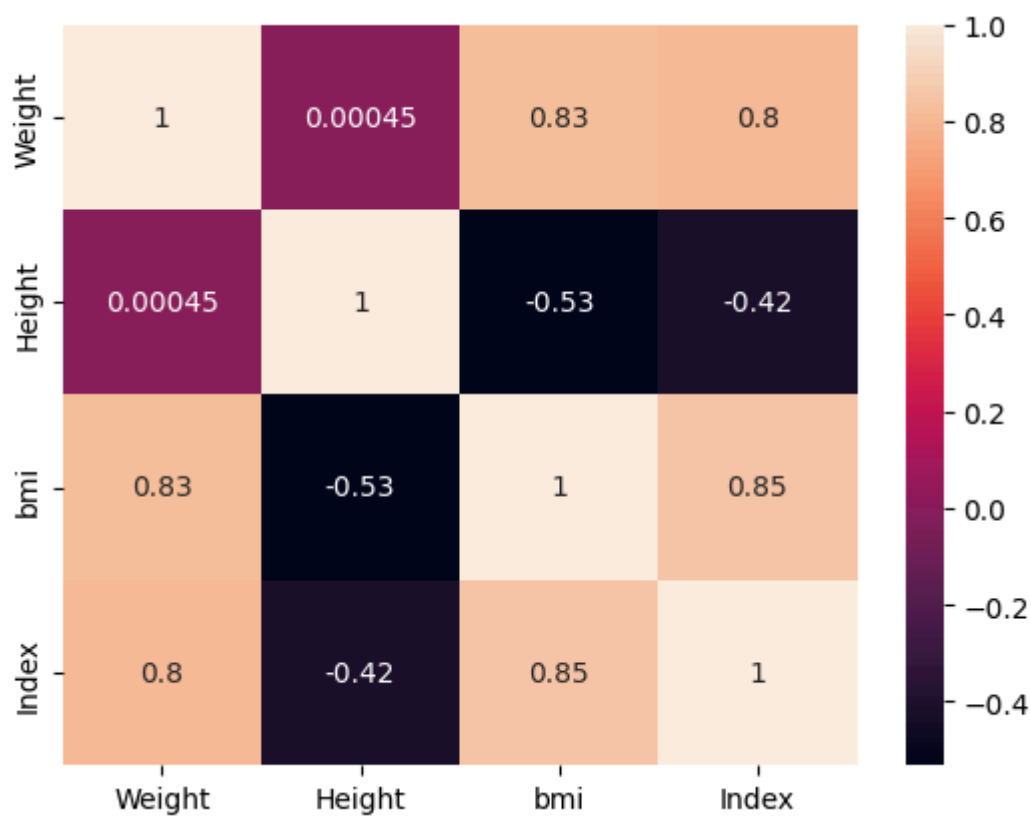
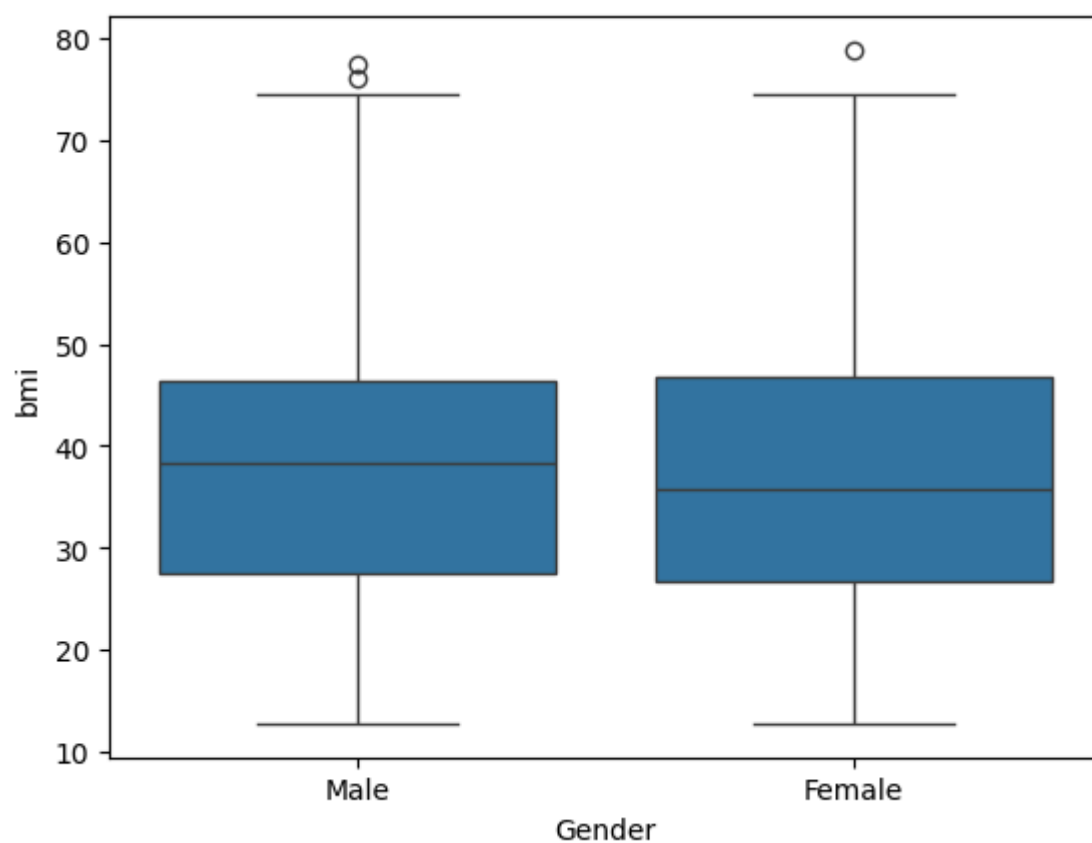
för att fylla kraven för denna punkten så skapades flera modeller och utvärderade dem och sen sist väljs den bästa modellen baserat från RMSE resultatet. men är osäker om det är tillräcklig för att fylla kraven?

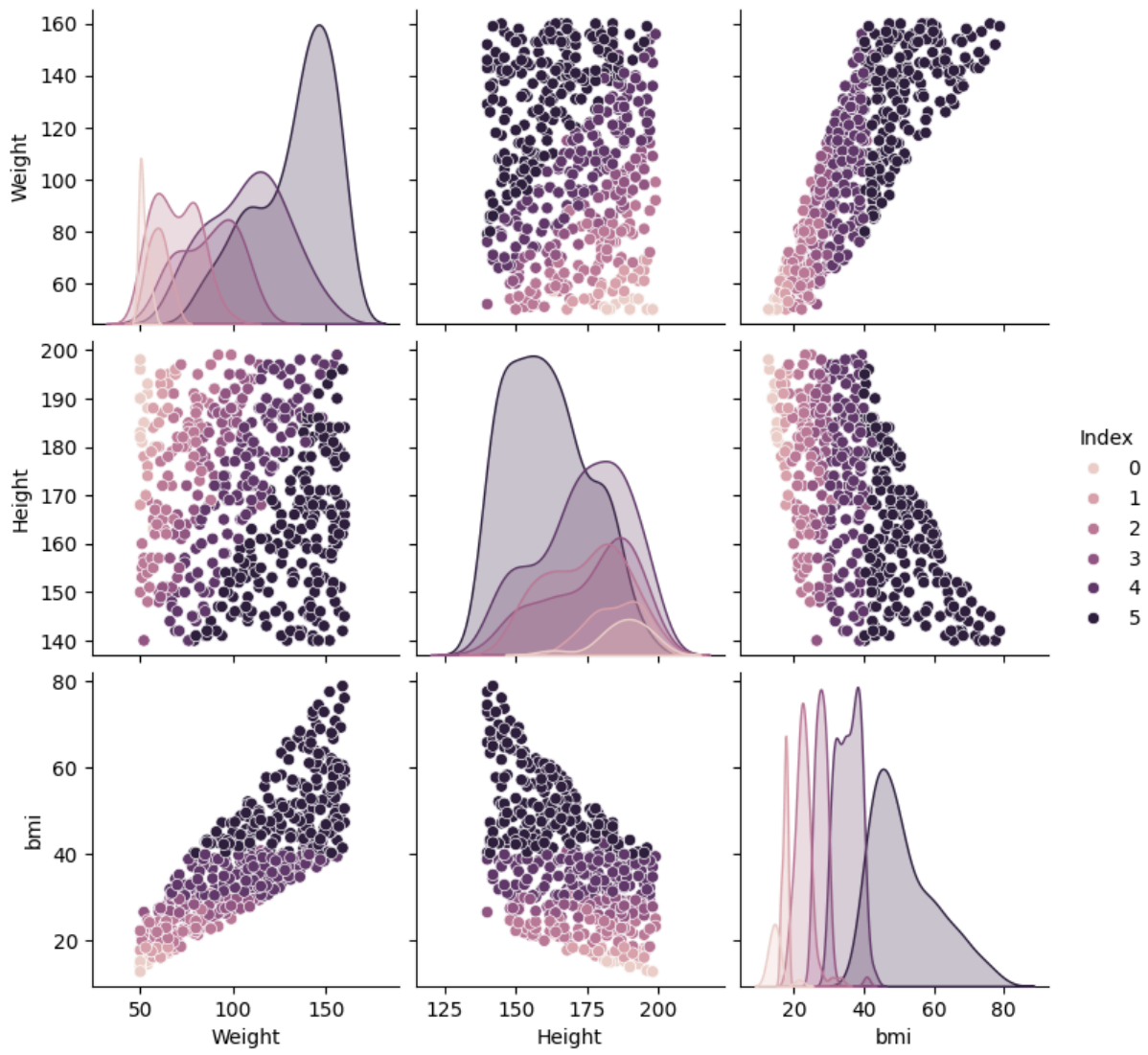
3. Något du vill lyfta fram till Antonio?

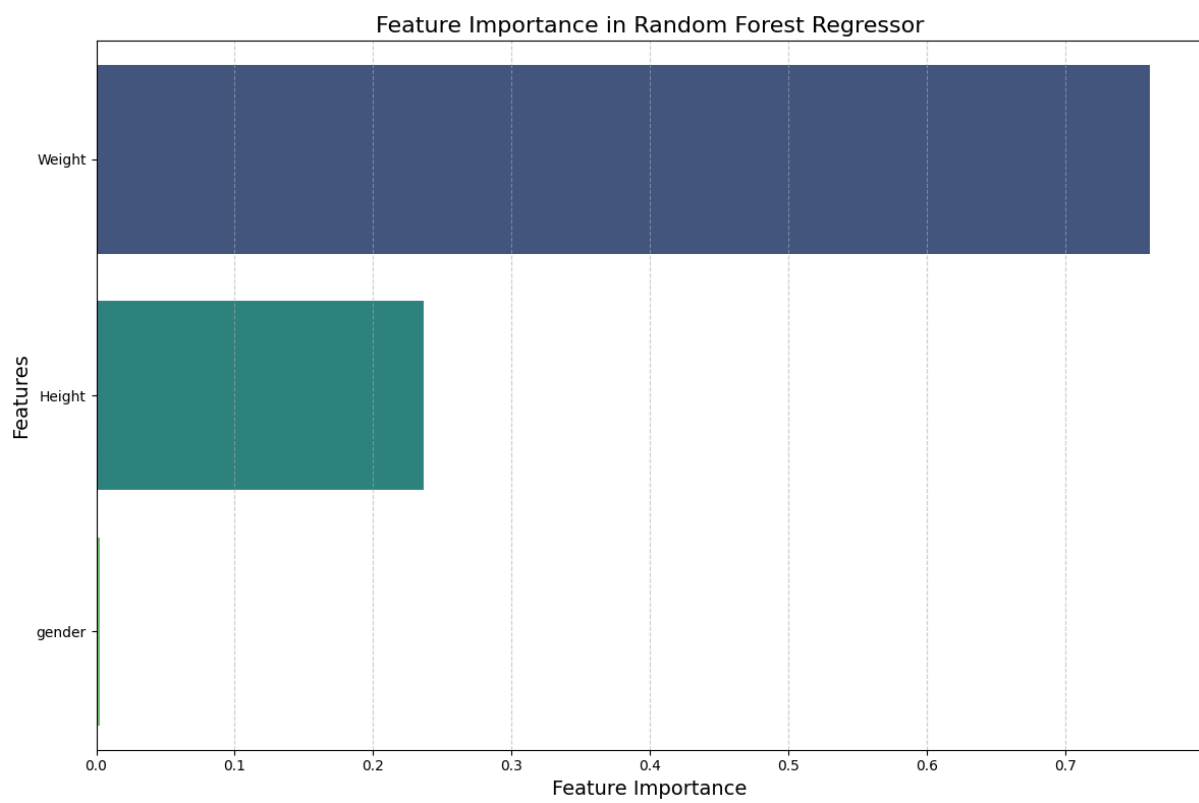
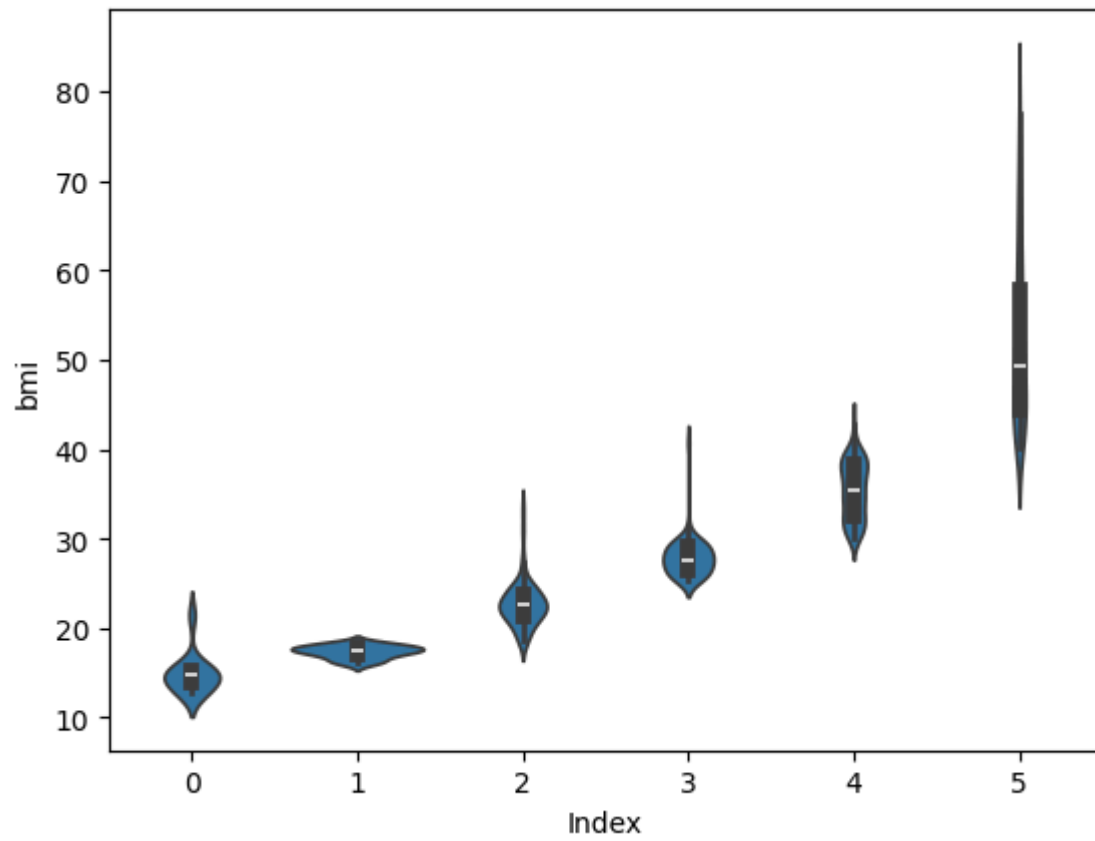
Utbildningen närma sig sitt slut och jag vill uttrycka mig att du har varit en riktig bra utbildare genom utbildningens gång. Jag tänker mig undra om du har/kommer ha en cybersäkerhet kurs i framtiden? om yes, sign me up som student igen :)

## Appendix A









## Källförteckning

Kaggle - <https://www.kaggle.com/>

BMI - [https://en.wikipedia.org/wiki/Body\\_mass\\_index](https://en.wikipedia.org/wiki/Body_mass_index)

Linear regression -  
[https://scikit-learn.org/stable/modules/generated/sklearn.linear\\_model.LinearRegression.html](https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html)

Decision tree regression - <https://scikit-learn.org/stable/modules/tree.html>

Random forest regression -  
<https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html>

XGBoost regression - <https://machinelearningmastery.com/xgboost-for-regression/>