Generalized Additive Model (GAM) dan Forward Selection pada Pemodelan Faktor - Faktor Yang Mempengaruhi Nilai Mahasiswa

Kanaya Dea Thalita
121450001

Sella Dianka Fitri
121450043

Halilah Roja Nasywa
121450046

Rizki Adrian Benovry
121450073

Anita Rahma Pramoda C
121450154

Pemodelan faktor-faktor yang mempengaruhi nilai mahasiswa telah menjadi perhatian utama dalam dunia pendidikan. Dalam rangka mengatasi tantangan yang terkait dengan hubungan non-linear antara variabel independen dan variabel dependen, penelitian ini mengusulkan penggunaan Generalized Additive Model (GAM) sebagai metode pemodelan yang fleksibel dan kuat. GAM memungkinkan identifikasi dan pemodelan hubungan non-linear antara faktor-faktor yang mempengaruhi nilai mahasiswa. Model ini dirancang untuk mengatasi ketidaklinieran dan interaksi kompleks antara variabel independen, yang tidak dapat ditangkap oleh model linear tradisional. Dalam GAM, variabel dependen, yaitu nilai mahasiswa, dijelaskan sebagai kombinasi linier dari faktor-faktor yang terlibat, dengan penambahan fungsi halus dari masing-masing faktor. Pendekatan ini memberikan kemampuan untuk mengeksplorasi hubungan yang lebih kompleks dan memahami pengaruh yang lebih akurat dari faktor-faktor yang mempengaruhi nilai mahasiswa. Dengan menggunakan GAM dalam pemodelan faktor-faktor yang mempengaruhi nilai mahasiswa, penelitian ini memberikan kontribusi penting dalam pemahaman tentang hubungan kompleks antara variabel independen dan variabel dependen dalam konteks pendidikan. Hasil penelitian ini dapat memberikan wawasan yang berharga bagi pendidik, penyedia kebijakan pendidikan, dan praktisi pendidikan dalam merancang intervensi yang lebih efektif untuk meningkatkan nilai mahasiswa.

Keywords - GAM, model, smooth, faktor, nilai, mahasiswa

I. Pendahuluan

Dalam dunia pendidikan, memahami faktor-faktor yang mempengaruhi hasil belajar siswa memegang peranan penting dalam meningkatkan mutu pendidikan. Model statistik yang tepat harus digunakan untuk mendapatkan pemahaman yang lebih lengkap. Salah satu pendekatan yang dapat digunakan adalah Generalized Additive Model (GAM). Dalam topik ini, kami fokus pada penerapan GAM untuk memodelkan faktor-faktor yang mempengaruhi nilai siswa. GAM adalah metode statistik yang dapat digunakan untuk mengidentifikasi hubungan non linier antara variabel independen dan variabel dependen. Dalam GAM, variabel terikat, dalam hal ini nilai siswa, diasumsikan dipengaruhi oleh kombinasi linear dari faktor-faktor yang berhubungan, dengan menambahkan fungsi smooth pada masing-masing faktor. Pendahuluan ini membahas tentang pentingnya dan manfaat penggunaan GAM dalam memodelkan faktor-faktor yang mempengaruhi nilai siswa. Kami juga menekankan pentingnya memahami dan mengidentifikasi faktor kunci yang mempengaruhi prestasi akademik siswa. Dengan menerapkan GAM, diharapkan dapat memperoleh pemahaman yang lebih mendalam tentang pola dan hubungan non-linear antara faktor-faktor tersebut dan hasil belajar siswa. Diharapkan jurnal ini akan memberikan kontribusi positif untuk pengembangan pendekatan analitis yang lebih kompleks dan akurat untuk memodelkan faktor-faktor yang mempengaruhi nilai siswa. Hasil penelitian ini dapat memberikan wawasan yang berharga bagi para pendidik, pembuat kebijakan pendidikan, dan penyelenggara pendidikan untuk meningkatkan kualitas belajar mengajar dan membantu siswa mencapai potensi akademiknya dengan sebaik-baiknya.

II. Metode penelitian

Data yang digunakan dalam penelitian merupakan dataset yang diperoleh dari publikasi pada website kaggle. Data ini mencakup 35 variabel yang terdiri atas banyak mata kuliah yang diambil, nilai semester, status perkawinan, angka gdp, status pembayaran ukt dan lain-lain. Variabel tersebut akan dibagi menjadi dua buah komponen yaitu variabel X (prediktor) dan variabel Y (respon). Dalam penelitian ini, terdapat beberapa langkah yang kami gunakan, yaitu :

1. Memilih data lalu menentukan variabel X dan Y.
2. Melakukan eksplorasi data untuk memahami karakteristik masing - masing variabel.
3. Melakukan pemrosesan data untuk membersihkan data dari missing values, variabel yang tidak digunakan, dan lain-lain.
4. Memisahkan dataset menjadi data training dan data validasi.
6. Melakukan pelatihan model dengan menggunakan algoritma yang sesuai dengan data yang dimiliki.
7. Mengevaluasi performa model.
8. Menyimpulkan hasil yang diperoleh dari proses analisis.

Adapun metode yang kami gunakan dalam proses analisis sebagai berikut.

A. *Generalized Linear Models (GLM)*

Dalam GLM, distribusi dari variabel dependen, yang juga dikenal sebagai variabel respons, dapat dengan jelas tidak normal dan dapat tidak bersifat kontinu (McCullagh, 1984, 2018,). Namun, variabel respons dapat mengikuti distribusi Poisson, berbeda dengan model linier umum. Selain itu, kombinasi linear dari variabel prediktor dapat memprediksi nilai variabel dependen, yang dapat "terhubung" dengan variabel dependen melalui fungsi link (Breslow & Clayton, 1993). Model linier umum dengan satu variabel dependen memberikan kesempatan untuk dipertimbangkan sebagai GLM dalam kasus tertentu. Dalam model linier umum, nilai dari variabel dependen diasumsikan mengikuti distribusi asimtotik, dan fungsi identitas adalah fungsi link. Kombinasi linear nilai tidak diubah untuk prediktor (Chatfield, 2010). Sebagai contoh, dalam model linier umum, ekspektasi bersyarat dari variabel dependen Y berkorelasi linear dengan ekspektasi bersyarat dari variabel independen X, sedangkan hubungan matematis dalam GLM dijelaskan sebagai berikut:

$$Y = g(b_0 + b_1X_1 + ... + b_mX_m)$$

B. *Generalized Additive Models (GAM)*

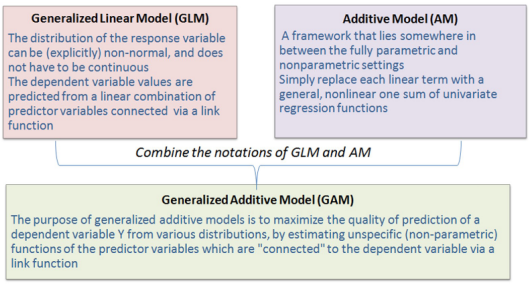
GAM (Generalized Additive Model) merupakan model yang dapat memberikan kerangka umum untuk memperluas model linier standar dengan mengizinkan fungsi non-linier dari masing-masing variabel, sambil mempertahankan aditivitas. Sama seperti model linier, GAM dapat diterapkan dengan respon kuantitatif dan kualitatif. Untuk response kontinu, Model GAM bisa ditulis sebagai:

$$y = \beta_0 \sum_{j=1}^{p} f(X_j) + \epsilon$$

Generalized additive models (GAM) memiliki bentuk $\eta(x) = \alpha + \sigma\, fj(xj)$, di mana η bisa menjadi fungsi regresi dalam regresi berganda atau transformasi logistik dari probabilitas posterior Pr(y = 1 x) dalam regresi logistik. Faktanya, model ini menggeneralisasi seluruh keluarga model linear umum $\eta(x) = \beta' x$, di mana $\eta(x) = g(\mu(x))$ merupakan transformasi dari fungsi regresi. Kami menggunakan algoritma scoring lokal untuk memperkirakan fungsi fj(xj) secara nonparametrik, dengan menggunakan smoother scatterplot sebagai blok bangunannya. Kami menunjukkan model ini dalam dua analisis yang berbeda: analisis kovarian nonparametrik dan regresi logistik. Prosedur ini bisa digunakan sebagai alat diagnostik untuk mengidentifikasi transformasi parametrik dari covariate dalam analisis linear standar. Berbagai alat inferensial telah dikembangkan untuk membantu analis dalam menilai relevansi dan signifikansi dari fungsi yang diperkirakan: ini termasuk kurva kepercayaan, estimasi derajat kebebasan, dan pengujian hipotesis aproksimatif. Algoritma scoring lokal analog dengan algoritma iterative least squares tertimbang ulang untuk memecahkan persamaan regresi likelihood dan nonlinear. Pada setiap iterasi, variabel penjelas diatur ulang dan model regresi tambahan dipasang menggunakan algoritma backfitting. Algoritma backfitting berulang kali memperkirakan setiap fungsi koordinasi dengan menghaluskan residu parsial. GAM merupakan tambahan klasik dari model linier umum yang diusulkan oleh (Hastie & Tibshirani, 1987) menunjukkan bahwa GLM yang memiliki prediktor linear berinteraksi dengan jumlah fungsi halus dari covariate. GAM memberikan struktur untuk memperluas model linier umum dengan memungkinkan aditivitas fungsi non-linear dari variabel-variabel. Konsep dari model aditif dengan GLM dapat digabungkan untuk menghasilkan konsep GAM, seperti berikut:

$$gi(\mu Y) = \sum_i (f_i(X_i))$$

Keuntungan dari GAM adalah membatasi kesalahan dalam memprediksi variabel dependen Y dari berbagai distribusi dengan mengevaluasi fungsi-fungsi yang tidak spesifik yang terhubung dengan variabel dependen melalui fungsi link. GAM menyediakan spesifikasi respons yang fleksibel dengan mendefinisikan model dalam bentuk fungsi halus sebagai pengganti hubungan parametrik terperinci pada covariate. Keluwesan dan kemudahan ini datang dengan biaya merepresentasikan fungsi-fungsi halus dalam pola yang serupa dan memilih tingkat kehalusan.



III. HASIL DAN PEMBAHASAN

3.1 Deskripsi Data

Var	Min	Mean	Maks
Y	0	10.23	18.57
X1	-	-	-
X2	1	1.179	6
X3	0	6.887	18
X4	0	10.64	18.88
X5	-4.06	0.002	3.51
X6	1	9.899	17
X7	-	-	-

Tabel 3.1 berisi deskripsi data yang akan digunakan dalam penelitian diantaranya variabel Y yaitu nilai semester 2 mahasiswa sebagai variabel respon dan beberapa variabel prediktor yaitu X1 (Target kelulusan) yang merupakan variabel kualitatif, X2 (Status perkawinan), X3 (Applicaton Mode), X4 (Nilai semester 1), X5 (Angka GDP), X6 (Banyak mata kuliah yang diambil) dan X7 (Status pembayaran UKT). Dengan variabel tersebut akan

dilakukan metode GAM untuk melihat hubungan linier antar variabel.

3.2 Pengolahan Data

Langkah awal yang harus dilakukan adalah membagi data menjadi data latih (training data) dan data uji (testing data) dengan menggunakan generalisasi model aditif seperti kode dibawah ini.

```
set.seed(721)
trainid <- sample(1:nrow(dataset), nrow(dataset)*0.8)
dataset.train <- dataset[trainid, ]
dataset.test <- dataset[-trainid, ]
```

Dalam kode diatas, langkah awal yang harus dibuat adalah mengatur biji acak (random seed) menjadi 721. Setelah itu pada baris kedua dalam kode berfungsi untuk menghasilkan vektor *`trainid`* yang berisi indeks acak dari baris data yang akan digunakan sebagai data latih. Fungsi `sample` digunakan untuk memilih sampel secara acak dari 1 hingga jumlah baris dalam dataset. Dalam kasus ini, 80% dari total baris dataset akan digunakan sebagai data latih. Pada baris ketiga dalam kode tersebut adalah membuat subset dataset dengan menggunakan indeks yang dihasilkan sebelumnya *(`trainid`)* untuk memilih baris-baris yang sesuai dari dataset. Ini menghasilkan dataset latih yang akan digunakan untuk melatih model. Dengan membagi dataset menjadi data latih dan data uji, kita dapat melatih model menggunakan data latih dan menguji model tersebut menggunakan data uji untuk melihat seberapa baik model tersebut bekerja pada data yang belum pernah dilihat sebelumnya.

Langkah selanjutnya adalah melakukan seleksi variabel menggunakan metode *"forward"* menggunakan fungsi *`regsubsets()`* pada dataset latih *(`dataset.train`)*.

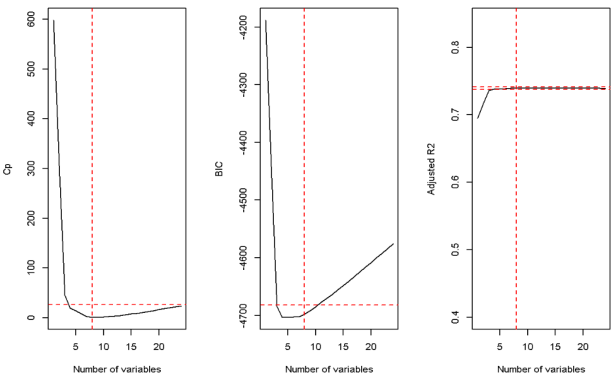
```
fit<- regsubsets(Curricular.units.2nd.sem..grade. ~ .,data=dataset.train, nvmax =24 ,method ="forward")
fit.summary<- summary(fit)
fit.summary
```

Output dari pemilihan subset variabel menggunakan algoritma seleksi maju (forward selection) untuk model regresi linear. Tertera pada output terdapat 25 variabel termasuk intercept dalam dataset ini. *"Forced in"* dan *"Forced out"* menunjukkan apakah suatu variabel dipaksa untuk dimasukkan atau dikeluarkan dari pemilihan subset. Dalam output ini, tidak ada variabel yang dipaksa masuk atau keluar.

penggunaan simbol ”*” dalam setiap baris yang menghubungkan variabel prediktor dengan variabel respons menunjukkan bahwa variabel prediktor tersebut searah dan berjalan secara linier terhadap variabel responnya sehingga fitur tersebut berefek signifikansi dan menunjukan variabel tersebut telah dipilih dalam langkah seleksi, sedangkan tanda " " menunjukkan bahwa variabel tersebut tidak dipilih.

Dari beberapa output diatas, kita dapat mengetahui variabel prediktor mana yang banyak memiliki simbol bintang “*” yang menyatakan hubungan linier terhadap variabel respon. dapat diketahui seperti variabel *“Target”* memiliki hubungan yang linier terhadap tabel responnya, sehingga fitur tersebut berefek signifikansi dan berjalan secara linier terhadap variabel respon. selain itu variabel “Curricular.units.1st.sem..grade.” juga memiliki nilai yang signifikan terhadap variabel respon, begitupun juga dengan variabel “course” dan *“Tuition.fees.up.to.date”*.

Langkah selanjutnya adalah membuat Plot Cp, BIC dan adj R2 untuk menentukan variabel terbaik dari 25 variabel yang ada pada dataset.



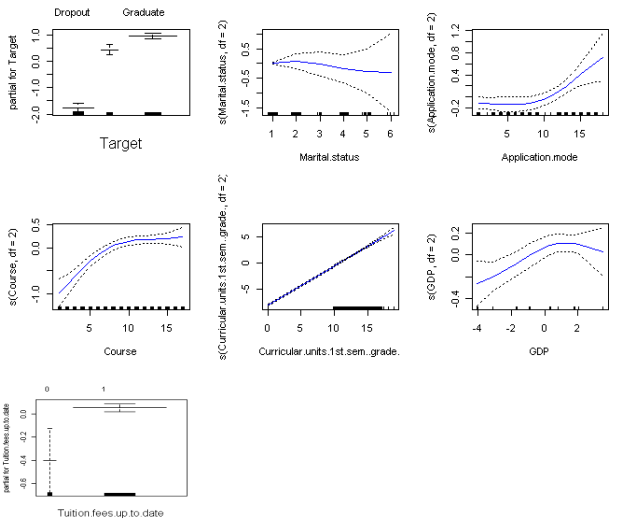
Dari ketiga plot tersebut dapat diketahui bahwa number of variabels terbaik ditunjukkan oleh angka 8. Jumlah variabel tersebut disarankan karena memiliki kecocokan yang paling baik dengan model. Oleh karena itu, kami memilih subset dari 8 variabel terbaik yang telah dipilih yaitu :

```
[1] "(Intercept)"                    "Marital.status"
[3] "Application.mode"               "Course"
[5] "Tuition.fees.up.to.date1"       "Curricular.units.1st.sem..grade."
[7] "GDP"                            "TargetEnrolled"
[9] "TargetGraduate"
```

Selanjutnya adalah pembuatan pemodelan GAM dengan menggunakan fungsi

```
```{r}
fit.gam <- gam(Curricular.units.2nd.sem..grade. ~ Target + s(Marital.status, df=2)
               + s(Application.mode, df=2) + s(Course, df=2)
               + s(Curricular.units.1st.sem..grade., df=2)+ s(GDP, df=2)
               + Tuition.fees.up.to.date,
               data=dataset.train)
par(mfrow = c(2, 3))
plot(fit.gam, se = T, col = "blue")
```
```

Sehingga, diperoleh plot dari masing-masing variabel berikut.



Plot GAM menunjukan hubungan antara variabel prediktor dengan variabel responnya. Dalam plot ini variabel application mode, course, 1st sem grade, course terlihat memiliki kemiringan positif yang menunjukkan bahwa variabel prediktor menunjukkan hubungan yang

positif, nilai variabel respon akan meningkat ketika nilai dari variabel prediktor meningkat. Sedangkan variabel GCP dan Marital status memiliki tren meningkat dan kemudian menurun.

Langkah selanjutnya adalah melihat nilai MSE dan R2. Output yang diberikan adalah nilai kesalahan kuadrat rata-rata sebesar 6.336417. Nilai ini mengindikasikan seberapa baik model GAM yang telah di-fit mampu memprediksi variabel respons pada dataset uji. Semakin rendah nilai kesalahan kuadrat rata-rata, semakin baik model dalam memprediksi nilai sebenarnya.

Selanjutnya,, menghitung nilai R-squared yang merupakan ukuran seberapa baik model GAM cocok dengan data pada dataset uji. Output yang diberikan adalah nilai RSS sebesar 0.7533333. Nilai ini mengindikasikan seberapa besar variabilitas yang tidak dapat dijelaskan oleh model GAM pada dataset uji. Semakin kecil nilai RSS, semakin baik model dalam menjelaskan variasi yang ada dalam variabel response. Nilai R-squared dapat dihitung sebagai 1 dikurangi dengan nilai RSS, sehingga semakin tinggi nilai R-squared, semakin baik model dalam menjelaskan variasi dalam data.

Langkah terakhir adalah mengevaluasi performa model dengan perintah summary(fit.gam) yang digunakan untuk menghasilkan ringkasan statistik dari model GAM yang telah di-fit.

```
Call: gam(formula = Curricular.units.2nd.sem..grade. ~ Target +
s(Marital.status,
    df = 2) + s(Application.mode, df = 2) + s(Course, df = 2) +
    s(Curricular.units.1st.sem..grade., df = 2) + s(GDP, df = 2) +
    Tuition.fees.up.to.date, data = dataset.train)
Deviance Residuals:
     Min       1Q   Median       3Q      Max
-12.8588  -0.5967   0.1105   0.9924  13.8978

(Dispersion Parameter for gaussian family taken to be 7.0988)

    Null Deviance: 97351.83 on 3538 degrees of freedom
Residual Deviance: 25023.18 on 3525 degrees of freedom
AIC: 16995.38

Number of Local Scoring Iterations: NA


 Anova for Parametric Effects
                                              Df Sum Sq Mean Sq   F value
 Target                                        2  33923   16962 2389.3785
 s(Marital.status, df = 2)                     1      7       7    1.0557
 s(Application.mode, df = 2)                   1     17      17    2.3646
 s(Course, df = 2)                             1   2719    2719  382.9915
 s(Curricular.units.1st.sem..grade., df = 2)   1  34339   34339 4837.2794
 s(GDP, df = 2)                                1     46      46    6.5222
 Tuition.fees.up.to.date                       1     61      61    8.5617
 Residuals                                  3525  25023       7
                                              Pr(>F)
 Target                                     < 2.2e-16 ***
 s(Marital.status, df = 2)                   0.304267
 s(Application.mode, df = 2)                 0.124205
 s(Course, df = 2)                          < 2.2e-16 ***
 s(Curricular.units.1st.sem..grade., df = 2) < 2.2e-16 ***
 s(GDP, df = 2)                              0.010695 *
 Tuition.fees.up.to.date                     0.003455 **
 Residuals
 ---
 Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

 Anova for Nonparametric Effects
                                              Npar Df  Npar F     Pr(F)
 (Intercept)
 Target
 s(Marital.status, df = 2)                          1  1.2621  0.261327
 s(Application.mode, df = 2)                        1 10.1484  0.001457 **
 s(Course, df = 2)                                  1 23.0578 1.637e-06 ***
 s(Curricular.units.1st.sem..grade., df = 2)        1  0.7773  0.378030
 s(GDP, df = 2)                                     1  5.1990  0.022657 *
 Tuition.fees.up.to.date
 ---
 Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Berdasarkan hasil analisis yang diberikan, terdapat model GAM yang dibuat untuk memodelkan hubungan antara variabel *Curricular.units. 2nd.sem..grade*. dengan beberapa variabel prediktor, yaitu *Target*, *Marital.status*, *Application.mode, Course*, *Curricular.units.1st.sem..grade.,* dan *GDP*. Model GAM ini menggunakan distribusi gaussian (normal). Nilai parameter dispersi yang digunakan dalam model ini adalah 7.0988.

Null Deviance menunjukkan nilai deviance ketika model hanya memprediksi variabel respon dengan menggunakan intercept saja, tanpa menggunakan variabel prediktor apapun. Nilai deviance ini dihitung dari jumlah kuadrat deviasi antara nilai aktual dari variabel respon dan nilai rata-rata dari variabel respon. Dalam hal ini, nilai deviance yang dihasilkan adalah sebesar 97351.83 pada 3538 derajat kebebasan.

Residual Deviance menunjukkan nilai deviance residu setelah model GAM yang dibuat diterapkan pada data. Nilai deviance residu dihitung dari jumlah kuadrat deviasi antara nilai aktual dari variabel respon dan nilai yang diprediksi oleh model GAM. Semakin kecil nilai residual deviance, semakin baik model GAM dalam memprediksi variabel respon. Dalam hal ini, nilai deviance residu yang dihasilkan adalah sebesar 25023.18 pada 3525 derajat kebebasan.

AIC menunjukkan nilai informasi *Akaike (Akaike Information Criterion)* dari model GAM yang dibuat. AIC digunakan sebagai metode untuk membandingkan model yang berbeda dan memilih model terbaik. Semakin kecil nilai AIC, semakin baik model dalam menjelaskan data. Dalam hal ini, nilai AIC yang dihasilkan adalah sebesar 16995.38.

Hasil analisis dari tabel ANOVA komponen nonparametrik menunjukkan adanya hubungan non linier yang kuat antara variabel course dengan variabel respon. Dan variabel application mode serta GDP menunjukkan hubungan linear dengan variabel respon namun tidak terlalu signifikan. Untuk melihat nilai kekuatan hubungan linear variabel digunakan matriks korelasi sebagai berikut.

	dataset.train.Curricular.units.2nd.sem..grade.
dataset.train.Curricular.units.2nd.sem..grade.	1.00000000
dataset.train.Course	0.17197987
dataset.train.Application.mode	-0.11752971
dataset.train.GDP	0.06625904
dataset.train.Marital.status	-0.07247665
dataset.train.Curricular.units.1st.sem..grade.	0.83378629

Setiap angka dalam matriks menunjukkan koefisien korelasi antara pasangan variabel tersebut. Nilai korelasi berkisar antara -1 hingga 1, di mana nilai 1 menunjukkan korelasi positif sempurna, nilai -1 menunjukkan korelasi negatif sempurna, dan nilai 0 menunjukkan tidak adanya korelasi. Misalnya, pada baris pertama dan kolom pertama, nilai korelasi adalah 1. Ini menunjukkan bahwa variabel *"Curricular.units .2nd.sem..grade."* memiliki korelasi positif sempurna dengan dirinya sendiri. Nilai korelasi 0.83 pada baris pertama dan kolom terakhir menunjukkan hubungan yang cukup kuat antara variabel *"Curricular.units .2nd.sem..grade."* dan *"Curricular.units.1st.sem..grade."*.

Matriks korelasi dapat memberikan informasi tentang hubungan linier antara variabel-variabel dalam dataset. Nilai korelasi yang tinggi antara dua variabel menunjukkan hubungan yang kuat di antara keduanya, sedangkan nilai korelasi yang rendah atau mendekati nol menunjukkan hubungan yang lemah atau tidak ada hubungan linier. Dengan melihat matriks korelasi kita dapat mengevaluasi kekuatan dan arah hubungan antara variabel-variabel dalam dataset. Hal ini dapat membantu dalam pemahaman karakteristik data dan pemilihan variabel yang relevan dalam analisis lebih lanjut. Tabel ini hanya memberikan tampilan yang lebih rapi dari matriks korelasi, memudahkan

untuk melihat hubungan korelasi antara variabel-variabel dalam dataset.

IV. KESIMPULAN

Dari analisis diatas, terlihat bahwa ketika memodelkan antara variabel Y dengan beberapa variabel X menggunakan metode *Generalize Additive Model (GAM)*, hasilnya adalah terdapat beberapa variabel prediktor yang memiliki hubungan positif dan kuat dengan variabel respon diantaranya variabel 1st sem grade (nilai semester 1), variabel course (banyak mata kuliah yang dipilih) serta GDP (angka perekonomian mahasiswa), nilai akan cenderung meningkat jika ketiga variabel tersebut meningkat. Sedangkan, terdapat juga variabel lain yang memiliki hubungan negatif dengan variabel respon yaitu marital status (status perkawinan) dan application mode yang berarti variabel ini dapat memperburuk kualitas nilai pada mahasiswa jika angkanya mengalami kenaikan meskipun tidak signifikan.

DAFTAR PUSTAKA

Breslow, & Clayton. (1993). Approximate inference in generalized linear mixed models.

Chatfield. (2010). An Introduction to Generalized Linear Models.

Hastie, & Tibshirani. (1987). Generalized additive models: some applications.

McCullagh. (1984, 2018). Generalized linear models.