FINAL PROJECT- DATA MINING C

KLASIFIKASI *NBA PLAYER* MENGGUNAKAN METODE REGRESI LOGISTIK BINER DAN *MACHINE LEARNING*

DisusunOleh :
Novita Dwi Utami NRP 062115 4000 0019
Iftitah Ayundari NRP 062115 4000 0039

Dosen Pembimbing :
Dr. Santi Wulan Purnami, S.Si., M.Si.
Erma Oktania Permatasari, S.Si., M.Si.

PROGRAM STUDISARJANA
DEPARTEMENSTATISTIKA
FAKULTAS MATEMATIKA,KOMPUTASI, DAN SAINS DATA
INSTITUT TEKNOLOGISEPULUH NOPEMBER
SURABAYA 2018

LAPORAN FINAL PROJECT

KLASIFIKASI *NBA PLAYER* MENGGUNAKAN METODE REGRESI LOGISTIK BINER DAN *MACHINE LEARNING*

Disusun Oleh :

Novita Dwi Utami NRP 062115 4000 0019
Iftitah Ayundari NRP 062115 4000 0039

Dosen Pembimbing :

Dr. Santi Wulan Purnami, S.Si., M.Si.
Erma Oktania Permatasari, S.Si., M.Si.

PROGRAM STUDI SARJANA
DEPARTEMEN STATISTIKA
FAKULTAS MATEMATIKA, KOMPUTASI, DAN SAINS DATA
INSTITUT TEKNOLOGI SEPULUH NOPEMBER
SURABAYA 2018

(Halaman ini sengaja dikosongkan)

KLASIFIKASI *NBA PLAYER* MENGGUNAKAN METODE REGRESI LOSGISTIK DAN METODE *MACHINE LEARNING*

Nama Mahasiswa 1 : Novita Dwi Utami
NRP Mahasiswa 1 : 06211540000019
Nama Mahasiswa 2 : Iftitah Ayundari
NRP Mahasiswa 2 : 06211540000039
Program Studi : S1-Statistika FMKSD-ITS

ABSTRAK

Permainan bola basket adalah permainan dua regu yang berlawanan, dimainkan dengan lima orang pemain yang bertujuan untuk memasukkan bola sebanyak-banyaknya ke keranjang lawan dan mencegah kemasukan di keranjangnya sendiri. Salah satu pertimbangan dalam memilih pemain basket selain dengan *skill* adalah usia karir.Untuk itu dilakukan klasifikasi menggunakan metode *machine learning* dan metode statistika. Tujuannya untuk melakukan analisis dan membantu pelatih atau pencari bakat dalam mencari iiiogistic pemain yang professional dengan mempertimbangkan usia karir. Berdasarkan hasil analisis dengan membandingkan antara metode statistika dan metode *machine learning* dimana metode statistika menggunakan regresi logistic biner dan *machine learning* menggunakan metode *naïve bayes, decision tree, SVM* dan *random forest,* didapatkan hasil bahwa metode klasifikasi terbaik untuk kasus NBA *player* dalam penelitian ini adalah metode *naïve bayes* dengan tingkat ketepatan klasifikasi atau akurasi sebesar 70,1%

Kata Kunci : Akurasi, Machine Learning, Metode Statistika, NBA player.

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

(Halaman ini sengaja dikosongkan)

BAB I

PENDAHULUAN

1.1 Latar Belakang

Permainan bola basket adalah permainan dua regu yang berlawanan, dimainkan dengan lima orang pemain yang bertujuan untuk memasukkan bola sebanyak-banyaknya ke keranjang lawan dan mencegah kemasukan di keranjangnya sendiri(Budiana & Lubay, 2013).Olah raga basket lahir pada tahun 1891 digagas oleh Dr. James A.Naissmith. Olah raga ini bertipe permainan. Pada awalnya ditujukan untuk menyegarkan pikiran dan perasaan orang yang terlibat dalam permainan itu. Namun hal ini berubah sejak munculnya berbagai liga bola basket di dunia, seperti liga basket di Amerika yaitu *National BasketballAssociation* (NBA), NBA adalah liga bola basket pria di Amerika Serikat dan merupakan liga basket paling bergengsi di dunia.

Basketball berubah menjadi suatu permainan profesional yang harus ditangani secara profesional pula Hal ini mengharuskan pengurus tim terutama pencari bakat *(scout)* dan pelatih *(coach)* untuk teliti dalam melihat bakat seorang pemain basket dan memutuskan posisinya di lapangan serta berapa lama pemain tersebut dimainkan. Salah satu cara untuk mendukung keputusantersebut adalah dengan melakukan analisis pada data statistik dari pemain tersebut. Secara umum, sebuah tim yang solid adalah tim yang terdiri dari 15 pemain dengan skill diatas rata– rata. Kategori kemampuan *(skill)* dari masing–masing pemain dapat diputuskan dengan menganalisa statistik selama pemain bersangkutan bertanding. Data statistik ini sangat detail, mulai dari ukuran fisik sampai perhitungan *foul*/pelanggaran dan perolehan point. Dari analisis statistik tentang karakteristik pemain ini dapat ditentukan berbagai macam skill, yang nantinya menjadi dasar pengambilan keputusan. Keputusan tersebut dapat berupa usulan perekrutan pemain oleh pemandu bakat. Selain itu hasil analisis dapat menjadi dasar bagi pelatih guna memutuskan penempatan posisi pemain serta lama pemain bermain pada saat pertandingan.

Salah satu pertimbangan dalam memilih pemain basket selain dengan *skill* adalah usia karir. Untuk itu peneliti akan melakukan klasifikasi menggunakan metode *machine learning* dan metode statistika. Tujuannya untuk melakukan analisis dan membantu pelatih atau pencari bakat dalam mencari 2ogistic pemain yang professional dengan mempertimbangkan usia karir.

1.2 Rumusan Masalah

Rumusan masalah dalam penelitian ini antara lain

1. Bagaimana karakteristik permainan basket dari *NBA Player*
2. Bagaimana prediksi pengelompokkan dari *NBA Player* usia karir berdasarkan metode 2ogistic2 dan *machine learning*
3. Bagaimana nilai akurasi, spensifitas, dan sensitifitas dari hasil pengelompokkan berdasarkan metode 2ogistic2 dan *machine learning*

1.3 Tujuan Penelitian

Berdasarkan rumusan masalah, tujuanm yang ingin dicapai pada penelitian ini adalah sebagai berikut

1. Mendeskripsikan karakteristik permainan basket dari *NBA Player*
2. Memprediksi pengelompokkan dari *NBA Player*usia karir berdasarkan metode statistik dan *machine learning.*
3. Mengetahui nilai akurasi, spensifitas, dan sensitifitas dari hasil pengelompokkan berdasarkan metode statistik dan *machine learning.*

1.4 Manfaat Penelitian

Manfaat yang ingin dicapai melalui penelitian ini adalah sebagai berikut.

1. Dapat mengetahui karakteristik permainan basket dari *NBA Player*
2. Memberikan informasi mengenai prediksi pengelompokkan dari *NBA Player*usia karir berdasarkan metode statistik dan *machine learning.*

3. Menambah pengetahuan mengenai nilai akurasi, spensifitas, dan sensitifitas dari hasil pengelompokkan berdasarkan metode statistik dan *machine learning*.
4. Menjadi bahan acuan bagi pelatih atau pencari bakat dalam mencari pemain basket yang professional.

1.5 Batasan Penelitian

Batasan masalah dalam penelitian ini dibatasi hanya pada musim 2015 hingga 2016.

Halaman ini sengaja dikosongkan

BAB II
TINJAUAN PUSTAKA

2.1 *Pre Processing Data*

Salah satu tahapan dalam melakukan *pre processing* data adalah melakukan imputasi pada data yang mengelami *missing value*

Missing value secara umum adalah keadaan dimana ada *value* (nilai) dari satu atau lebih variabel yang hilang / tidak tersedia untuk analisis.*Missing value* terjadi karena informasi untuk sesuatu tentang objek tidak diberikan,sulit dicari, atau memang informasi tersebut tidak ada(Hair,dkk, 2010). *Missing value* dapat menyebabkan data menjadi bias sehingga memungkinkan hasil dari nalisis data tersebut menjadi tidak valid. Oleh karena itu, jika terjadi *missing value*, maka harus dilakukan perlakuan untuk mengatasinya.

Berbagai perlakuan *(treatment)* yang dapat dilakukan pada data-data yang *missing at random :*

1. Membuang baris (kasus) yang mengandung *missing value*, menghapus variabel (kolom) yang mengandung *missing value.*
   a. Apabila *missing value* kurang < 10% bisa diabaikan atau dengan kata lain bisa dilanjutkan ke analisis selanjutnya tanpa imputasi
   b. Apabila *missing value* lebih dari 15% bisa menjadi kandidat untuk dilakukan penghapusan variabel ataupun observasi, tetapi apabila nilai missing value 20% hingga 30% masih dapat diatasi(tergantung dengan peneliti)
   c. Apabila *missing value*>50% bisa dilakukan penghapusan data(Hair,dkk, 2010).
2. Mengisi sel (data) yang *missing* dengan nilai tertentu yang dianggap bisa mendekati kenyataan sebenarnya jika data terisi. Hal ini lebih baik dan rasional daripada membuang satu baris (data konsumen) hanya karena usia konsumen tidak terdata, atau bahkan satu variabel hanya karena satu dua sel tidak terisi. Cara mengisi data yang *missing* bisa

bermacam-macam, dan yang populer adalah mengisi dengan rata-rata keseluruhan data(Santoso, 2010).

2.2 *Feature Selection*

Feature Selection adalah suatu proses yang mencoba untuk menemukan sub himpunan dari himpunan fitur yang tersedia untuk meningkatkan aplikasi dari suatu algoritma pembelajaran. Feature Selection digunakan dibanyak area aplikasi sebagai alat untuk menghilangkan fitur yang tidak relevan dan atau fitur berlebihan. Sebuah fitur dikatakan tidak relevan jika memberikan sedikit informasi, sedangkan sebuah fitur dikatakan berlebihan jika informasi yang diberikan adalah informasi yang terkandung dalam fitur lain (tidak memberikan informasi baru)(Dash & Liu, 1997).

2.3 Statistika Deskriptif

Statistika deskriptif merupakan metode-metode yang berkaitan dengan pengumpulan dan penyajian suatu gugus data sehingga memberikan informasi yang berguna. Statistika deskriptif dalam pratikum ini, terdiri dari *mean* dan standar deviasi(Walpole, 2013).

a. *Mean* atau rata-rata adalah jumlah semua data yang ada dibagi dengan banyaknya data.

Data tunggal :

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \ldots\ldots + x_n}{n} \qquad (2.1)$$

b. Standar Deviasi untuk sebuah peubah acak $x_1, x_2, \ldots\ldots, x_n$ didefinisikan sebagai berikut.

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} \qquad (2.2)$$

2.4 Regresi Logistik Biner

Regresi 6ogistic biner adalah suatu metode analisis data yang digunakan untuk mencari hubungan antara variabel respon (y) yang bersifat biner dengan variabel prediktor (x) (Hosmer & Lemeshow, 2000).

2.5 *Naïve Bayes*

Naïve Bayes adalah metode klasifikasi yang berdasarkan probabilitas dan *Teorema Bayesian* dengan asumsi bahawa setiap varaibel *X* bersifat bebas atau independen. Dengan kata lain, *Naïve Bayes* mengasumsikan bahwa keberadaan sebuahatribut tidak ada kaitannya dengan keberadaan atribut yang lain (Abidin,2012)

2.6 *Decision Tree*

Decision Tree atau pohon keputusan adalah pemetaan mengenai alternative alternative pemecahan masalah yang dapat diambildari masalah tersebut. Pohon keputusan memadukan antara eksplorasi data dan pemodelan, sehingga sangat bagus sebagai langkah awal dalam proses pemodelan bahkan ketika dijadikan sebagai model akhir dari beberapa teknik lain (Kusrini & E.T, 2009)

2.7 SVM

Support Vector Machine (SVM) dikembangkan oleh Boser, Guyon, Vapnik, dan pertama kali dipresentasikan pada tahun 1992 di Annual Workshop on Computational Learning Theory. Konsep dasar SVM sebenarnya merupakan kombinasi harmonis dari teori-teori komputasi yang telah ada puluhan tahun sebelumnya, seperti margin hyperplane (Duda & Hart (1973), Cover (1965), Vapnik (1964), dan sebagai nya.), kernel (Aronszajn, 1950) dan konsep-konsep pendukung yang lain. Belum pernah ada upaya merangkaikan komponen-komponen tersebut hingga tahun 1992.

2.8 *Random Forest*

Random Forest pertama kali dikenalkan oleh Breiman pada Tahun 2001. Dalam penelitiannya menunjukkan kelebihan Random Forest antara lain dapat menghasilkan error yang lebih rendah, memberikan hasil yang bagus dalam klasifikasi, dapat mengatasi data training dalam jumlah sangat besar secara efisien, dan metode yang efektif untuk mengestimasi missing data (Breiman, 2001)

2.9 *Confussion Matrix*

Confussion Matrix adalah sebuah tabel yang menyatakan jumlah data uji yang benar diklasifikasikan dan jumlah data uji

yang salah diklasifikasikan. Contoh *confusion matrix* untuk klasifikasi biner ditunjukkan pada Tabel 2.1

Tabel 2.1 *Confussion Matrix*

		Prediksi		Total
		1	0	
Aktual	1	TP	FN	P
	0	FP	TN	N
Total		P’	N’	P+N

Keterangan untuk tabel 1dinyatakan sebagai berikut:

a) *True Posstive* (TP), yaitu jumlah dokumen dari kelas 1 yang benar dan diklasifikan sebagai kelas1.
b) *True Negative* (TN), yaitu jumlah dokumen dari kelas 0 yang benar diklasifikasikan sebagai kelas0.
c) *False Positive* (FP), yaitu jumlah dokumen dari kelas 0 yang salah diklasifikasikan sebagai kelas1.
d) *False Negative* (FN) yaitu jumlah dokumen dari kelas 1yang salah diklasifikasikan sebagai kelas0.

Perhitugan akurasi, *sensitivity*, *specivity* dinyatakan dalam persamaan berikut (Han, Kamber, & Pei, 2012)

$$Akurasi = \frac{TP + TN}{TP + FN + FP + TN} \times 100\%$$

$$Sensitivity = \frac{TP}{P} \times 100\%$$

$$Specivity = \frac{TN}{N} \times 100\%$$

2.10 Olah Raga Basket

Permainan bola basket adalah permainan dua regu yang berlawanan, dimainkan dengan lima orang pemain yang bertujuan untuk memasukkan bola sebanyak-banyaknya ke keranjang lawan dan mencegah kemasukan di keranjangnya sendiri(Budiana & Lubay, 2013).

BAB III
METODOLOGI PENELITIAN

3.1 Sumber Data

Data yang digunakan dalam penelitian ini adalah data sekunder. Data sekunder yang digunakan yaitu data *NBA Player* tahun 2015 hingga 2016 yang diambil dari website data.world.

3.2 Variabel Penelitian

Di dalam penelitian ini variabel yang digunakan berasal dari 1340 pemain basket di *NBA*. Macammacam variabel yang digunakan dapat dilihat pada Tabel 3.1

Tabel 3.1 Variabel Penelitian

Variabel		Skala Pengukuran
GP	*Games Played*	Rasio
MIN	*Minuted Played*	Rasio
PTS	*Points Per Game*	Rasio
FGM	*Field Goals Made*	Rasio
FGA	*Field Goal Attempts*	Rasio
FG%	*Filed Goal Percent*	Rasio
3P Made	*3 Point Made*	Rasio
3PA	*3 Point Attemps*	Rasio
3P%	*3 Point Percent*	Rasio
FTM	*Free Throw Made*	Rasio
FTA	*Free Throw Attempts*	Rasio
FT%	*Free Throw Percent*	Rasio
OREB	*Offensive Rebounds*	Rasio
DREB	*Defensive Rebounds*	Rasio
REB	*Rebound*	Rasio
AST	*Assists*	Rasio
STL	*Steals*	Rasio

BLK	*Blocks*	Rasio
TOV	*Turnovers*	Rasio
		Ordinal
TARGET	*Outcome*	0=Usia Karir< 5
		1=Usia Karir ≥5

Berikut merupakan penjelasan singkat dari variabel variabel yang digunakan dalam penelitian ini.

1. *Games Played,* Jumlah pertandingan yang telah dialami pemain dalam 1 musim kompetisi
2. *Minuted Played,* Durasi waktu dimana pemain bermain dalam bertandingan
3. *Points Per Game*, point yang dihasilkan pada setiap permainan
4. *Field Goals Made,* keberhasilan usaha seorang pemain untuk mencetak poin
5. *Field Goal Attempts*, Usaha yang dilakukan pemain untuk mencetak point
6. *Filed Goal Percent,* Persentase keberhasilan usaha seseorang pemain untuk mencetak poin pada pertandingan baik itu perolehan 2 poin atau 3 poin
7. *3 Point Made,* Keberhasilan usaha seorang pemain untuk mencetak 3 poin
8. *3 Point Attemps,* Jumlah tembakan 3 point
9. *3 Point Percent,* Persentase keberhasilan seorang pemain untuk mencetak poin melalui tembakan 3 angka pada pertandingan
10. *Free Throw Made,* Keberhasilan seorang pemain membuat tembakan bebas
11. *Free Throw Attempts*, Usaha seorang pemain membuat tembakan bebas
12. *Free Throw Percent,* Persentase keberhasilan seorang pmeain dalam mengeksekusi tembakan bebas atau tembakan hukuman di pertandingan
13. *Offensive Rebounds,* Perolehan bola netral atau bola mentah hasil tembakan tim serang pada saat posisi menyerang pada peetandingan

14. *Defensive Rebounds*,Perolehan bola netral atau bola mentah hasil tembakan tim serang pada saat bertahan pada pertandingan
15. *Rebound*, Jumlah perolehan bola netral atau bola mentah yang dilakukan seorang pemain pada pertandingan
16. *Assists*, Hasil *passing* atau operan bola yang membuahkan poin dalampertandingan
17. *Steals*, Pencurianboladaritangantimlawan dalampertandingan
18. *Blocks*, Hasil menggagalkan tembakan pemain daritimlawandalampertandingan
19. *Turnover*, Kesalahan-kesalahan yang dilakukan seorang pemain dalampertandingan

3.3 Struktur Data Penelitian

Pada penelitian ini terdapat satu variabel respon (Y) dan 19 prediktor. Struktur data dalam penelitian ini ditampilkan pada Tabel 3.2 sebagai berikut

Tabel 3.2 Struktur Data

Nama	***Y***	***X₁***	***X₂***	**…**	***X₁₈***	***X₁₉***
Brandon	Y_1	X_{11}	X_{21}	…	$X_{18;1}$	$X_{19;1}$
Andrew	Y_2	X_{12}	X_{22}	…	$X_{18;2}$	$X_{19;2}$
Jakarr	Y_3	X_{13}	X_{23}	…	$X_{18;3}$	$X_{19;3}$
…	…	…	…	…	…	…
Jon Barry	Y_{1034}	$X_{1;1034}$	$X_{2;1034}$	…	$X_{18;1034}$	$X_{19;1034}$

3.4 Metode Analisis Data

Metode yang digunakan pada penelitian antara lain metode statistika dan *machine learning.* Pada metode statistika menggunakan regresi logistic biner. Sedangkan untuk metode *machine learning* digunakan metode *naïve bayes, decision tree, SVM,* dan *random forest.* Berikut adalah langkah langkah yang digunakan dalam melakukan penelitian ini adalah

1. Melakukan *pre processing* pada data *NBA Player*. Pada tahap *pre processing* data hal yang dilakukan adalah mengatasi *missing value*
2. Melakukan *Feature Selection*.

3. Mendeskripsikan data *NBA Player*. Mendiskripsikan dalam bentuk visualisasi *scatter plot, bar char*, korelasi, dan*box plot.*
4. Membagi data menjadi data testing dan data training
5. Melakukan analisis regresi 12ogistic ordinal
6. Melakukan klasifikasi dengan metode *machine learning*
7. Membandingkan nilai akurasi, spesifitas,dan sensitivitas antar metode baik regresi logistic ordinal dan masing masing metode dalam *machine learning*
8. Menarik kesimpulan.

BAB IV
ANALISIS DAN PEMBAHASAN

4.1. *Pre-Processing Data*

Pre-processing dilakukan dengan mengatasi *missing value* yang terdapat pada data. Untuk data NBA *player*, *missing value* hanya terdapat pada variabel 3P% atau *3 Points Percents* yaitu terdapat 11 *missing value.* Karena variabel 3P% merupakan variabel dengan data kontinyu, maka akan dilakukan imputasi *mean* untuk *missing value*nya. Setelah dilakukan imputasi data atau *cleaning data*, data telah siap untuk dianalisis pada tahapan selanjutnya.

Tahapan selanjutnya yaitu dengan melakukan *feature selection. Feature selection* dilakukan untuk memilih *features* atau variabel yang berpengaruh pada analisis data yang akan dilakukan oleh peneliti. Dari data awal yang memiliki 19 *features*, setelah dilakukan *feature selection* dapat mereduksi sebanyak 11 *features* dan menyisakan 8 *features* yaitu GP, MIN, PTS, FG%, 3PA, 3P%, FT% dan REB. Tahapan analisis selanjutnya akan menggunakan 8 *features* hasil dari *feature selection* sebagai variabel prediktor.

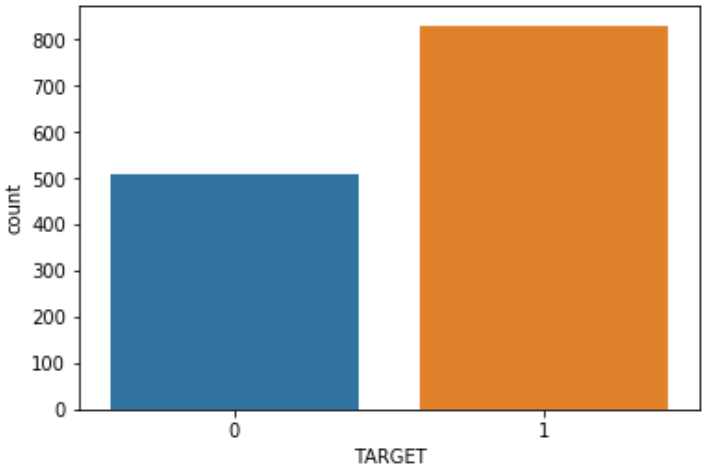
4.2. Karakteristik Data

Berikut merupakan tabel statistika deskriptif dari variabel prediktor yang mempengaruhi klasifikasi pemain NBA dikategorikan sebagai pemain dengan usia karir dibawah atau diatas 5 tahun.

Tabel 4.1 Statistika Deskriptif Variabel Penelitian

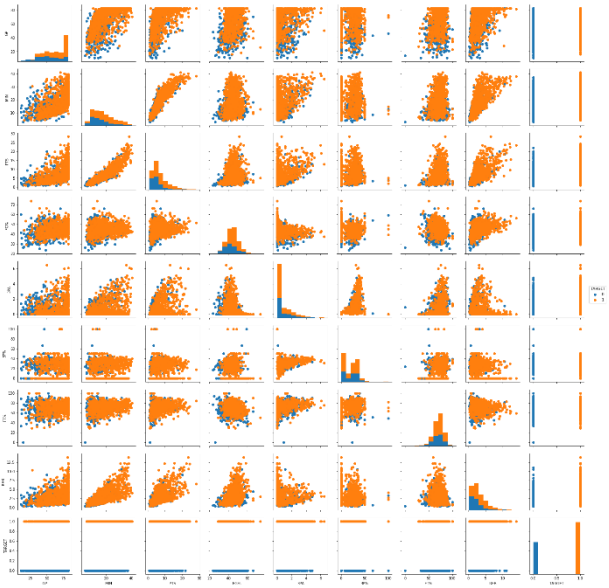
Variabel	Mean	Standar Deviasi
GP	60,414	17,433
MIN	17,625	8,308
PTS	6,801	4,357
FG%	44,169	6,138
3PA	0,779	1,062
3P%	19,308	15,956
FT%	70,300	10,578
REB	3,034	2,057

Berdasarkan tabel dapat diketahui bahwa rata-rata jumlah pertandingan yang telah dialami pemain dalam satu musim kompetisi (GP) adalah sebanyak 60,414 atau 60 kali. Rata-rata durasi waktu pemain bermain dalam permainan (MIN) adalah 17,625 menit. Variabel yang memiliki sebaran data atau standar deviasi tertinggi adalah jumlah pertandingan yang telah dialami pemain dalam satu musim kompetisi (GP) dan yang memiliki standar deviasi terendah adalah jumlah tembakan 3*point* yang dilakukan (3PA). Berikut ditampilkan grafik jumlah NBA *player* dengan usia karir diatas dan dibawah 5 tahun.



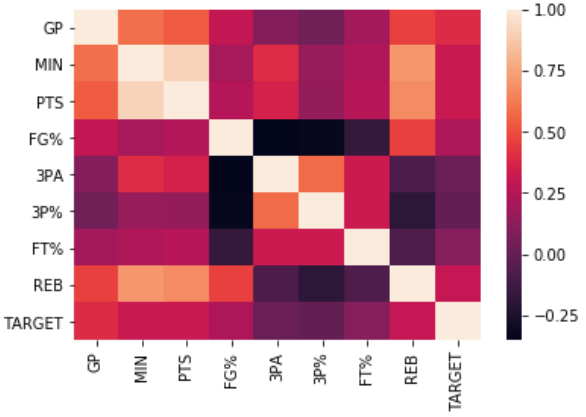
Gambar 4.1. Jumlah Pemain NBA Berdasarkan Usia Karir

Dari plot diatas, dapat dilihat bahwa dari NBA *player* yang diperoleh, sebanyak 509 pemain merupakan NBA *player* dengan usia karir dibawah 5 tahun dan sebanyak 831 pemain merupakan NBA *player* dengan usia karir diatas 5 tahun. Selanjutnya ditampilkan sebaran data penelitian melalui *scatterplot*. Berikut merupakan *scatterplot* per variabel dari data yang digunakan dalam penelitian.



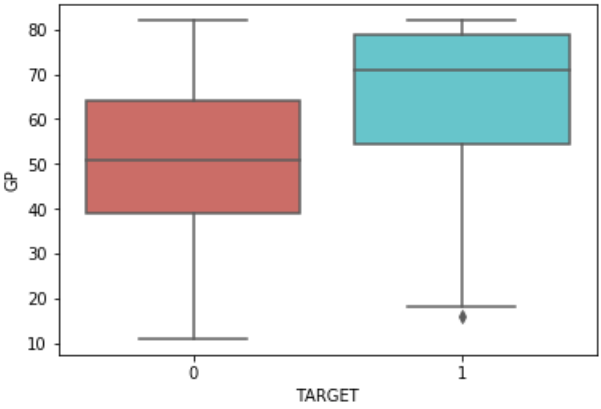
Gambar 4.2. *Scatterplot* Data Penelitian Per Variabel

Berdasarkan *scatter plot* di atas, diketahui bahwa data karakteristik NBA *player* tersebar secara acak. Untuk mengetahui hubungan antara variabel yang digunakan dalam penelitian, berikut ditampilkan visualisasi dari korelasi yang menampilkan besar korelasi antar variabel.



Gambar 4.3. KorelasiAntar Variabel

Dari visualisasi korelasi antar variabel diketahui bahwa korelasi tertinggi terjadi pada variabel MIN dan PTS, namun besarnya korelasi tidak lebih dari 0,75. Sedangkan untuk variabel lainnya memiliki korelasi antar variabel yang rendah. Berikut merupakan visualisasi *boxplot* dari variabel GP (jumlah pertandingan yang telah dialami pemain dalam satu musim kompetisi) untuk masing-masing target.



Gambar 4.4. *Boxplot* Variabel GP Berdasarkan Target

Dari grafik diatas diketahui bahwa NBA *player* dengan usia karir diatas 5 tahun mengalami jumlah pertandingan dalam satu musim kompetesi lebih banyak dibandingkan NBA *player* dengan usia karir dibawah 5 tahun.

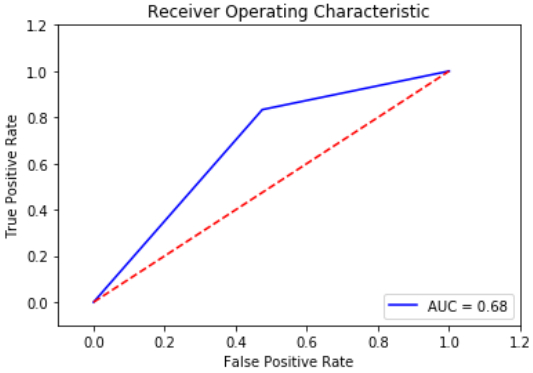
4.3. Klasifikasi Data Menggunakan Metode Statistika

Metode statistika yang digunakan dalam analisis klasifikasi NBA *player* berdasarkan usia karir dalam penelitian ini adalah regresi 16ogistic biner. Dalam analisis, data dibagi menjadi data *training* sebanyak 80% dan data *testing* sebanyak 20% dari data total. Regresi 16ogistic biner baik digunakan saat tidak ada multikolinearitas pada data. Berdasarkan korelasi antar variabel yang telah dijelaskan diatas, tidak ada variabel yang memiliki korelasi lebih dari 0,95 (multikolinearitas terjadi saat korelasi antar variabel lebih dari 0,95). Berikut adalah hasil klasifikasi dari *teting* data yang disajikan dalam *confussion matrix*.

Tabel 4.2*Confussion Matrix* Metode Reglog Biner

Aktual	**Prediksi**	
	0	**1**
0	62	56
1	25	125

Hasil klasifikasi dengan metode regresi logistic biner terdapat 56 pemain dengan usia karir dibawah 5 tahun yang salah terklasifikasi menjadi pemain dengan usia karir diatas 5 tahun. Sedangkan untuk pemain dengan usia karir diatas 5 tahun yang salah diklasifikasi menjadi pemain dengan usia karir dibawah 5 tahun adalah sebanyak 25 pemain. Berikut merupakan kurva ROC (*Receiver Operating Characteristic*) dari model yang telah didapatkan.



Gambar 4.5. Kurva *ROC* Model Regresi Logistik Biner

Dari kurva ROC diatas, dapat diketahui bahwa model regresi logistic biner yang didapatkan untuk pengklasifikasian sudah cukup baik karena garis kurva lebih cenderung mendekati titik 1,0 dibandingkan dengan garis *baseline.*

4.4. Klasifikasi Data Menggunakan Metode *Machine Learning*

Metode *machine learning* yang digunakan dalam analisis klasifikasi usia karir NBA *player* dalam penelitian ini adalah sebanyak empat metode yaitu metode *Naïve Bayes, Decision Tree, Support Vector Machine (SVM)* dan *Random Forest.*

Berikut adalah analisis untuk masing-masing metode yang digunakan.

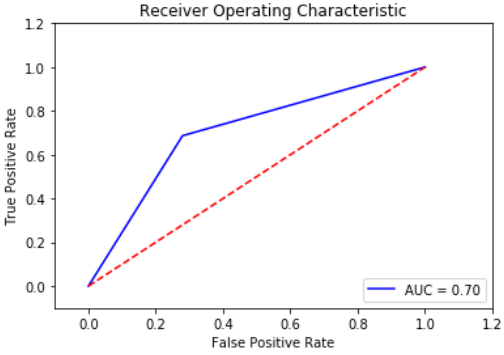
a) *Naïve Bayes*

Metode *machine learning* pertama yang digunakan untuk mengklasifikasikan NBA *player* berdasarkan usia karir (diatas atau dibawah 5 tahun) adalah metode *Naïve Bayes*. Dengan menggunakan data *testing* sebanyak 268 data NBA *players*, berikut merupakan *confussion matrix* hasil klasifikasi dari data *testing*.

Tabel 4.3*Confussion Matrix* Metode Naïve Bayes

Aktual	**Prediksi**	
	0	**1**
0	85	33
1	47	103

Hasil klasifikasi dengan metode *Naïve Bayes* terdapat 33 pemain dengan usia karir dibawah 5 tahun yang salah terklasifikasi menjadi pemain dengan usia karir diatas 5 tahun. Sedangkan untuk pemain dengan usia karir diatas 5 tahun yang salah diklasifikasi menjadi pemain dengan usia karir dibawah 5 tahun adalah sebanyak 47 pemain. Berikut merupakan kurva ROC (*Receiver Operating Characteristic*) dari model yang telah didapatkan.



Gambar 4.6. Kurva *ROC* Model Naïve Bayes

Dari kurva ROC diatas, dapat diketahui bahwa model *Naïve Bayes* yang didapatkan untuk pengklasifikasian sudah cukup baik karena garis kurva lebih cenderung mendekati titik

1,0dibandingkan dengan garis *baseline*.Hasil ini dapat dilihat dari nilai AUC (*Area Under Curve*) sebesar 0,7 dimana nilai tersebut mendekati 1.

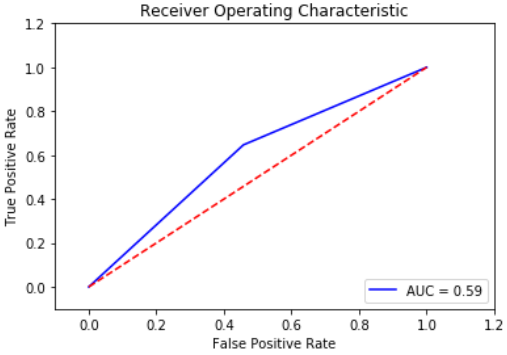
b) *Decision Tree*

Metode klasifikasi kedua yang digunakan dalam pengklasifikasian NBA *player* berdasarkan usia karir mereka adalah metode *Decision Tree*. Berikut merupakan hasil klasifikasi data *testing* berdasarkan metode klasifikasi *decision tree* yang disajikan dalam *confussion matrix*.

Tabel 4.4*Confussion Matrix* Metode *Decission Tree*

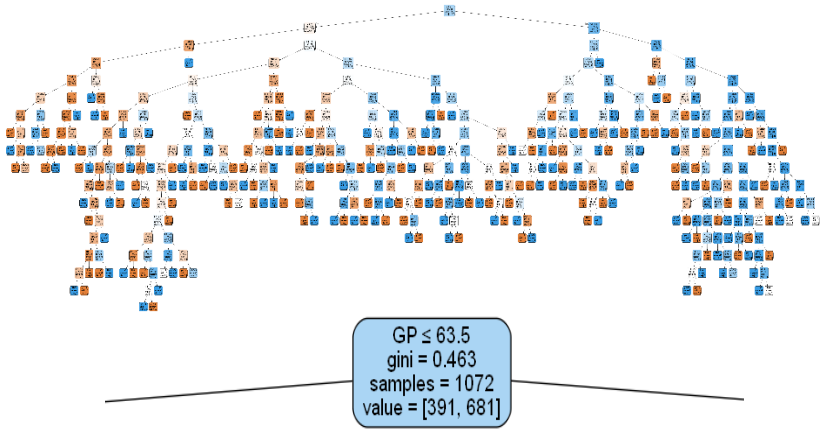
Aktual	**Prediksi**	
	0	**1**
0	64	54
1	53	97

Hasil klasifikasi dengan metode *Decisoin Tree* terdapat 54 pemain dengan usia karir dibawah 5 tahun yang salah terklasifikasi menjadi pemain dengan usia karir diatas 5 tahun. Sedangkan untuk pemain dengan usia karir diatas 5 tahun yang salah diklasifikasi menjadi pemain dengan usia karir dibawah 5 tahun adalah sebanyak 55 pemain. Hasil klasifikasi dengan metode ini menghasilkan lebih banyak misklasifikasi dibandingkan metode sebelumnya yaitu *Naïve Bayes*. Berikut merupakan kurva ROC (*Receiver Operating Characteristic*) dari model yang telah didapatkan.



Gambar 4.7. Kurva *ROC* Model *Decission Tree*

Model *Decission Tree* yang digunakan untuk klasifikasi NBA *player* tidak lebih baik daripada model yang sebelumnya didapatkan dari metode *Naïve Bayes.* Hal ini dapat dilihat dari kurva ROC dimana garis kurva lebih mendekati ke garis *baseline* jika dibandingkan dengan garis kurva dari mode *Naïve Bayes*, selain itu nilai AUC yang didapatkan adalah 0,59. Berikut ditampilkan grafik pohon keputusan (*decision tree)* dari data yang dianalisis.



Gambar 4.8. Visualisasi Pohon Keputusan

Dari visualisasi pohon keputusan di atas, dapat diketahui bahwa variabel yang menjadi *main root* atau akar utama dalam klasifikasi adalah variabel GP (jumlah pertandingan yang telah dialami pemain dalam satu musim kompetisi) dimana pemisahnya adalah berupa jumlah pertandingan yang telah dialami pemain dalam satu musim kompetisi apakah kurang atau dari 63,5 kali pertandingan.

c) *Support Vector Machine (SVM)*

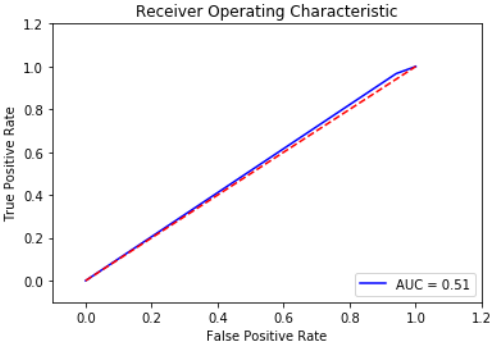
Metode klasifikasi selanjutnya yang digunakan adalah metode *Support Vector Machine (SVM).* Dengan menggunakan

data *testing* sebanyak 268 data NBA *players*, berikut merupakan *confussion matrix* hasil klasifikasi dari data *testing.*

Tabel 4.5 *Confussion Matrix* Metode SVM

Aktual	**Prediksi**	
	0	**1**
0	7	111
1	5	145

Hasil klasifikasi dengan metode SVM terdapat 111 pemain dengan usia karir dibawah 5 tahun yang salah terklasifikasi menjadi pemain dengan usia karir diatas 5 tahun. Hal ini berbanding terbalik dengan pemain dengan usia karir dibawah 5 tahun yang tepat diklasifikasikan yaitu hanya sebanyak 7 pemain. Sedangkan untuk pemain dengan usia karir diatas 5 tahun yang salah diklasifikasi menjadi pemain dengan usia karir dibawah 5 tahun hanya sedikit yaitu sebanyak 5 pemain. Hal ini menunjukkan bahwa model tidak baik digunakan untuk mengklasifikasi pemain dengan usia karir dibawah 5 tahun. Berikut merupakan kurva ROC (*Receiver Operating Characteristic*) dari model yang telah didapatkan.



Gambar 4.9. Kurva *ROC* Model SVM

Dapat dilihat dari kura ROC diatas bahwa garis kurva yaitu garis yang berwarna biru hamper sejajar dengan garis *baseline* yang ditandai dengan garis putus-putus warna merah. Garis kurva yang mendekati garis *baseline* menunjukkan bahwa model yang didapat dari metode SVM tidak baik digunakan untuk klasifikasi. Hal ini sejalan dari hasil klasifikasi yang

dapat dilihat pada *confussion matrix* diatas. Banyaknya misklasifikasi lebih banyak daripada klasifikasi yang tepat pada kelompok pemain dengan usia karir dibawah 5 tahun.

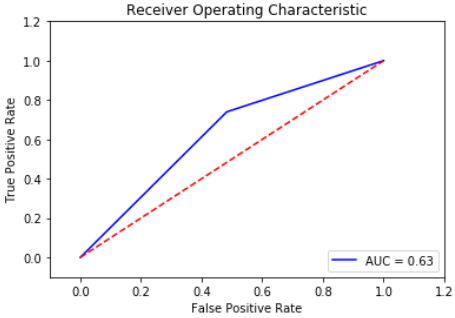
d) *Random Forest*

Metode klasifikasi terakhir yang digunakan dalam pengklasifikasian NBA *player* berdasarkan usia karir mereka adalah metode *Random Forest.* Berikut merupakan hasil klasifikasi data *testing* berdasarkan metode klasifikasi *Random Forest* yang disajikan dalam *confussion matrix*

Tabel 4.6*Confussion Matrix* Metode *Random Forest*

Aktual	**Prediksi**	
	0	**1**
0	61	57
1	39	111

Hasil klasifikasi dengan metode *Random Forest* terdapat 57 pemain dengan usia karir dibawah 5 tahun yang salah terklasifikasi menjadi pemain dengan usia karir diatas 5 tahun. Sedangkan untuk pemain dengan usia karir diatas 5 tahun yang salah diklasifikasi menjadi pemain dengan usia karir dibawah 5 tahun adalah sebanyak 39 pemain. Hasil klasifikasi dengan metode ini menghasilkan lebih sedikit misklasifikasi dibandingkan metode sebelumnya yaitu SVM. Berikut merupakan kurva ROC (*Receiver Operating Characteristic*) dari model yang telah didapatkan.



Gambar 4.10. Kurva *ROC* Model *Random Forest*

Dari kurva ROC diatas, dapat diketahui bahwa model *Random Forest* yang didapatkan untuk pengklasifikasian sudah cukup baik karena garis kurva lebih cenderung mendekati titik 1,0 dibandingkan dengan garis *baseline*.Hasil ini dapat dilihat dari nilai AUC (*Area Under Curve*) sebesar 0,63 dimana nilai tersebut mendekati 1.

4.5. Perbandingan Akurasi, Sensitifitas dan Specifisitas

Berikut merupakan perbandingan nilai akurasi, sensitifitas dan specifisitas dari metode-metode yang telah digunakan dalam analisis baik metode statistika maupun metode *machine learning.*

Tabel 4.7 Perbandingan Akurasi, Sensitifitas dan Specifisitas

	Regresi Logistik Biner	***Naïve Bayes***	***Decision Tree***	***SVM***	***Random Forest***
Akurasi	0,698	0,701	0,600	0,567	0,642
Sensifitas	0,525	0,525	0,542	0,525	0,517
Specifisitas	0,833	0,687	0,647	0,967	0,74

Berdasarkan tabel perbandingan nilai diatas, didapatkan kesimpulan bahwa metode yang paling baik digunakan untuk klasifikasi usia karir NBA *player* apakah diatas atau dibawah 5 tahun adalah dengan metode *Machine Learning Naïve Bayes* karena dapat menghasilkan ketepatan klasifikasi atau akurasi sebesar 0,701 atau sebesar 70,1%.

Halaman ini sengaja dikosongkan

BAB V
PENUTUP

5.1 Kesimpulan

Berdasarkan hasil analisis yang telah dilakukan maka kesimpulan dari hasil analisis adalah sebagai berikut.

1. *Features* atau variabel yang digunakan dalam analisis adalah variabel GP, MIN, PTS, FG%, 3PA, 3P%, FT% dan REB yang merupakan hasil dari *feature selection.*
2. Dari kedelapan variabel yang digunakan dalam analisis, antar variabel memiliki korelasi yang rendah (kurang dari 0,95) yang artinya tidak terdapat multikolinearitas.
3. Klasifikasi dengan metode regresi 25ogistic biner menghasilkan 81 misklasifikasi dari total 268 pengamatan.
4. Klasifikasi dengan metode *machine learning* dilakukan dengan empat metode yaitu metode *naïve bayes*, *decision tree*, *SVM* dan *random forest.* Dari hasil klasifikasi, diketahui bahwa metode naïve bayes merupakan metode dengan hasil misklasifikasi paling sedikit yaitu sebanyak 80 dari 268 pengamatan.
5. Dari kelima metode yang digunakan, metode *Naïve Bayes* merupakan metode terbaik untuk pengklasifikasian usia karir dari NBA *player* dengan hasil ketepatan klasifikasi atau akurasi sebesar 70,1%

5.2 Saran

Untuk mencari NBA *player* yang professional berdasarkan pengalaman permainannya dengan usia karir diatas 5 tahun, maka direkomendasikan para pelatih atau pencari bakat untuk menggunakan metode *Naïve Bayes* dalam penentuan klasifikasi.

Halaman ini sengaja dikosongkan

DAFTAR PUSTAKA

Abidin, T. F. (2012). *Naive Bayes Classifier. Bahan Kuliah Data Mining.* Program Studi Informatika FMIPA Universitas Syiah Kuala.

BMKG. (2017, Mei 7). Retrieved Juli 18, 2018, from http:/juanda.jatim.bmkg.go.id

Budiana, D., & Lubay, L. (2013). *Modul Pembelajaran Permainan Bola Basket.* Bandung: Fakultas Pendidkan Olahraga dan Kesehatan, Universuitas Pendidikan Indonesia.

Dash, M., & Liu, H. (1997). Feature Selection for Classification, Intelligent Data Analysis. 131-156.

Hair, J. F., Willian C.Black, B. J., Babin, & Anderson, R. E. (2010). *Multivariate Data Analysis* (Seventh Edition ed.). New Jersey: Pearson Prentice Hall.

Han, J., Kamber, M., & Pei, J. (2012). *Data Mining Concept and Techniques Third Edition.* USA: Morgan Kaufmann.

Hosmer, D., & Lemeshow, S. (2000). *Aplied Logistic Regression.* USA: John Wiley & Sons.

Kusrini, & E.T, L. (2009). *Algoritma Data Mining Edisi 1*. Yogyakarta: Andi Offset.

L, B., & Breiman, L. (2001). *Random Forest Machine Learning.* Belanda: Kluwer Academic Publisher.

Santoso, S. (2010). *Statistik Multivariat.* Jakarta: PT Elex Media Komputindo.

Walpole, E. R. (2013). *Pengantar Statistika.* Bandung: Alfabeta.

Halaman ini sengaja dikosongkan

LAMPIRAN

Lampiran

GP	MIN	PTS	FG%	3PA	3P%	FT%	REB	TARGET
36	27.4	7.4	34.7	2.1	25	69.9	4.1	0
35	26.9	7.2	29.6	2.8	23.5	76.5	2.4	0
74	15.3	5.2	42.2	1.7	24.4	67	2.2	0
58	11.6	5.7	42.6	0.5	22.6	68.9	1.9	1
48	11.5	4.5	52.4	0.1	0	67.4	2.5	1
75	11.4	3.7	42.3	1.1	32.5	73.2	0.8	0
62	10.9	6.6	43.5	0.1	50	81.1	2	1
48	10.3	5.7	41.5	1.5	30	87.5	1.7	1
65	9.9	2.4	39.2	0.5	23.3	71.4	0.8	0
42	8.5	3.7	38.3	0.3	21.4	67.8	1.1	0
35	6.9	2.3	36.5	0.1	33.3	81.8	0.9	0
40	6.7	3.6	39.8	0.6	13.6	77.6	1.2	1
27	6.6	1.3	47.2	0	0	28.6	2	1
45	15.3	5.6	32.3	3.6	30.1	86.1	2	0
44	6.4	2.4	53.7	0	0	50	1.4	1
40	6.1	2.6	51.4	0.4	14.3	68.4	0.4	1
49	5.3	2.1	37.6	0	0	64.2	1.2	0
41	4.2	1.7	34.8	0.3	21.4	73.1	0.3	0
82	37.2	19.2	49	0.3	22.7	82.9	11	0
82	37.2	19.2	49	0.3	22.7	82.9	11	1
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
80	15.8	4.3	43.3	0.2	14.3	79.2	1.2	0
68	12.6	3.9	35.8	0.7	16.7	79.4	1.5	1
43	12.1	5.4	55	0	0	64.3	3.8	0
52	12	4.5	43.9	0.2	10	62.5	0.7	1
47	11.7	4.4	36.9	1.3	33.3	67.3	0.9	1

Lampiran Syntax Python Untuk Pre Processing dan Feature Selection

```
import pandas as pd
import numpy as np
nba = pd.read_csv("nba.csv")
nba.head(n=20)
nba.isnull().sum()
nba.describe()
#imputasi data dengan mean
nba['3P%']=nba['3P%'].fillna(nba['3P%'].mean())
nba.isnull().sum()
nba.to_csv("nbaclean.csv",sep=',')
nba1= pd.read_csv("nbaclean.csv")
nba1.head()
X=nba1.iloc[:,2:21]
Y=nba1.iloc[:,21]
X, y= X, Y
print(X, y)
X.shape
from sklearn.svm import LinearSVC
from sklearn.feature_selection import SelectFromModel
from sklearn.ensemble import ExtraTreesClassifier
from sklearn.feature_selection import SelectKBest
from sklearn.feature_selection import chi2
clf = ExtraTreesClassifier(n_estimators=50)
lsvc = LinearSVC(C=0.01, penalty="l1", dual=False).fit(X, y)
model = SelectFromModel(lsvc, prefit=True)
X_new = model.transform(X)
X_new.shape
X=pd.DataFrame(X_new)
print(X)
X.to_csv("nba_neww.csv",sep=',')
```

Lampiran Syntax Python Untuk Statistika Deskriptif dan Klasifikasi

EKSPLORASI

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import seaborn as sb
data = pd.read_csv('nba_new.csv')
data.head()
sns.countplot(x=data['TARGET'])
plt.show()
g = sns.PairGrid(data, hue="'TARGET'")
g.map_diag(plt.hist)
g.map_offdiag(plt.scatter)
g.add_legend()
plt.show()
sb.heatmap(data.corr())
sb.boxplot(x='TARGET', y='GP', data=data, palette='hls')
sb.boxplot(x='TARGET', y='MIN', data=data, palette='hls')
sb.boxplot(x='TARGET', y='PTS', data=data, palette='hls')
sb.boxplot(x='TARGET', y='FG%', data=data, palette='hls')
sb.boxplot(x='TARGET', y='3PA', data=data, palette='hls')
sb.boxplot(x='TARGET', y='3P%', data=data, palette='hls')
sb.boxplot(x='TARGET', y='FT%', data=data, palette='hls')
sb.boxplot(x='TARGET', y='REB', data=data, palette='hls')
data.describe()
```

METODE STATISTIKA : REGLOG BINER

```
import numpy as np
import pandas as pd
import warnings
warnings.filterwarnings('ignore')
import matplotlib.pyplot as plt
import seaborn as sns
```

```
%matplotlib inline
dataset = pd.read_csv("nba_new.csv")
dataset.head()
dataset.shape
dataset.info()
x = dataset.iloc[:,0:8].values
print(x)
y= dataset.iloc[:,8].values
print(y)
# Split the data into Training and Testing set
from sklearn.cross_validation import train_test_split
x_train,x_test,y_train,y_test =
train_test_split(x,y,test_size=0.2,random_state=0)
#Fitting logistic regression to the training set
from sklearn.linear_model import LogisticRegression
classifier = LogisticRegression(random_state = 0)
classifier.fit(x_train,y_train)
# Predicting the Test set results
y_pred = classifier.predict(x_test)
y_pred
# Making the confusion matrix
from sklearn.metrics import confusion_matrix
cm = confusion_matrix(y_test,y_pred)
cm
from sklearn import metrics
print (metrics.accuracy_score(y_test,classifier.predict(x_test)))
total=sum(sum(cm))
#####from confusion matrix calculate accuracy
accuracy=(cm[0,0]+cm[1,1])/total
print ('Accuracy : ', accuracy)

sensitivity = cm[0,0]/(cm[0,0]+cm[0,1])
```

```
print('Sensitivity : ', sensitivity )

specificity = cm[1,1]/(cm[1,0]+cm[1,1])
print('Specificity : ', specificity)
from sklearn.metrics import confusion_matrix,
accuracy_score, precision_score, recall_score, auc, roc_curve
fpr, tpr, thresholds = roc_curve(y_test, y_pred)
auc = auc(fpr, tpr)
print('auc =', auc)
plt.figure()
plt.title('Receiver Operating Characteristic')
plt.plot(fpr, tpr, 'b',
label='AUC = %0.2f'% auc)
plt.legend(loc='lower right')
plt.plot([0,1],[0,1],'r--')
plt.xlim([-0.1,1.2])
plt.ylim([-0.1,1.2])
plt.ylabel('True Positive Rate')
plt.xlabel('False Positive Rate')
plt.show()
```

MACHINE LEARNING

NAIVE BAYES

```
from sklearn.naive_bayes import GaussianNB
gnb = GaussianNB()
x_train.shape, y_train.shape
x_test.shape, y_test.shape
gnb = GaussianNB().fit(x_train, y_train)
gnb_predictions = gnb.predict(x_test)
#Confusion matrix, Accuracy, sensitivity and specificity
from sklearn.metrics import confusion_matrix
```

```
cm1 = confusion_matrix(y_test, gnb_predictions)
print('Confusion Matrix : \n', cm1)
total1=sum(sum(cm1))
#####from confusion matrix calculate accuracy
accuracy1=(cm1[0,0]+cm1[1,1])/total1
print ('Accuracy : ', accuracy1)

sensitivity1 = cm1[0,0]/(cm1[0,0]+cm1[0,1])
print('Sensitivity : ', sensitivity )

specificity1 = cm1[1,1]/(cm1[1,0]+cm1[1,1])
print('Specificity : ', specificity1)
from sklearn.metrics import confusion_matrix,
accuracy_score, precision_score, recall_score, auc, roc_curve
fpr, tpr, thresholds = roc_curve(y_test, gnb_predictions)
auc = auc(fpr, tpr)
print('auc =', auc)
plt.figure()
plt.title('Receiver Operating Characteristic')
plt.plot(fpr, tpr, 'b',
label='AUC = %0.2f'% auc)
plt.legend(loc='lower right')
plt.plot([0,1],[0,1],'r--')
plt.xlim([-0.1,1.2])
plt.ylim([-0.1,1.2])
plt.ylabel('True Positive Rate')
plt.xlabel('False Positive Rate')
plt.show()
```

DECISION TREE

```
import pydotplus
from sklearn import tree
import collections
```

```
import graphviz
from sklearn.externals.six import StringIO
from IPython.display import Image
from sklearn.tree import export_graphviz
import pydotplus
data_feature_names=['GP','MIN','PTS','FG%','3PA','3P%','FT
%','REB']
print(data_feature_names)
#Training
clf=tree.DecisionTreeClassifier()
clf=clf.fit(x_train,y_train)
clfpredictions = clf.predict(x_test)
#Confusion matrix, Accuracy, sensitivity and specificity
from sklearn.metrics import confusion_matrix

cm2 = confusion_matrix(y_test, clfpredictions)
print('Confusion Matrix : \n', cm2)
total2=sum(sum(cm2))
#####from confusion matrix calculate accuracy
accuracy2=(cm2[0,0]+cm2[1,1])/total2
print ('Accuracy : ', accuracy2)

sensitivity2 = cm2[0,0]/(cm2[0,0]+cm2[0,1])
print('Sensitivity : ', sensitivity2 )

specificity2 = cm2[1,1]/(cm2[1,0]+cm2[1,1])
print('Specificity : ', specificity2)
dot_data=StringIO()
export_graphviz(clf,out_file=dot_data,feature_names=data_fe
ature_names,filled=True,rounded=True,special_characters=Tr
ue)
graph=pydotplus.graph_from_dot_data(dot_data.getvalue())
```

```
Image(graph.create_png())
from sklearn.metrics import confusion_matrix,
accuracy_score, precision_score, recall_score, auc, roc_curve
fpr, tpr, thresholds = roc_curve(y_test, clfpredictions)
auc = auc(fpr, tpr)
print('auc =', auc)
plt.figure()
plt.title('Receiver Operating Characteristic')
plt.plot(fpr, tpr, 'b',
label='AUC = %0.2f'% auc)
plt.legend(loc='lower right')
plt.plot([0,1],[0,1],'r--')
plt.xlim([-0.1,1.2])
plt.ylim([-0.1,1.2])
plt.ylabel('True Positive Rate')
plt.xlabel('False Positive Rate')
plt.show()
```

SVM

```
from sklearn.model_selection import train_test_split,
StratifiedKFold, GridSearchCV
from sklearn import datasets
from sklearn.metrics import confusion_matrix
from sklearn.ensemble import RandomForestClassifier,
BaggingClassifier, AdaBoostClassifier,
GradientBoostingClassifier
from sklearn import svm
clf =
svm.SVC(class_weight=None,C=1,gamma=0.1,kernel='rbf',ra
ndom_state=100).fit(x_train, y_train)
svm_predictions=clf.predict(x_test)
#Confusion matrix, Accuracy, sensitivity and specificity
from sklearn.metrics import confusion_matrix
```

```
total3=sum(sum(cm3))
#####from confusion matrix calculate accuracy
accuracy3=(cm3[0,0]+cm3[1,1])/total3
print ('Accuracy : ', accuracy3)

sensitivity3 = cm3[0,0]/(cm3[0,0]+cm3[0,1])
print('Sensitivity : ', sensitivity )

specificity3 = cm3[1,1]/(cm3[1,0]+cm3[1,1])
print('Specificity : ', specificity3)
cm3 = confusion_matrix(y_test, svm_predictions)
print('Confusion Matrix : \n', cm3)
from sklearn.metrics import confusion_matrix,
accuracy_score, precision_score, recall_score, auc, roc_curve
fpr, tpr, thresholds = roc_curve(y_test, svm_predictions)
auc = auc(fpr, tpr)
print('auc =', auc)
plt.figure()
plt.title('Receiver Operating Characteristic')
plt.plot(fpr, tpr, 'b',
label='AUC = %0.2f'% auc)
plt.legend(loc='lower right')
plt.plot([0,1],[0,1],'r--')
plt.xlim([-0.1,1.2])
plt.ylim([-0.1,1.2])
plt.ylabel('True Positive Rate')
plt.xlabel('False Positive Rate')
plt.show()
```

RANDOM FOREST

```
RF = RandomForestClassifier(random_state=123)
RF.fit(x_train,y_train)
rf_pred=RF.predict(x_test)
```

```
rf_pred
#Confusion matrix, Accuracy, sensitivity and specificity
from sklearn.metrics import confusion_matrix
cm4 = confusion_matrix(y_test, rf_pred)
print('Confusion Matrix : \n', cm4)
total4=sum(sum(cm4))
#####from confusion matrix calculate accuracy
accuracy4=(cm4[0,0]+cm4[1,1])/total4
print ('Accuracy : ', accuracy4)
sensitivity4 = cm4[0,0]/(cm4[0,0]+cm4[0,1])
print('Sensitivity : ', sensitivity4 )
specificity4 = cm4[1,1]/(cm4[1,0]+cm4[1,1])
print('Specificity : ', specificity4)
from sklearn.metrics import confusion_matrix,
accuracy_score, precision_score, recall_score, auc, roc_curve
fpr, tpr, thresholds = roc_curve(y_test, rf_pred)
auc = auc(fpr, tpr)
print('auc =', auc)
plt.figure()
plt.title('Receiver Operating Characteristic')
plt.plot(fpr, tpr, 'b',
label='AUC = %0.2f'% auc)
plt.legend(loc='lower right')
plt.plot([0,1],[0,1],'r--')
plt.xlim([-0.1,1.2])
plt.ylim([-0.1,1.2])
plt.ylabel('True Positive Rate')
plt.xlabel('False Positive Rate')
plt.show()
```