Analyzing Penguin Types Using the Ensemble Method

Menganalisa Jenis - Jenis Penguin Menggunakan Metode Ensemble.

1. Steven Ricardo Upa, 2. Leadericko Wisnu Wira Kejora, 3. Mentari Estefin M. Pangkey, 4. Christophel Juniar Ignatius Soge
Dept. of Electrical Engineering, Sam Ratulangi University Manado, Kampus Bahu St., 95115, Indonesia
e-mails : stevenupa026@student.unsrat.ac.id, leaderickokejora026@student.unsrat.ac.id, mentaripangkey026@student.unsrat.ac.id, christophelsoge026@student.unsrat.ac.id

Abstract* — There are many types of penguins living in the world and the dataset we use classifies 4 types of penguins based on their physical differences. We use 3 kinds of base modules, namely; K-Nearest Neighbor Method, Naive Bayes Classifier Method and Regression Linear Method. The three methods will produce a percentage of accuracy, precision, recall and F1 score. After that, the dataset is processed using Ensemble Techniques which consists of; Averaged and Voting. The outputs of the two Ensemble Techniques used are accuracy, precision, recall and F1 Score which is a combination of the 3 Base Modules used. In the percentage generated by some of the methods mentioned above, we can see that some of the methods in the Ensemble Techniques will show more accurate results.

***Key words* — Machine Learning, Ensemble Techniques, Base Module, Voting Classifier, Averaged Method.**

Abstrak* — Ada banyak jenis penguin yang hidup di dunia dan dataset yang kita gunakan mengklasifikasikan 4 jenis penguin berdasarkan perbedaan fisiknya. Kami menggunakan 3 macam *base module* yaitu; *K-Nearest Neighbor Method, Naive Bayes Classifier Method* dan *Regression Linear Method.* Ketiga metode tersebut akan menghasilkan persentase dari akurasi, presisi, *recall* dan Skor F1. Setelah itu dataset diolah dengan menggunakan *Ensemble Techniques* yang terdiri dari; *Averaged* dan *Voting.* Keluaran dari kedua *Ensemble Techniques* yang digunakan adalah akurasi, presisi, *recall* dan Skor F1 yang merupakan gabungan dari 3 *Base Module* yang digunakan. Dalam persentase yang dihasilkan oleh beberapa metode yang sudah disebutkan diatas, kita bisa melihat bahwa beberapa metode dalam *Ensemble Techniques* akan memperlihatkan hasil yang lebih akurat.

***Kata kunci* — Machine Learning, Ensemble Techniques, Base Module, Voting Classifier, Averaged Ensemble Method.**

I. PENDAHULUAN

Terdapat banyak jenis Penguin didunia dan perbedaan yang sangat nampak bisa terlihat pada bentuk fisik dari jenis penguin yang ada. Bentuk fisik dapat berupa panjang sirip, bentuk paruh, berat badan, habitat tinggal dan masih banyak lagi. Dalam dataset yang kita gunakan ciri-ciri tersebut diklasifikasikan agar bisa menentukan jenis penguin yang tepat.

Didalam laporan ini, kita akan memprediksi jenis-jenis Penguin dengan melihat faktor – faktor dari yang diberikan dari dataset. Setelah melihat beberapa ciri – ciri tersebut, maka dataset yang dibagi menjadi data training dan data test tersebut akan memprediksi dalam data test sesuai dengan ciri – ciri yang diberikan maka kemungkinan jenis atau *Species* penguin tersebut.

Dengan metode tersebut, maka kita bisa memprediksi spesies penguin apa seteleh ciri – cirinya disebutkan. Namun tidak hanya menggunakan satu metode saja, tapi dengan tiga metode *Base Module* serta digabungkan dengan metode *Ensemble Techniques* agar akurasi dan presisinya dapat dimaksimalkan lebih lagi.

II. DATASET YANG DIGUNAKAN

Dataset yang kita gunakan berupa beberapa ciri – ciri dari penguin – penguin yang sudah diambil datanya berupa spesies, pulau, Panjang dan dalamnya paruh, Panjang sayap, berat badan dan jenis kelamin. Seperti yang bisa dilihat pada contoh berikut ini.

species	island	bill_length	bill_depth_	flipper_len	body_mass	sex
Adelie	Torgersen	39.1	18.7	181	3750	Male
Adelie	Torgersen	39.5	17.4	186	3800	Female
Adelie	Torgersen	40.3	18	195	3250	Female
Adelie	Torgersen	36.7	19.3	193	3450	Female
Adelie	Torgersen	39.3	20.6	190	3650	Male
Adelie	Torgersen	38.9	17.8	181	3625	Female
Adelie	Torgersen	39.2	19.6	195	4675	Male
Adelie	Torgersen	41.1	17.6	182	3200	Female
Adelie	Torgersen	38.6	21.2	191	3800	Male
Adelie	Torgersen	34.6	21.1	198	4400	Male
Adelie	Torgersen	36.6	17.8	185	3700	Female
Adelie	Torgersen	38.7	19	195	3450	Female

Dataset tersebut akan dibagi menjadi *variable* x yang memegang nilai semua kolom kecuali *species* dan *varible* y yang memegang nilai dari kolom *species*. Dengan jumlah data sebanyak 333 baris dengan data test akan mengambil sebanyak 10% dari jumlah data, maka akan terbagi menjadi 299 baris data training dan 34 baris data test

III. METODE YANG DIGUNAKAN & HASIL

Dua jenis metode akan digunakan dalam menguji dataset dari penguin yang ada. Kita akan menggunakan *Base Module* yang didalamnya terdapat *K-Nearest Neighbor Method, Naive Bayes Classifier Method* dan *Regression Linear Method*. Lalu *Ensemble Techniques* yang menggunakan *Average Method* dan *Voting Classifier*.

a. Base Module

- *K-Nearest Neighbor Method*

Dengan Menggunakan metode KNN, kita mengambil parameter data dari data training, data test dari *variable* **x** lalu masukan juga data training dari *variable* **y**. Lalu dengan memasukkan parameter K = 1 maka kita akan mendapatkan hasil sebagai berikut yang ditampilkan dalam Matrix

prediction actual	Adelie	Chinstrap	Gentoo
Adelie	14	0	0
Chinstrap	1	4	0
Gentoo	0	0	15

Pada gambar diatas, kita dapat melihat bahwa hasil prediksi dari data test menghasilkan 15 Adelie, 4 Chinstrap dan 15 Gentoo. Namun pada hasil *acutal*-nya terdapat 14 Adelie, 5 Chinstrap dan 15 Gentoo. Dimana hasil tersebut memiliki satu prediksi yang salah maka persentase dari gambar diatas adalah :

```
Akurasi model:  97.05882352941177 %
Presisi model:  97.05882352941177 %
Recall:  0.9705882352941176
Skor F1:  0.9705882352941176
```

	precision	recall	f1-score	support
Adelie	0.93	1.00	0.97	14
Chinstrap	1.00	0.80	0.89	5
Gentoo	1.00	1.00	1.00	15
accuracy			0.97	34
macro avg	0.98	0.93	0.95	34
weighted avg	0.97	0.97	0.97	34

- *Naive Bayes Classifier Method*

Untuk menggunakan metode *Naïve Bayes*, kita mengimpor *MultinomialNB* yang akan melakukan *pre-processing* dataset yang kita punya. Lalu kita hanya perlu memanggil metode *predict* dari library tersebut dengan memasukkan parameter data test dari *variable* **x**, maka akan menghasilkan hasil sebagai berikut.

prediction actual	Adelie	Chinstrap	Gentoo
Adelie	11	1	2
Chinstrap	1	4	0
Gentoo	1	0	14

Dapat dilihat pada *confusion matrix* diatas, hasil prediksi yang dilakukan metode ini memperlihatkan 13 Adelie, 5 Chinstrap dan 16 Gentoo. Sedangkan pada *actual*-nya memperlihatkan 14 Adelie, 5 Chinstrap dan 15 Gentoo. Hasil prediksi dari metode ini memperlihatkan beberapa prediksi yang salah sehingga persentase dari metode ini dapat dilihat pada gambar dibawah ini.

```
Akurasi model:  85.29411764705883 %
Presisi model:  85.29411764705883 %
Recall:  0.8529411764705882
Skor F1:  0.8529411764705882
```

	precision	recall	f1-score	support
Adelie	0.85	0.79	0.81	14
Chinstrap	0.80	0.80	0.80	5
Gentoo	0.88	0.93	0.90	15
accuracy			0.85	34
macro avg	0.84	0.84	0.84	34
weighted avg	0.85	0.85	0.85	34

- *Regression Linear Method*

Lalu *Base Module* yang digunakan yaitu *Regression Linear*. Menggunakan library *LogisticRegression* yang akan melakukan *pre-processing* dari data training *variable* **x** dan **y** lalu memanggil metode *predict* dengan parameter data test *variable* **x**. Hasilnya muncul sebagai berikut.

prediction actual	Adelie	Chinstrap	Gentoo
Adelie	14	0	0
Chinstrap	0	5	0
Gentoo	0	0	15

Dapat kita lihat bahwa hasil dari prediksi sesuai dengan hasil dari *actual*-nya maka hasil persentasi yang muncul akan menjadi seperti berikut.

```
Akurasi model:  100.0 %
Presisi model:  100.0 %
Recall:  1.0
Skor F1:  1.0
```

	precision	recall	f1-score	support
Adelie	1.00	1.00	1.00	14
Chinstrap	1.00	1.00	1.00	5
Gentoo	1.00	1.00	1.00	15
accuracy			1.00	34
macro avg	1.00	1.00	1.00	34
weighted avg	1.00	1.00	1.00	34

b. Ensemble Techniques

- *Average Method*

  Metode *Average* akan menggabungkan hasil prediksi dari ketiga *Base Module* yang telah kita gunakan sebelumnya lalu menghasilkan hasil dari rata – rata ketiga metode tersebut. Maka kita hanya perlu menggabungkan hasil akurasi dan presisi dari ketiga metode dan membaginya dengan jumlah metode yang digunakan. Maka kita akan menghasilkan persentase seperti berikut.

```
Akurasi model:  94.11764705882352 %
Presisi model:  94.11764705882352 %
Recall:  0.9411764705882353
Skor F1:  0.9411764705882353
```

  Hasil dari menggunakan *Average Method* bisa kita lihat pada gambar diatas, dengan persentasi dari akurasi dan presisi sebesar 94%.

- *Voting Classifier Method*

  Metode *Voting Classifier* akan menggunakan library yang tersedia yang diimpor dari *sklearn.ensemble* dengan nama impor yaitu *VotingClassifier*. Setelah kita menentukan beberapa parameter yang digunakan, maka akan tampil sebagai berikut.

```
ACCURACY SCORE:
0.9706
CLASSIFICATION REPORT:
              Adelie  Chinstrap  Gentoo  accuracy  macro avg  weighted avg
precision   0.933333   1.000000     1.0  0.970588   0.977778      0.972549
recall      1.000000   0.800000     1.0  0.970588   0.933333      0.970588
f1-score    0.965517   0.888889     1.0  0.970588   0.951469      0.969461
support    14.000000   5.000000    15.0  0.970588  34.000000     34.000000
```

  Hasil akurasi dan presisi sebesar 97% dengan rincian dari masing – masing spesies terlihat pada gambar diatas

IV. KESIMPULAN

Dapat disimpulkan bahwa dengan menggunakan metode diatas dengan menggunakan dataset jenis-jenis dari penguin dapat dibedakan dengan memanfaatkan perbedaan fisiknya. Hal ini dapat dilakukan dengan membandingkan penguin satu sama lain agar kita dapat memprediksi Penguin mana yang cocok dengan bentuk fisik yang di uji.

Metode yang digunakan yaitu *Base Module* serta *Ensemble Techniques* juga terbukti akurat dan presisi jika diimplementasikan kedalam pengujian yang dilakukan dimana kita dapat melihat kecocokan dari Akurasi model dan Presisi model memiliki kecocokan 100% atau mendekati 100%.

V. REFERENSI

[1] Ardabili S., Mosavi A., Várkonyi-Kóczy A.R. (2020) *Advances in Machine Learning Modeling Reviewing Hybrid and Ensemble Methods.* In: Várkonyi-Kóczy A. (eds) Engineering for Sustainable Future. INTER-ACADEMIA 2019. Lecture Notes in Networks and Systems, vol 101. Springer, Cham. https://doi.org/10.1007/978-3-030-36841-8_21USA: Abbrev. of Publisher, year, ch.x, sec. x, pp. xxx–xxx.

[2] *Ensemble Machine Learning* "Cha Zhang Yunqian Ma" DOI https://doi.org/10.1007/978-1-4419-9326-7 eBook ISBN 978-1-4419-9326-7 Edition Number 1 Number of Pages VIII, 332W.-K.Chen,*LinearNetworksandSystems.*Belmont, CA:Wadsworth, 1993, pp. 123–135.

[3] Yong Zhang,1 Hongrui Zhang,1 Jing Cai ,1 and Binbin Yang1 *Artificial Intelligence and Data Mining 2014* https://doi.org/10.1155/2014/376950

[4] Liyang Wei, Yongyi Yang, R. M. Nishikawa and Yulei Jiang, "*A study on several Machine-learning methods for classification of Malignant and benign clustered microcalcifications,*" in IEEE Transactions on Medical Imaging, vol. 24, no. 3, pp. 371-380, March 2005, doi: 10.1109/TMI.2004.842457.

[5] Hussain, M., Zhu, W., Zhang, W. et al. *Using machine learning to predict student difficulties from learning session data*. Artif Intell Rev 52, 381–407 (2019). https://doi.org/10.1007/s10462-018-9620-8