**BIOINFORMATIKA:**

**PREDIKSI STRUKTUR SEKUNDER PROTEIN**

**MENGGUNAKAN SUPPORT VECTOR MACHINE PADA**

**DATASET RS126**



Muhammad Ilham Ibadurrohman

14/366093/PA/16211

**PROGRAM STUDI ILMU KOMPUTER**

**DEPARTEMEN ILMU KOMPUTER DAN ELEKTRONIKA**

**FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM**

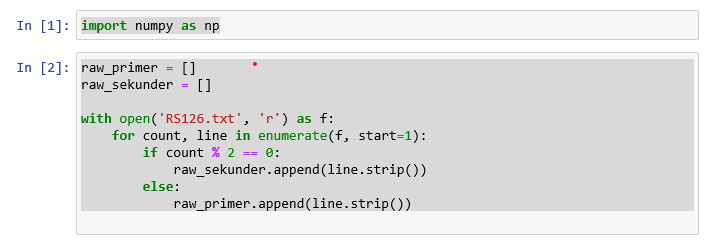
**UNIVERSITAS GADJAH MADA**

2019

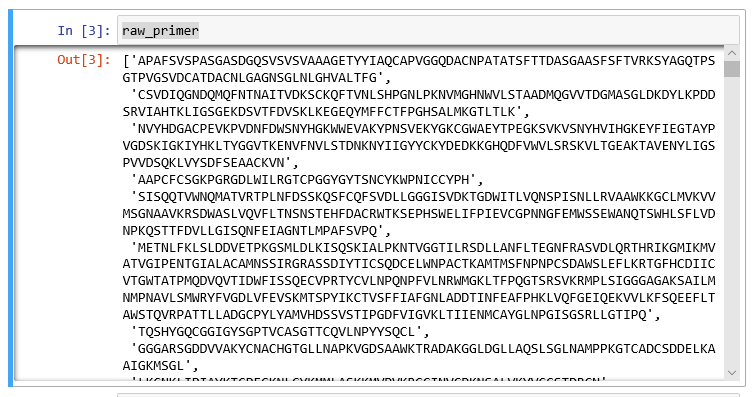
GitHub Link: <https://github.com/nopedude/SVM_PSSP/tree/master/SVM_PSSP>

**PENGOLAHAN DATA :**

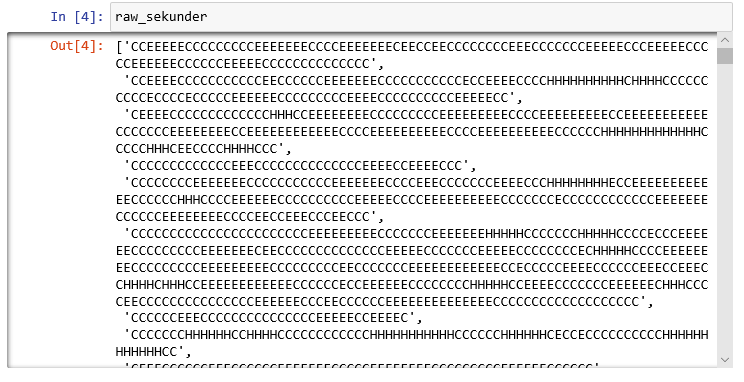
Dataset RS126 dibuka dan diolah dengan menggunakan bahasa pemrograman python. dari dataset tersebut, line ganjil merujuk pada struktur primer protein dan line genap merujuk pada struktur sekunder dari struktur primer.



Contoh pembacaan data struktur primer:



Contoh pembacaan data struktur sekunder:

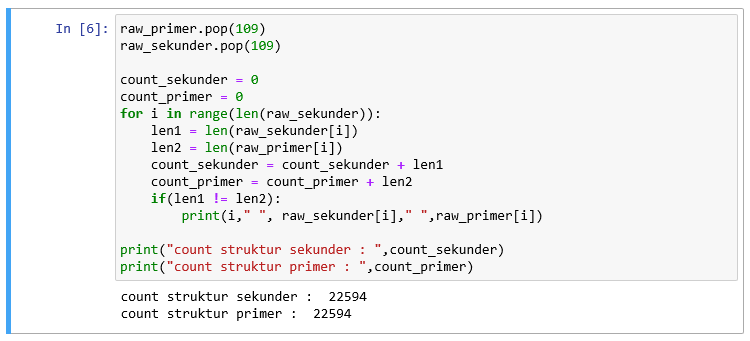


**PENCARIAN DATA YANG TIDAK LENGKAP:**

Pengecekan terhadap data dilakukan untuk melihat apakah terdapat perbedaan jumlah karakter struktur primer dan sekunder (karena prediksi struktur sekunder termasuk ke dalam permasalahan sequence labelling, maka dipastikan jumlah karakter dari struktur primer dan struktur sekunder selalu sama).

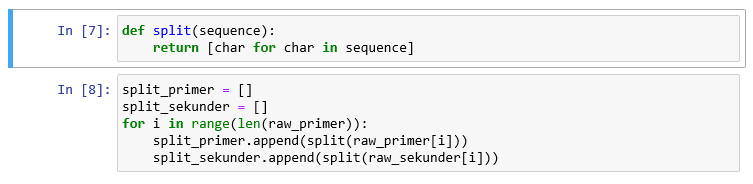


Setelah dilakukan pengecekan, ditemukan bahwa data pada urutan urutan 109 memiliki panjang struktur primer dan sekunder yang berbeda. Karena hanya ada 1 data maka data struktur primer dan struktur sekunder dari urutan ke 109 tersebut kita hapus dengan cara melakukan operasi pop.



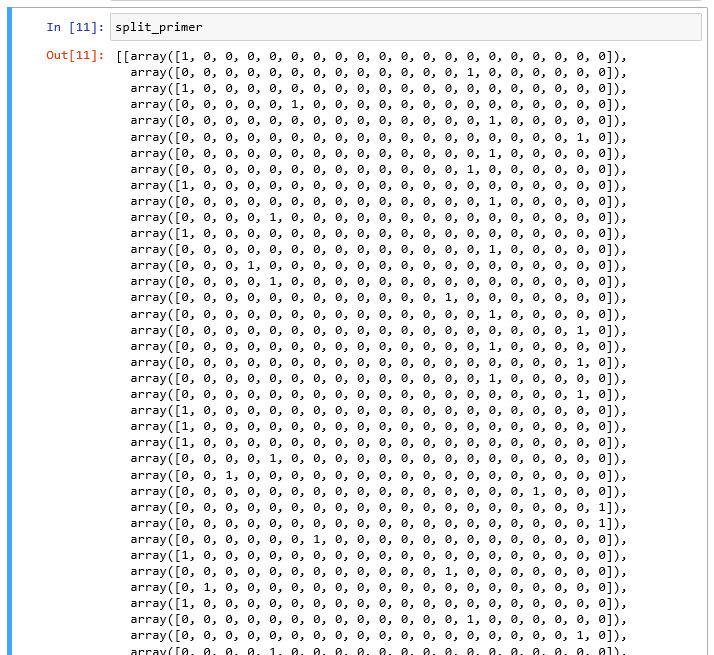
**ORTHOGONAL ENCODING**:

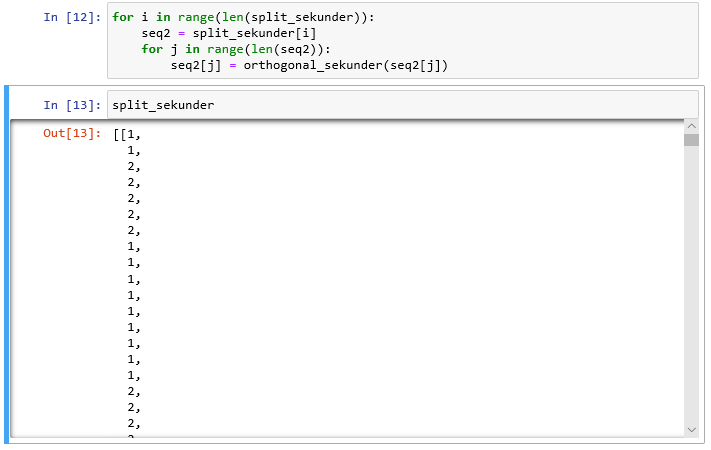
Target Labeling Untuk setiap data struktur primer dan sekunder dilakukan split sehingga dapat diencode kedalam bentuk orthogonal.



Hasil operasi split struktur primer dan struktur sekunder protein kemudian diubah kedalam bentuk Orthogonal Encoding dan Target Labeling. Cuplikan switch case untuk setiap asam amino pada struktur primer protein ditampilkan sebagai berikut: 







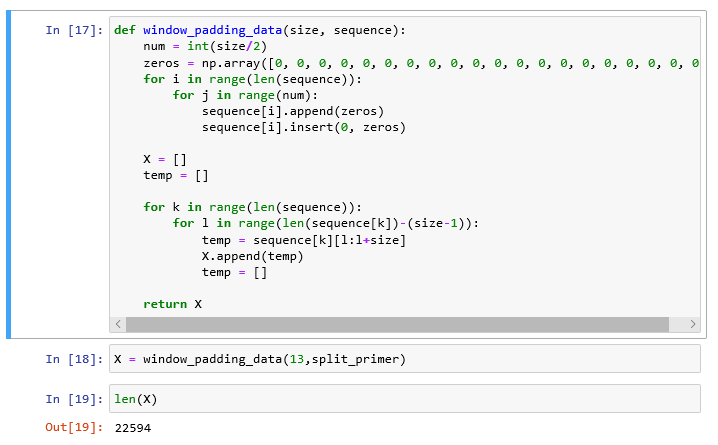


**DATASET**

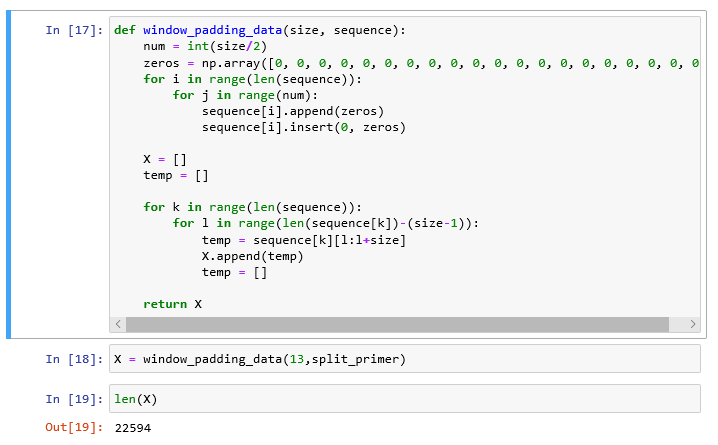
Label dan Fitur Setiap hasil split dari struktur sekunder protein akan dijadikan label/target class. Dari hasil operasi tersebut didapatkan total keseluruhan label berjumlah 22.594. Angka ini akan digunakan untuk memastikan bahwa dataset juga memiliki jumlah fitur yang sama.



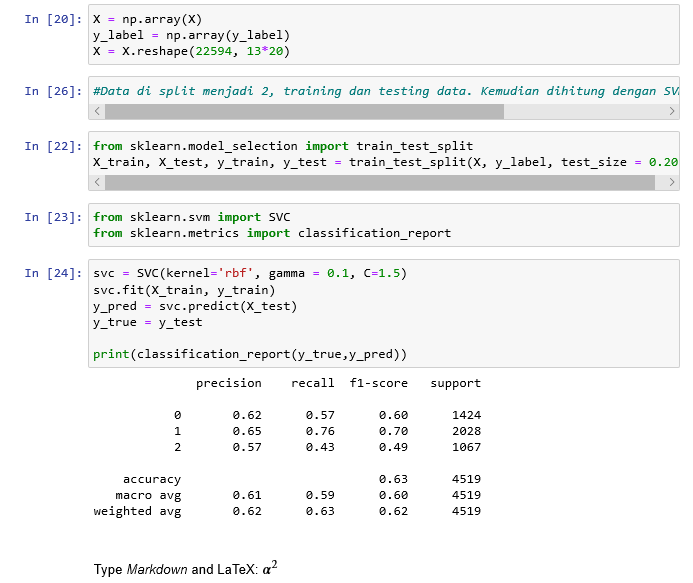
Pembentukan fitur data dilakukan dengan menggunakan fungsi window\_padding\_data. Fungsi ini akan menerima ukuran dari sliding window dan sekuens struktur primer protein. Pada fungsi ini, fitur akan diolah seperti penambahan padding 0 di awal dan diakhir dan pengambilan fitur-fitur hasil windowing sehingga data keluaran bisa langsung dilatih pada model SVM.



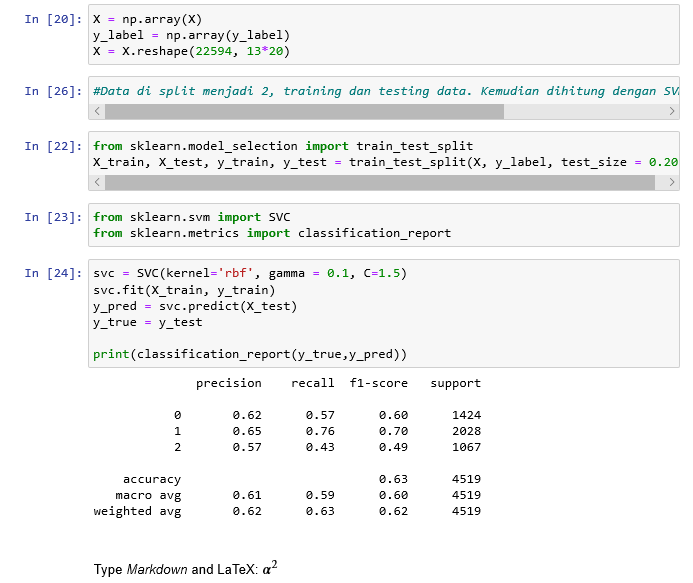
Contoh penggunaan fungsi padding :



Dimana data struktur primer protein akan diolah dengan variabel ukuran sliding window SVM. Sebelum dimasukkan ke model SVM Scikit-Learn, data di reshape untuk mengikuti ukuran input dari model.



Sebelum data dimasukkan ke model SVM Scikit-Learn, data di reshape untuk mengikuti ukuran input dari model. Data direshape menjadi ukuran 22.594 data dikali dengan 100 (x ukuran window dan 20 ukuran orthogonal encoding.)



**HASIL:**

|  |  |
| --- | --- |
| Ukuran Window | Classification Report |
| 3 |  |
| 5 |  |
| 7 |  |
| 9 |  |
| 11 |  |
| 13 |  |
| 15 |  |
| **17**  **(Terbaik)** |  |
| 19 |  |