Tập dữ liệu Yeast trên UCI Machine Learning Repository chứa thông tin về khả năng sinh tồn của một loại men vi sinh trong các điều kiện thích hợp và không thích hợp. Tập dữ liệu này bao gồm 1484 mẫu và 8 thuộc tính số. Các thuộc tính trong tập dữ liệu bao gồm: 1. Mcg: tỷ lệ GC (guanine và cytosine) của chuỗi nucleotide 2. Gvh: số lượng acid amin trong protein được mã hoá bởi DNA có tính chất kém oxi hóa (hay glutamine-histidine-valine) 3. Alm: độ sâu của cluster hydrophobic amino acids 4. Mit: xác suất rằng gen có liên quan đến mitochondria 5. Erl: xác suất rằng gene này là ER-localized. 6. Pox: Tỷ lệ proline, phenylalanine, and tyrosine residues; 7. VAC : Probability of nuclear localization signal in the N-terminal 8. Nuc : Score the potential for non-classical nuclear localization signals Như vậy, tập dữ liệu Yeast cung cấp thông tin chi tiết về các yếu tố di truyền quan trọng ảnh hưởng đến khả năng sinh tồn của men vi sinh trong môi trường thử nghiệm,nó có thể được áp dụng để huấn luyện các model phân loại để xác định khả năng sống sót của men vi sinh trong các điều kiện khác nhau..

Tập dữ liệu Yeast bao gồm 8 thuộc tính sau:

mcg (Mean of positional weighted matrix score): Điểm số trung bình của ma trận trọng số vị trí tính toán từ các quá trình nhận dạng mô-típ trên chuỗi protein.

gvh (Gradient vector of hydrophobicity): Độ lớn của gradient của hàm thủy phân trên chuỗi protein.

alm (Score of the ALOM membrane spanning regions prediction program): Điểm số của chương trình dự đoán các khu vực bám màng của ALOM trên chuỗi protein.

mit (Score of discriminant analysis of the amino acid content of the N-terminal region): Điểm số của phân tích phân biệt nội dung axit amin trong khu vực N-đầu của chuỗi protein.

erl (Presence of "HDEL" substring (thought to act as a retention signal for ER membrane)) : Sự hiện diện của chuỗi con "HDEL" (được cho là tác dụng như một tín hiệu giữ lại màng ER) trên chuỗi protein.

pox (Score of human overlapped homologous extension): Điểm số của phần mở rộng đồng hợp của người chồng lên nhau trên chuỗi protein.

vac (Score of the discriminant analysis of the amino acid content of the vacuolar and extracellular proteins): Điểm số của phân tích phân biệt nội dung axit amin của các protein ngoài tế bào và protein thuỷ phân trên chuỗi protein.

nuc (Score of discriminant analysis of nuclear localization signals of nuclear and non-nuclear proteins): Điểm số của phân tích phân biệt tín hiệu địa phương hóa hạt nhân của protein hạt nhân và không phải hạt nhân trên chuỗi protein.

Tất cả các thuộc tính này đều liên quan đến các đặc điểm về cấu trúc và tính chất của các protein. Các thuộc tính này có thể được sử dụng để dự đoán chức năng của protein, vị trí tế bào của chúng, và các đặc tính khác của chúng. Tập dữ liệu này có thể được sử dụng trong nhiều lĩnh vực, bao gồm sinh học phân tử, tế bào học, y học, và khoa học máy tính.

0: CYT

1: NUC

2: ERL

3: MIT

4: ME3

5: ME2

6: ME1

7: EXC

8: POX

9:VAC