NATIONAL INSTITUTE OF TECHNOLOGY HAMIRPUR (H.P.)

Structural Protein Sequences

Submitted To:
Dr. Robin Singh Badhoria

Submitted By:
Ritu Mehta

CONTENTS:

INTRODUCTION:

1) The Protein Data Bank (PDB) is a repository for atomic coordinates and other information about proteins and biological macromolecules.
2) Structural biologists use techniques like X-ray crystallography, NMR spectroscopy, and cryo-electron microscopy to determine atomic locations in molecules.
3) The archive includes diverse structures such as ribosomes, oncogenes, drug targets, and viruses.

DATA DESCRIPTION:

1) **Data size:**
   27.29MB data

1) **Structure:**
   140911 rows
   14 columns

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	structureI	classificati	experimer	macromol	residueCo	resolution	structureM	crystalliza	crystalliza	densityMa	densityPer	pdbxDetai	phValue	publicationYear	
2	100D	DNA-RNA	X-RAY DIF	DNA/RNA	20	1.9	6360.3	VAPOR DIF	FUSION, H	1.78	30.89	pH 7.00, V	7	1994	
3	101D	DNA	X-RAY DIF	DNA	24	2.25	7939.35			2	38.45			1995	
4	101M	OXYGEN T	X-RAY DIF	Protein	154	2.07	18112.8			3.09	60.2	3.0 M AMI	9	1999	
5	102D	DNA	X-RAY DIF	DNA	24	2.2	7637.17	VAPOR DII	277	2.28	46.06	pH 7.00, V	7	1995	
6	102L	HYDROLA	X-RAY DIF	Protein	165	1.74	18926.6			2.75	55.28			1993	

TOOLS AND TECHNOLOGIES:

- **Python**: Primary programming language for data analysis, model development, and processing.
- **Pandas and NumPy**: Libraries for data manipulation, preprocessing, and handling large datasets efficiently.
- **Sci-kit Learn**: Used for implementing classical machine learning models like Logistic Regression and SVM.
- **TensorFlow and Keras**: Frameworks for building, training, and evaluating deep learning models, Used for developing the sequential neural network model with layers optimized for high-dimensional genetic data.

- **Matplotlib and Seaborn**: Visualization libraries used to graphically represent model accuracy, loss, and performance comparisons across different diseases.
- **StandardScaler** (from Sci-kit Learn): Standardizes features by removing the mean and scaling to unit variance, which improves model convergence.
- **CSV File Handling**: Exported processed and transformed datasets to CSV format for ease of re-use in model training and testing.

EVALUATION METRICS:

1. Accuracy: The percentage of correctly predicted instances (both true positives and true negatives) out of all instances.

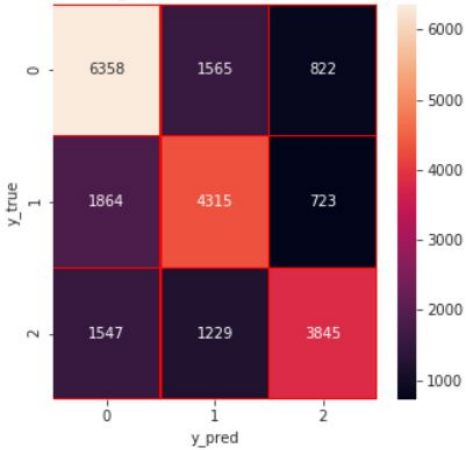
2. Precision: The ratio of true positives to the sum of true positives and false positives. Measures the model's ability to correctly identify positive instances. It's important when false positives have significant consequences.

3. Recall (Sensitivity): The ratio of true positives to the sum of true positives and false negatives. Measures the model’s ability to correctly identify all positive instances. It is crucial when false negatives are costly.

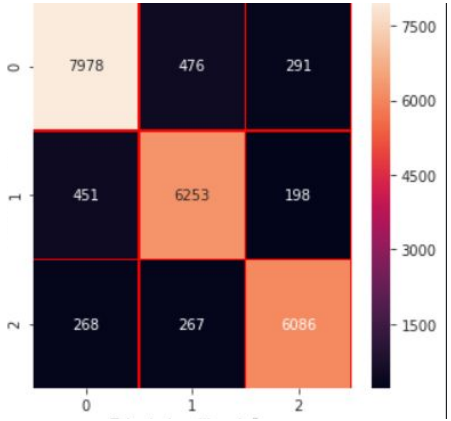
4. F1-Score: The harmonic mean of precision and recall, calculated as 2 * (Precision * Recall) / (Precision + Recall). A balanced metric that is particularly useful when the class distribution is imbalanced. It provides a single value that captures both precision and recall.

MODELS:

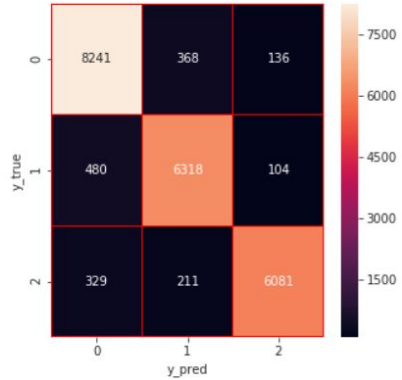
1) KNN Algorithm
2) Decision Trees
3) Random Forests Algorithm
4) Naive Bayes
5) Logistic Regression
6) SVM



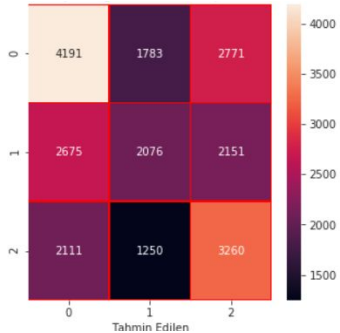
KNN



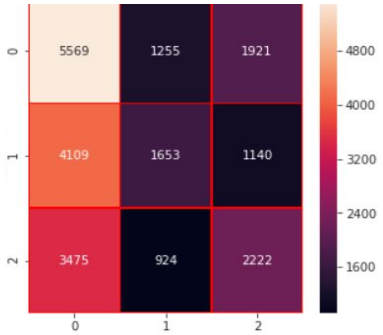
Decision Tree



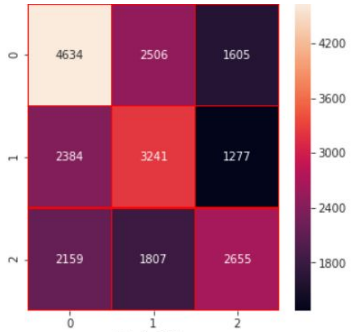
Random Forests



Naive Bayes



Logistic Regression



SVM

RESULTS:

1) KNN

	precision	recall	f1-score	support
1	0.65	0.73	0.69	8745
2	0.61	0.63	0.62	6902
3	0.71	0.58	0.64	6621
accuracy			0.65	22268
macro avg	0.66	0.64	0.65	22268
weighted avg	0.66	0.65	0.65	22268

1) Decision Tree

	precision	recall	f1-score	support
1	0.92	0.91	0.91	8745
2	0.89	0.91	0.90	6902
3	0.93	0.92	0.92	6621
accuracy			0.91	22268
macro avg	0.91	0.91	0.91	22268
weighted avg	0.91	0.91	0.91	22268

3) Random Forests

	precision	recall	f1-score	support
1	0.91	0.94	0.93	8745
2	0.92	0.92	0.92	6902
3	0.96	0.92	0.94	6621
accuracy			0.93	22268
macro avg	0.93	0.93	0.93	22268
weighted avg	0.93	0.93	0.93	22268

4) Naive Bayes

	precision	recall	f1-score	support
1	0.47	0.48	0.47	8745
2	0.41	0.30	0.35	6902
3	0.40	0.49	0.44	6621
accuracy			0.43	22268
macro avg	0.42	0.42	0.42	22268
weighted avg	0.43	0.43	0.42	22268

5) Logistic Regression

	precision	recall	f1-score	support
1	0.42	0.64	0.51	8745
2	0.43	0.24	0.31	6902
3	0.42	0.34	0.37	6621
accuracy			0.42	22268
macro avg	0.43	0.40	0.40	22268
weighted avg	0.43	0.42	0.41	22268

6) SVM

	precision	recall	f1-score	support
1	0.50	0.53	0.52	8745
2	0.43	0.47	0.45	6902
3	0.48	0.40	0.44	6621
accuracy			0.47	22268
macro avg	0.47	0.47	0.47	22268
weighted avg	0.47	0.47	0.47	22268

CONCLUSIONS:

Random Forest: Accuracy of 93%, with high precision, recall, and f1-scores across all classes, particularly excelling in class 3. This model consistently shows strong performance with balanced results across the metrics.

Decision Trees: Accuracy of 91%, also performing very well with high precision and recall for all classes, though slightly lower than Random Forest.

KNN: Accuracy of 65%, with balanced precision and recall, but the performance is lower compared to both Decision Trees and Random Forest.

Naive Bayes: Accuracy of 43%, with poor performance overall, especially for class 2, indicating that it is not well-suited for this particular dataset.

Logistic Regression: Accuracy of 42%, with low f1-scores and recall, indicating that this model is also not effective.

SVM: Accuracy of 47%, with average performance and poor recall for class 3, making it less competitive compared to the other models.

REFERENCES:

1) https://www.sciencedirect.com/science/article/abs/pii/S0300908412000405
1) https://www.sciencedirect.com/science/article/abs/pii/S0031320306000756
1) https://www.sciencedirect.com/science/article/abs/pii/S0006291X06017050
1) https://www.sciencedirect.com/science/article/abs/pii/S0022519303931791

THANK YOU!