REPORT

Hello seniors!!  
This assignment did challenge me to the best of my abilities. Managing this assignment, INCIDENT and the quizzes was a tough job to do.

I had to push through nights to absorb all of this massive information and complete all of the tasks.

The fun thing is that I loved every part of it. Day by day I just got more and more curious about machine learning. I couldn’t wait to hop on StatQuest and learn all the Machine Learning algorithms and concepts. But I couldn’t do them completely. I covered everything until Random Forest Classifiers and a few basics of XGBoost.

The models I learned are LogisticRegression, LinearRegression, KNN Classifier, Decision Trees, RandomForest Classifier.

I enjoyed learning the mathematics which goes behind these models and understanding them deeper, I got to know many many more concepts and terminologies used in machine learning.

I relied on StatQuest for understanding the models and the underlying concepts.

And for the implementation I would refer to GeeksForGeeks and Kaggle.

Overall, I absolutely loved investing my time in this assignment, getting infodumped from all the sources I could find and just getting to know more of the world of Machine Learning.

Evaluation Reports

For Task 1: Firstly I trained a Logistic regression model.

But it did not perform that well.

It had an accuracy of 18.75 percent.

Then I thought I should train other advanced models if they can make better predictions. But I obtained around same accuracies for other models too :

1)KNN Classifier: 23.125%

2)Decision Tree: 20%

3)Random Forest Classifier: 28%

Random Forest Classifier captured the relationship the best, but still it performed poor on the dataset.

Implying that the feature variables(stats) does not much ties with the type of pokemon.

For Task 2:   
I got an MSE of 78.18, which is also poor and suggests that the model is struggling to capture the relationship.

For Task 3:

I trained several models in order to predict Legendary status of pokemons and all of the models performed quite well.

Random Forest Classifier performed the best with an accuracy of 94.58% and an F1 Score of 0.682

Then comes KNN Classifier with an accuracy score of 94.1% and an F1 score of 0.533

Then is Logistic regression with an accuracy of 93.33% and an F1 score of 0.428

I infer that an F1 score of 0.682 for random forest is okayish (not so bad) considering the high imbalance of the data.

INSIGHTS

By plotting CAS Histogram of Legendary vs Non-Legendary pokemons, we found out that many non legendary pokemons had stats close to those of legendary pokemons.

These pokemons may perform like Legendary pokemons in battles, even though they are not classified so.

Another insight gained from these models is that there was a huge data imbalance in the dataset.

i.e a very small portion of the dataset was legendary making it difficult for the models to understand and predict them.

And CAS was indeed a good metric to predict whether a pokemon is legendary or not, but it too fails at capturing some relations.