PREDICTION OF HIV-1 PROGRESSION USING K NEAREST NEIGHBOR MACHINE LEARNING ALGORITHM

*Mohammad Nazmul Hasan Maziz[1], Mohammad Abdur Rashid[2], Farzana Yasmin[3]

[1]Faculty of Medicine, Bioscience and Nursing, MAHSA University, Bandar Saujana Putra, Jenjarom, 42610 Selangor, Malaysia

[2]Faculty of Medicine, AIMST University, Bedong, Kedah, Malaysia

[3]Faculty of Science, Lincoln University, 47301 Petaling Jaya, Selangor, Malaysia.

ABSTRACT

In the health context, the method of a teaching machine was used to identify possible candidates for preliminary prevention (preliminary preliminary). The technology of artificial intelligence has greatly affected the progress of the medical area. In the ConnectionOfConcept phase, the use of other types contains machine learning with Smart Componregated and Social Media, and guarantees real-time risk to ensure the disclosure of HIV Sero, as well as virtual reality technology, as well as HIV chatbots. Studies on HIV ML's prevention are also used in duke production. progress. HIV can be predicted using algorithms and machine learning methodologies.

KEYWORDS: MACHINE LEARNING, POPULATION ,ALGORITHMS, PREVENTION, HEALTHCARE

1 INTRODUCTION

The amount of “social big data” generated by technologies such as social media, wearable devices, and online searches is growing and may benefit HIV research.Researchers can identify patterns and understandings of HIV infection and trends, but the identification process takes time and resources.Machine learning strategies based on laptop technology could be used to help HIV professionals find unexpected and reliable ways to detect HIV-related styles in vast amounts of social data.[1]

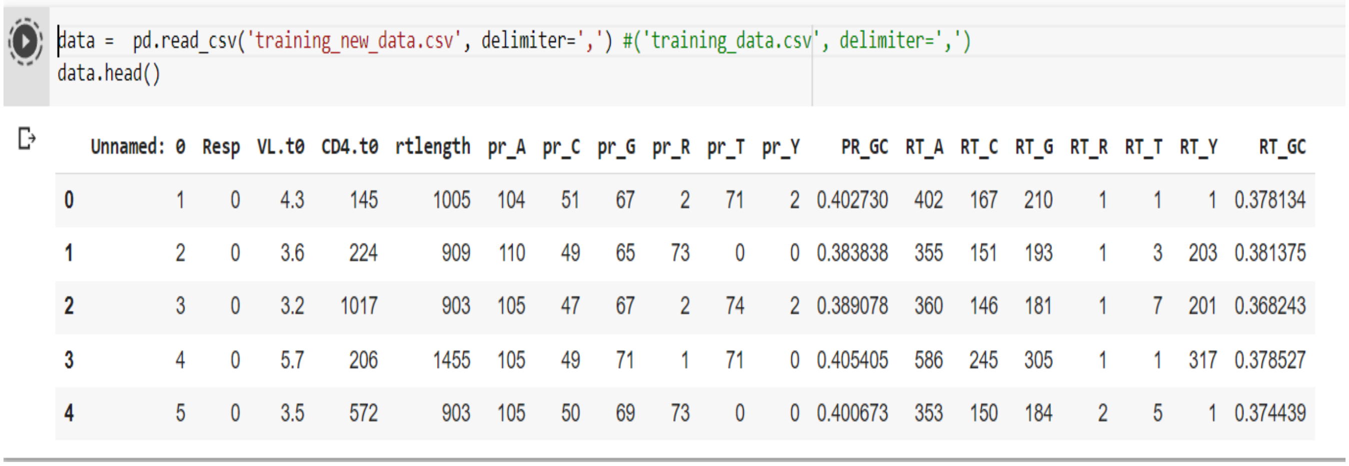
This competition focuses on forecasting the patient's short-term progression utilizing Reverse Transcriptase (RT) and Protease nucleotide sequences (PR). Nucleotide sequences are the blueprint for proteins that are the driving force of cells for non-biologists. The RT enzyme is responsible for the replication of the HIV1 genome within the cell. The HIV1 genome is translated into long chains of amino acids, and the PR protein is cleaved into the various functional components required for the HIV life cycle. Because they are predominantly exclusive to the HIV-1 life cycle, most HIV-1 medicines target these proteins.

I've included the two most common clinical signs for determining an HIV-1 infected person's "general health": viral load and CD4+ cell counts, in addition to the HIV-1 viral sequences. The CD4+ cell count is the number of white blood cells per milliliter of blood, whereas the viral load is the number of viral particles per milliliter of blood. The viral load is measured on a log-10 scale in this dataset. When the number is higher, the immune system becomes more "active." Faster CD4 levels, on the other hand, indicate a healthier individual as well as a higher viral reproduction rate (the virus primarily replicates in CD4 cells).[2]

2 HIV PROGRESSION PREDICTION USING THE DATA

The dataset from kaggle is taken to predict the HIV progression. The features of the data are VL.t0','CD4.t0', 'rtlength', 'pr_A', 'pr_C','pr_G', 'pr_R', 'pr_T','pr_Y','PR_GC','RT_A', 'RT_C',
'RT_G', 'RT_R', 'RT_T', 'RT_Y', 'RT_GC'. These features are dependent on the result of HIV .

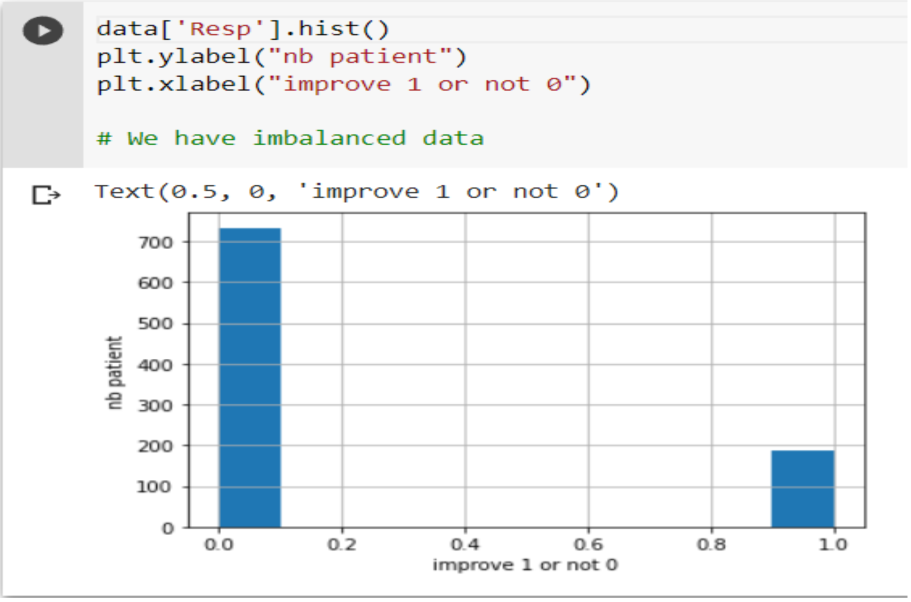
doi.org/10.31838/hiv21.01.13 Received: 10.07.21, Revised: 22.01.22, Accepted: 01.03.22



```
data =  pd.read_csv('training_new_data.csv', delimiter=',') #('training_data.csv', delimiter=',')
data.head()
```

	Unnamed: 0	Resp	VL.t0	CD4.t0	rtlength	pr_A	pr_C	pr_G	pr_R	pr_T	pr_Y	PR_GC	RT_A	RT_C	RT_G	RT_R	RT_T	RT_Y	RT_GC
0	1	0	4.3	145	1005	104	51	67	2	71	2	0.402730	402	167	210	1	1	1	0.378134
1	2	0	3.6	224	909	110	49	65	73	0	0	0.383838	355	151	193	1	3	203	0.381375
2	3	0	3.2	1017	903	105	47	67	2	74	2	0.389078	360	146	181	1	7	201	0.368243
3	4	0	5.7	206	1455	105	49	71	1	71	0	0.405405	586	245	305	1	1	317	0.378527
4	5	0	3.5	572	903	105	50	69	73	0	0	0.400673	353	150	184	2	5	1	0.374439

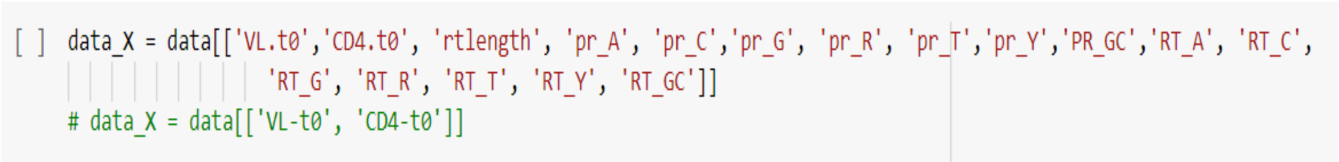
The dataset contains 920 instances, with 187 patients infected with HIV and 733 non-infected people.



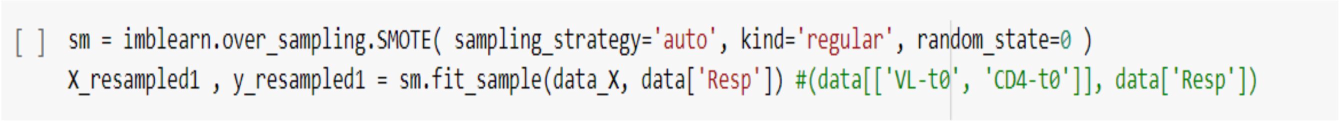
```
data['Resp'].hist()
plt.ylabel("nb patient")
plt.xlabel("improve 1 or not 0")

# We have imbalanced data
```

```
Text(0.5, 0, 'improve 1 or not 0')
```

The dataset's features columns are displayed, and they are referred to as data X and the y label as the result. SMOTE preprocesses the data before using the resampled data to train the model and forecast the outcome.



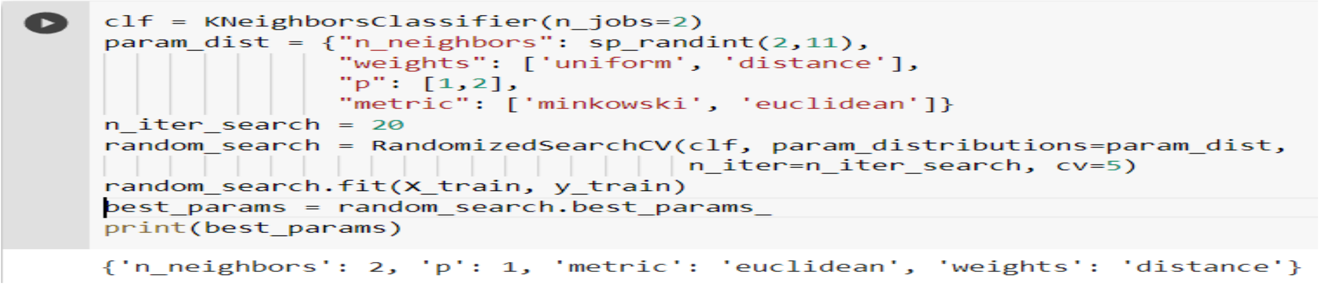
```
[ ]  data_X = data[['VL.t0','CD4.t0', 'rtlength', 'pr_A', 'pr_C','pr_G', 'pr_R', 'pr_T','pr_Y','PR_GC','RT_A', 'RT_C',
                    'RT_G', 'RT_R', 'RT_T', 'RT_Y', 'RT_GC']]
     # data_X = data[['VL-t0', 'CD4-t0']]
```



```
[ ]  sm = imblearn.over_sampling.SMOTE( sampling_strategy='auto', kind='regular', random_state=0 )
     X_resampled1 , y_resampled1 = sm.fit_sample(data_X, data['Resp']) #(data[['VL-t0', 'CD4-t0']], data['Resp'])
```

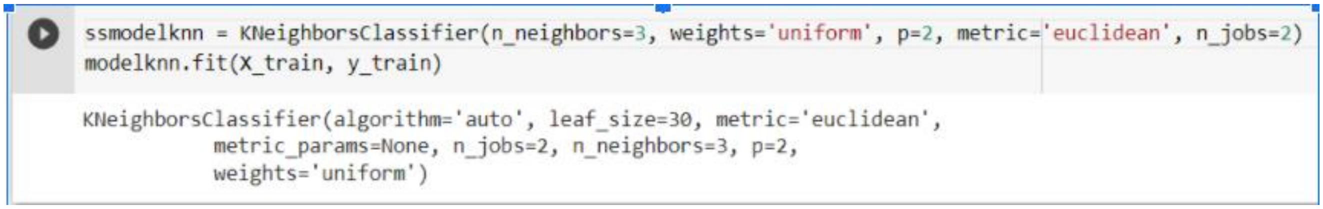
2.1 Using the K-Nearest Neighbor Model for predicting the HIV progression .

The K-Nearest Neighbor classifier is used, and the parameters affecting this model are n neighbors, weights, metrics, and the number of iterations is 20. The percentage of probability is calculated by searching the data using cross validation. It is 5 fold cross validation because the cross validation for random search is set to 5.



```
clf = KNeighborsClassifier(n_jobs=2)
param_dist = {"n_neighbors": sp_randint(2,11),
              "weights": ['uniform', 'distance'],
              "p": [1,2],
              "metric": ['minkowski', 'euclidean']}
n_iter_search = 20
random_search = RandomizedSearchCV(clf, param_distributions=param_dist,
                                   n_iter=n_iter_search, cv=5)
random_search.fit(X_train, y_train)
best_params = random_search.best_params_
print(best_params)

{'n_neighbors': 2, 'p': 1, 'metric': 'euclidean', 'weights': 'distance'}
```

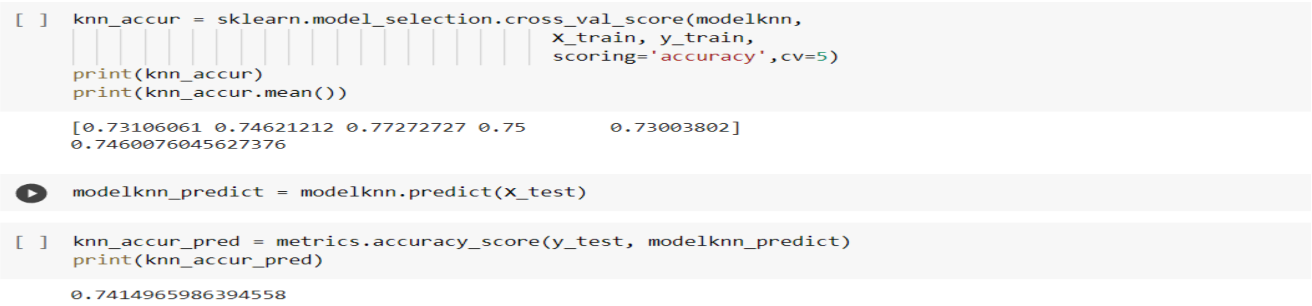
The training data is fitted in the K-Nearest Neighbor model, with euclidean metrics, three neighbors (nearest neighbors), and uniform weights.



```
ssmodelknn = KNeighborsClassifier(n_neighbors=3, weights='uniform', p=2, metric='euclidean', n_jobs=2)
modelknn.fit(X_train, y_train)

KNeighborsClassifier(algorithm='auto', leaf_size=30, metric='euclidean',
           metric_params=None, n_jobs=2, n_neighbors=3, p=2,
           weights='uniform')
```

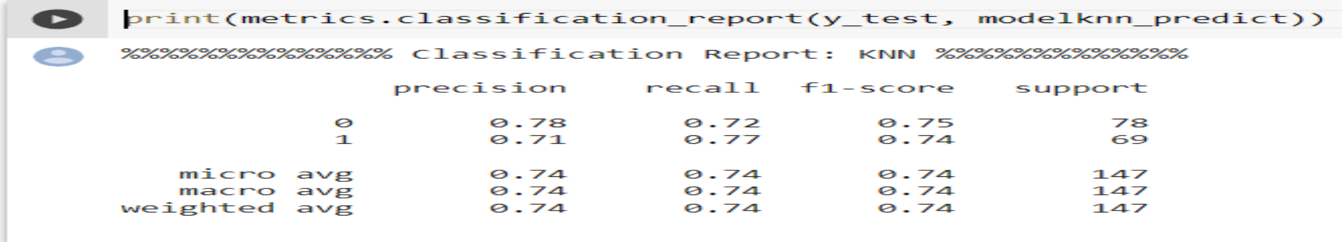
3 RESULTS AND DISCUSSIONS

The mean accuracy is required for the accuracy score obtained after fitting the training data in the k-nearest neighbour classifier. And the average precision was 74.6 percent. The accuracy score obtained for the test data in the classifier was 74.14 percent.



```
[ ] knn_accur = sklearn.model_selection.cross_val_score(modelknn,
                                                        X_train, y_train,
                                                        scoring='accuracy',cv=5)
    print(knn_accur)
    print(knn_accur.mean())

    [0.73106061 0.74621212 0.77272727 0.75       0.73003802]
    0.7460076045627376

    modelknn_predict = modelknn.predict(X_test)

[ ] knn_accur_pred = metrics.accuracy_score(y_test, modelknn_predict)
    print(knn_accur_pred)

    0.7414965986394558
```

The f1-score, precision, and recall are calculated by looking at the categorization report. The model's performance can be determined by knowing this value. The model's performance is good based on these numbers, and accuracy can be improved by constructing some classifiers and using neural networks.



```
print(metrics.classification_report(y_test, modelknn_predict))

%%%%%%%%%%%%%%%%% Classification Report: KNN %%%%%%%%%%%%%%%

              precision    recall  f1-score   support

           0       0.78      0.72      0.75        78
           1       0.71      0.77      0.74        69

   micro avg       0.74      0.74      0.74       147
   macro avg       0.74      0.74      0.74       147
weighted avg       0.74      0.74      0.74       147
```

CONCLUSION

Artificial intelligence advancements lead to automation in a variety of industries. Based on the findings, it can be inferred that by upgrading the model otr through the use of deep neural networks, the accuracy of the model can be raised, and these models can be used to forecast HIV progression.

REFERENCES

1. Young SD, Yu W, Wang W. Toward automating HIV identification: machine learning for rapid identification of HIV-related social media data. Journal of acquired immune deficiency syndromes (1999). 2017;74(Suppl 2):S128. https://doi.org/10.1097/QAI.0000000000001240.

2. Ironson GH. Do positive psychosocial factors predict disease progression in HIV-1? A review of the evidence. Psychosomatic medicine. 2008;70(5):546. https://doi.org/10.1097/PSY.0b013e318177216c.