**Company- Happymonk Technology Pvt. Ltd (Bangalore, India)**

**Candidate Name**- Mohammad Juned Khan

**Subject**- Test Assignment for the Junior Data Scientist Position

**Code link-** [**https://github.com/Mohammad-juned-khan/Cancer\_prediction\_using\_ANN/tree/main**](https://github.com/Mohammad-juned-khan/Cancer_prediction_using_ANN/tree/main)

**Objective –** Task is to build an Artificial Neural network using Wisconsin Breast Cancer dataset to correctly predict the class of tumor. Use different parameters for model such as activation function, epoch and others, compare the models using loss values, F1-Score, classification report and other measures.

**Abstract**- Wisconsin Breast Cancer dataset is the most commonly used data set for research and experiment. The data set is used to build an Artificial Neural Network (ANN) with one hidden layer. The ANN model will help prediction of diagnosis (M = malignant, B = benign) using unknown data/ information from patient report.

**The project includes following process: -**

1. **Download** Wisconsin Breast Cancer data set.
2. **Preprocess** the data set.
3. **Feature Engineering.**
4. **Model Building and Evaluation**

**Step 1. Download Data set: -** Download the data set from Kaggle or UCI Machine Learning Repository.

**Step 2. Preprocess data**: - Preprocessing is the most important step in any machine learning project. The data we have consist of 569 records with 32 columns/features. Most of the data is numeric and one categorical feature which is the target class.

Following are the operations performed in preprocessing.

* Data set consist of one extra column “Unnamed: 32” with 569 NaN values which is removed.
* ID column is set as index.

**Step 3. Feature Engineering**: - Feature engineering is the process of preparing datasets and features in a machine compatible format.

Following are the operations performed in Feature Engineering.

* **Encoding Categorical Feature**- Encoded the class label using LabelEncoder.
* **Feature Scaling**- Scaled numerical features using standard scaler from Sklearn .
* **Split data**- Split data into train and test data set.
* **Feature Selection**- Feature selection is an important step to identify features that are more helpful in predicting the class of diagnosis. For feature selection we have used Embedded technique by using RandomForestClassifier to get important features.

**Lastly create data set with top important features**. We have more than 30 features of which top 12 features are selected to create a training data set for our model.

**Step 4**. **Model Building: -** For building **ANN** we have used Google Colab as web IDE for python and machine learning. And multiple libraries to create model such as Keras.

Keras to create Sequential model. And Dense to add hidden layer.

The model consists of input layer, one hidden layer and an output layer. These layers consist of parameters such as for classification problem we have used

Input layer – It consist of 12 input and we have used “relu” as activation function with “he\_uniform” as weight initialization. This is because relu works very well with he\_uniform.

Hidden layer- It consist of “relu” as activation function with “he\_uniform” as weight initialization and different number of neurons in different models.

Output laye- The output layer consists of only a single neuron. With activation function and “sigmoid” and weight initialization as “glorot\_uniform”. The combination of sigmoid and glorot\_uniform is good for binary classification.

To compile the model, we have used multiple optimizers such as Adamax, and Adam along with the loss function “binary\_crossentropy” for binary class classification and metrics as “accuracy”.

Also, we have tried different combination of parameters to train/fit the model for better performance and lower value of loss as much as possible. With different values of “epochs”, validation split and batch size we can fit the model.

Model Evaluation: - To evaluate the accuracy of the classification model we have F1 score, Classification Report, Confusion Matrix and others.

In the Code file we have two plots for each of the model that shows: -

* Training Loss Vs Validation Loss
* Training Accuracy Vs Validation Accuracy

For better understanding of the problem and search for optimal performance we can also use Hyperparameter Optimization to find best parameters.

We can also use the Keras Tuner for best optimal parameters to het higher accuracy.

We tried to build some model with different combination of parameters and measured their accuracy for search of optimal model.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Performance Record | | | | | | | | | | | | | | |
|
| SI.NO | Model Name | Neurons Layers | | | Optimizer | Activation in last layer | Epochs | Training Loss | Val Loss | Training Accuracy | Val Accuracy | Test | | F1 Score |
| Input | Hidden | Output | Loss | Accuracy |
| 1 | Model1 | 5 | 5 | 1 | Adamax | sigmoid | 100 | 0.0901 | 0.1383 | 0.9606 | 0.978 | 0.2537 | 0.9474 | 0.95 |
| 2 | Model2 | 5 | 5 | 1 | Adam | sigmoid | 150 | 0.0568 | 0.1647 | 0.9839 | 0.978 | 0.317 | 0.9474 | 0.95 |
| 3 | Model3 | 8 | 4 | 1 | Adamax | sigmoid | 150 | 0.1093 | 0.1315 | 0.9445 | 0.967 | 0.2027 | 0.9474 | 0.95 |
| 4 | Model4 | 6 | 3 | 1 | Adamax | sigmoid | 200 | 0.101 | 0.1265 | 0.9433 | 0.956 | 0.1987 | 0.9474 | 0.95 |
| 5 | Model5 | 5 | 5 | 1 | Adamax | Tanh | 200 | 0.2836 | 0.313 | 0.9139 | 0.9451 | 0.5462 | 0.9298 | 0.93 |

**Code link-** [**https://github.com/Mohammad-juned-khan/Cancer\_prediction\_using\_ANN/tree/main**](https://github.com/Mohammad-juned-khan/Cancer_prediction_using_ANN/tree/main)

**The link contains-**

* Cancer\_ANN\_M\_Juned\_Khan\_\_All\_Model.ipynb
* Code Implementation Document “Implementation-Code-Document-Cancer-ANN.doc”
* Assignment-Mohammad.Juned Khan.pdf
* Dataset “Cancer.csv”
* Performance Table.Xlsx

**Submitted by – Mohammad Juned khan**

**Mail-** [**Mohammad.juned.z.khan@gmail.com**](mailto:Mohammad.juned.z.khan@gmail.com)

**Contact- +91 9834454754**

**LinkedIn-** [**https://www.linkedin.com/in/md-juned-khan/**](https://www.linkedin.com/in/md-juned-khan/)