**Introduction to Data Mining and Machine Learning Report by Kelvin Kipkemboi, SID- 2161812.**

**Link to Labwork ipynb files is:**

[**https://github.com/Kevcoding/Data-Mining-Labworks**](https://github.com/Kevcoding/Data-Mining-Labworks)**:**

**Week 1 Lab work report**

Our course tutor: Dr. Imran Ahmed introduced us to the second trimester’s workload by outlining the course materials, resources, and additional websites to be used. These included *Kaggle.com, Github.com, and Medium.com*. I created new accounts on the three platforms for the purpose of the trimester onwards.

We started Exploratory Data Analysis of the Titanic dataset.

**Week 2 Lab work report**

Introduction to ***data pre-processing*** using Data.csv

From what we did, these steps were undertaken:

1. Importing libraries
2. Import data
3. Dealing with missing values using dropna : fillna with NAN; and fillna with mean
4. Data encoding categorical variables using OneHotEncoder and LabelEncoder
5. Split the data into test and train sets using train\_test\_split
6. Data scaling using StandardScaler and MinMaxScaler
7. Model training and fitting
8. Model testing on the X\_test
9. Model Visualization of test sets
10. Model Evaluation using evaluation matrices such as MSE, RMSE, MAE, R2

**Week 3 Lab work report**

***Simple Linear Regression*** using [PracticeData.csv](https://canvas.anglia.ac.uk/courses/30984/modules/items/1686692). Under this regression model, we found that outcome –Y (Dependent/Response variable) is obtained from Independent/ explanatory variable X, under the equation Y0 = B0 + B1X1 + Ethis equation has:

Bo – as Y-intercept; B1 – As coefficient; X1 – The Independent variable; Y0 – The Dependent variable and E – This is the error term

After exploring the regression line, we studied the assumptions of the model:

1. Errors, E is normally distributed and
2. There are no clear outliers

We undertook Hypothesis Testing using H0(Null Hypothesis) and H1(Alternative Hypothesis), with rejection rules given the desired level of significance using the p-value.

In so doing, we either fail to reject the null hypothesis or accept the alternative hypothesis.

Finally, we found out that *regression and correlation analysis* was *used to study the association of continuously measured outcomes and determinants*, where:

*Correlation analysis*: Used to measure the strength and direction of the association between variables. The correlation of X and Y (Y and X), values of *r* falling in the range (–1<0<1) while

*Linear regression*: Used in predicting the value of one variable based on (given) the value of the other variable. The regression of Y on X.

**Week 4 Lab work report**

Finalizing Simple Linear Regression and diving into ***Multiple Linear Regression*** using [Churn\_Modelling2.csv](https://canvas.anglia.ac.uk/courses/30984/modules/items/1724714) and [Salary\_Data.csv](https://canvas.anglia.ac.uk/courses/30984/modules/items/1724715)

Under Multiple Linear Regression, the steps covered were:

1. Importing libraries,
2. Importing the dataset,
3. Encoding categorical data,
4. Splitting the dataset into the Training and Test set
5. Training the Multiple Linear Regression model on the Training set
6. Predicting the Test set results
7. Making a single prediction
8. Deriving linear equation with values of the variable coefficients

**Week 5 Lab work report**

Undertook ***Polynomial regression*** using [50\_Startups.csv](https://canvas.anglia.ac.uk/courses/30984/modules/items/1686715) and [Salary\_Data.csv](https://canvas.anglia.ac.uk/courses/30984/modules/items/1686717)

The regression line for the polynomial had its second and subsequent independent variables raised to powers greater than 1. As such, the regression line assumed the shape:

Y0 = B0 + B1X1 + B2X12+ B3X13 + E:

Bo – as Y-intercept

B1 – As the coefficient

X1 – The Independent variables

Y0 – The Dependent variable

E – This is the error term

The steps covered for this lab week were:

1. Importing libraries,
2. Importing the dataset,
3. Encoding categorical data,
4. Splitting the dataset into the Training and Test set
5. Training the Linear Regression model on the Training set with varying degrees of the coefficients.
6. Predicting the Test set results
7. Making a single prediction with linear regression and with polynomial regression
8. Evaluation of the model using evaluation matrices.

**Week 6 Lab work report**

Studied ***Logistic Regression*** using [DataModelSelection.csv](https://canvas.anglia.ac.uk/courses/30984/modules/items/1686728), [and Position\_Salaries.csv](https://canvas.anglia.ac.uk/courses/30984/modules/items/1686729). From there, we moved on to check the steps necessary to conduct ***Support Vector Regression***, ***Decision Trees***

Logistic regression is one of the Machine Learning algorithms for making classifications; It is based on modeling the probability of a random variable Y being an outcome of either a 1 or a 0.

We used a sigmoid math function fitted to a simple regression line to create a sigmoid curve for logistic regression. In so doing, we noticed that a threshold must be specified for a particular case. For the lecture, we picked on 0.5, such that values below the threshold value were assumed to be 0. Values above the threshold were taken to be 1.

**Week 7 Lab work report**

**KNN,** **SVM, and SOM:**

K-Nearest Neighbors Algorithm utilizes data items around them to define datapoint to be used as defining points for classification. This is based on the Euclidean distances of the data items. In the model, one picks odd numbers as the number of k to avoid ties during the voting process in the assigning of centroids. Locations can be updated.

Under Support Vector Machines we have the creation of a hyperplane that acts as a categorizer for data points that are not linearly separable. Thus creating a higher dimension and fitting the hyperplane. SVM can be used for regression, classification, and detecting outliers in datasets.

SVM, unlike most machine learning methods, not only compares data points to existing standards or structures but makes reference to extreme points (outliers) using them as support vectors.

Self-Organizing Maps, a deep learning method for clustering (unsupervised learning) is based on dimensionality reduction, in a way that instead of taking new features and creating new dimensions, the model re-creates the initial input maps and then self-organizes outputs into original maps created by self. This helps reveal correlations not easily visible.

**Week 8 Lab work report**

For ***Random Forests***, whose modeling is based on choosing n-data points randomly from the training set and building a decision tree based on the n-data points. To get a new data point, we use our n-trees to predict the value of Y for the new data point, then assign the average across all predicted Y-values to the new data point.

In modeling the random forest using RandomForestRegressor, we choose n estimators and state our random state, before we fit the model to our dataset.

**Week 9 Lab work report**

During this session, Hierarchical clustering: K-means clustering, and decision tree classification models were featured.

For K-means clustering, the steps covered were:

1. Data pre-processing
2. Finding the optimal number of clusters using the elbow method
3. Training the K-Means algorithm on the training data set
4. Visualizing the clusters

For the decision tree classification model;

1. Importing libraries,
2. Importing the dataset,
3. Splitting the dataset into the Training and Test set
4. Feature scaling using the standard scaler
5. Training the decision tree classifier model on the Training set
6. Predicting the Test set results
7. Evaluation of the model using evaluation matrices: confusion matrices and classification report
8. Visualization of train and test sets

Some uses of the classification models were:

#Sentiment Analysis.

#Email Spam Classification.

#Document Classification.

#Image Classification

**Week 10 Lab Report**

**ANN CNN and RNN:**

Artificial Neural Network used in pattern recognition, works by taking inputs, learning by more inputs, and learning by adjusting the weights of the hidden nodes to achieve the desired output. In the process, the neural network learns by backpropagation of errors so that the system learns to achieve higher accuracy.

For Convolutional Neural Networks, utilized in computer vision and works through layers below:

1. Convolution layer: Generation of feature maps
2. Max pooling: ensuring features are invariant of transformations
3. Flattening: vectorizing of inputs
4. Full connection: a running CNN

Additional filters create additional feature maps that create a convolution layer

Recurrent Neural Networks utilize LSTM (Long Short-Term Memory) and thus can process both past data and input data making precise predictions. Though slower to learn, once trained it is highly efficient in that it can take multiple inputs and map them to many outputs and one input to many or many inputs to one output.

**Week 11 Lab Report**

Recommendation systems using Boltzmann Machines:

1. Restricted Boltzmann Machine (RBM)
2. Deep Belief Network (DBN)
3. Deep Boltzmann Machine (DBM)
4. GANs
5. Autoencoders (Stacked and Deep)

For these recommender systems, we analyzed its architecture from Visible input nodes, hidden layers, visible output layers, and activation layers using SoftMax and reLu activators to generate a recommendation based on the output of the neural network

|  |  |
| --- | --- |
| **Data science model** | **Dataset** |
| Simple Linear Regression | **Salary Data** |
| Multiple Linear Regression | **50 Startups** |
| Polynomial Regression | **Position Salaries** |
| Random Forest Regression | **Position Salaries** |
| Decision Tree Regression | **Position Salaries** |
| SVR | **Position Salaries** |
| Logistic Regression | **Social Network Ads** |
| SVM | **Social Network Ads** |
| Decision Tree Classification | **Social Network Ads** |
| Self-Organizing Maps | **Credit Card Applications** |
| K-Means Clustering | **Mall Customers** |
| Kernel SVM | **Social Network Ads** |
| KNN | **Social Network Ads** |
| CNN | **sample\_submission.csv** |
| DBN Recommender Systems | **Movie Lens** |
| Tokyo Olympics Visualization | **Olympics** |
| Titanic EDA | **Titanic** |
| Diabetes EDA | **Diabetes** |
| Time Series Analysis | **Superstore.xls** |