Configuration Manual

MSc Research Project
Data Analytics

Bharat Bhardwaj
Student ID: X18186424

School of Computing
National College of Ireland

Supervisor: Dr Rashmi Gupta

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Bharat Bhardwaj		
Student ID:	X18186424		
Programme:	Master's in Data Analytics	**Year:**	2019-2020
Module:	Research Project		
Supervisor:	Dr Rashmi Gupta		
Submission Due Date:	25 September 2020		
Project Title:	Prediction of Charged-off Loans Using Classification Models and Artificial Neural Network for P2P Online Banking		
Word Count:	929	**Page Count**	8

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Bharat Bhardwaj

Date: 25 September 2020

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Configuration Manual

Bharat Bhardwaj
Student ID: X18186424

1 Hardware/Software Requirements

1.1 Hardware Requirements

Hardware details are as below to run the development smoothly.

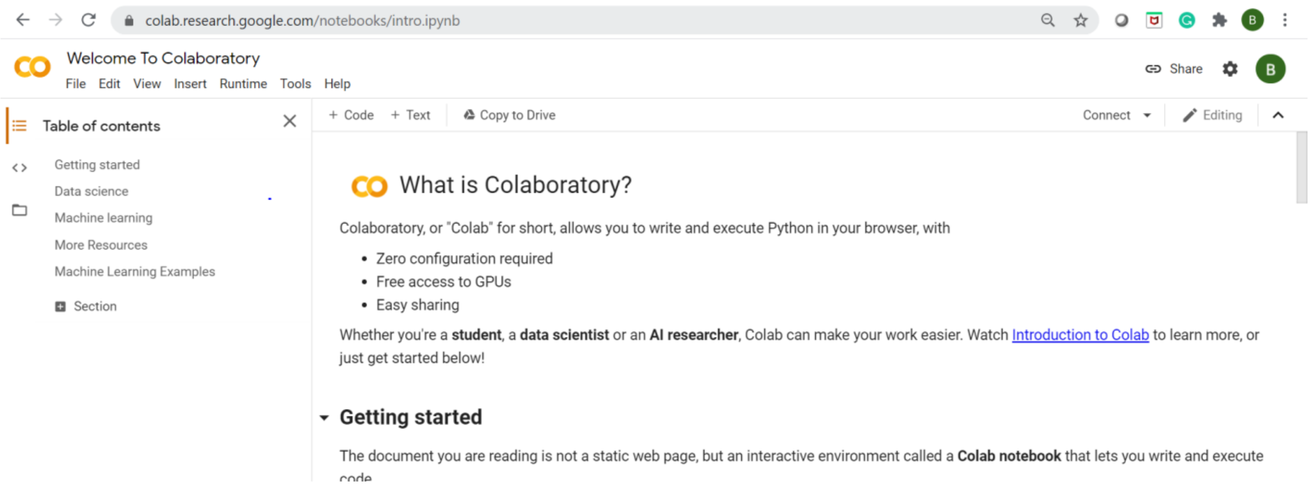
Operating System	Windows 10
RAM (Read only memory)	8GB
Hard Disk	50 GB

1.2 Software Requirements

Programming Language Tools	Google Collaboratory, Python version 3, Microsoft Excel
Web Browser	Google Chrome or Mozilla
Email	Gmail account

2 Google Collaboratory Environment Setup

This section explains the google colab environment set up to perform the development activity. Respective screenshot for set-up is given below. A Gmail account as bharat.bhardwaj2014@gmail.com has been used to create a account on google colab notebook.



What is Colaboratory?

Colaboratory, or "Colab" for short, allows you to write and execute Python in your browser, with

- Zero configuration required
- Free access to GPUs
- Easy sharing

Whether you're a **student**, a **data scientist** or an **AI researcher**, Colab can make your work easier. Watch Introduction to Colab to learn more, or just get started below!

Getting started

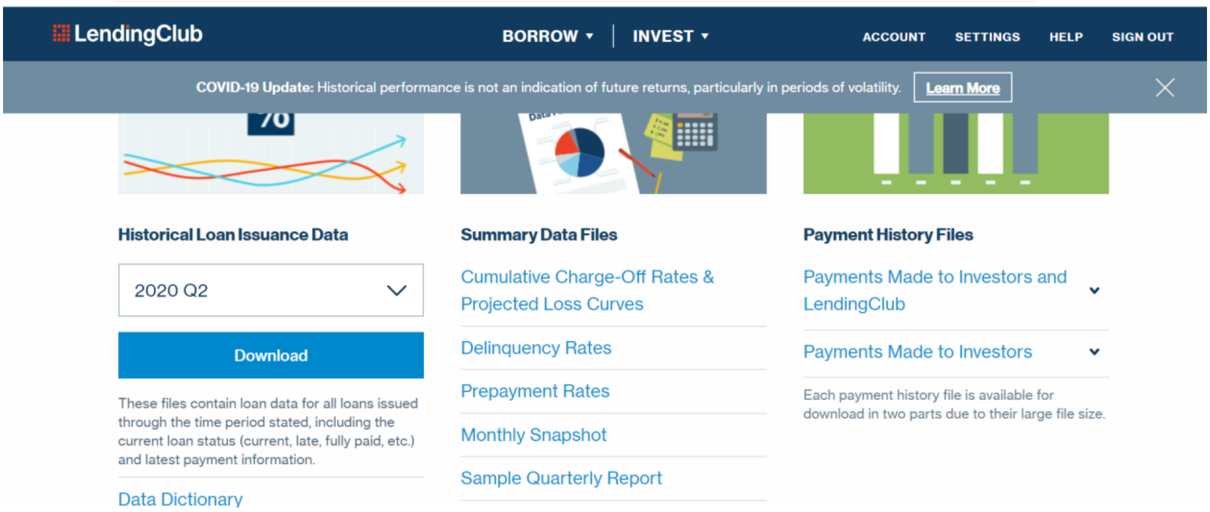
The document you are reading is not a static web page, but an interactive environment called a **Colab notebook** that lets you write and execute code

Figure 1: Sign-in to Google Collaboratory

3 Data preparation for Experiments

3.1 Data upload over google drive:

Lending club data collected from lending club official site and uploaded over google drive as per Figure 3.



COVID-19 Update: Historical performance is not an indication of future returns, particularly in periods of volatility.

These files contain loan data for all loans issued through the time period stated, including the current loan status (current, late, fully paid, etc.) and latest payment information.

Each payment history file is available for download in two parts due to their large file size.

Figure 2: Data collection from LendingClub site

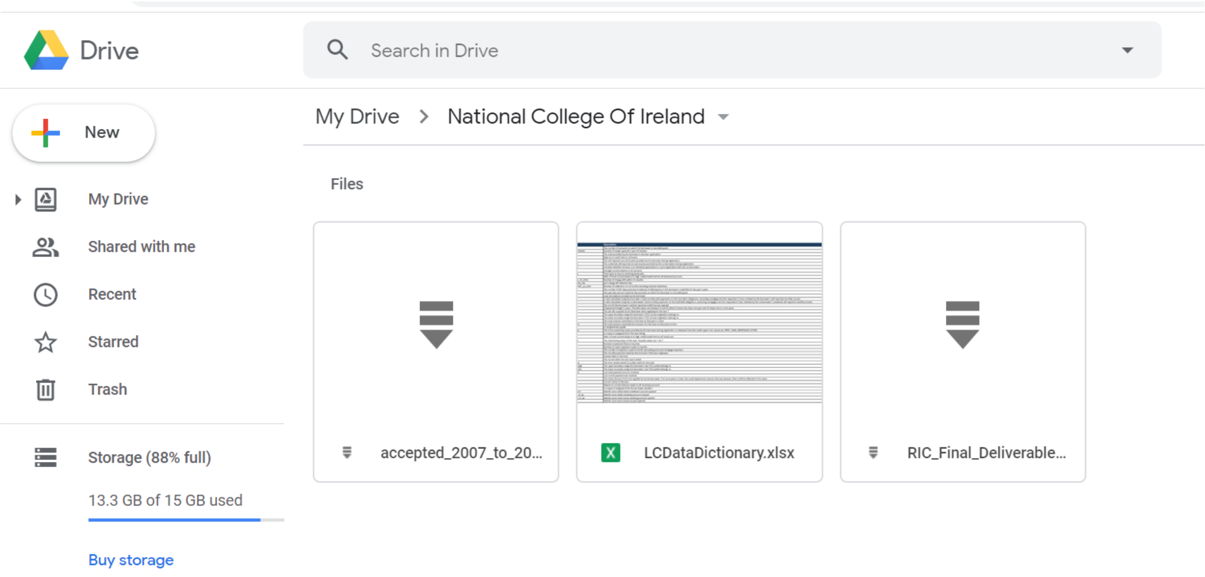


Figure 3: Data upload over google drive

3.2 Mount google drvie over google colab

Google drive is mounted over google colab platform. While mounting the drive over colab a gmail authentication is required. Once the authintication is done drive will be mounted over the colab and data can be assessed over the colab platform.

```
from google.colab import drive
drive.mount('/content/drive')
```

3.3 Unzip LendingClub data

The mounted data is in zip file over google colab hence to access the data over colab platform data need to unzip as below.

```
data = pd.read_csv('/content/drive/My Drive/National College Of Ireland/accepted_2007_to_2018Q4.csv.gz', compression='gzip', low_memory=True)
```

3.4 Import python librariese

To progress the devlopment acitivity further, python file need to import on google colab as shown below.

```
import numpy as np
import scipy as sp
import pandas as pd
import matplotlib as mpl
import matplotlib.pyplot as plt
import seaborn as sns
import tensorflow as tf
from tensorflow import keras
from sklearn.model_selection import train_test_split
from keras import backend as K
from sklearn.metrics import accuracy_score
from sklearn.metrics import confusion_matrix
from sklearn.metrics import classification_report
from sklearn.metrics import precision_score
from sklearn.metrics import recall_score
from sklearn.metrics import f1_score
from sklearn.metrics import matthews_corrcoef
from keras import optimizers
from keras.models import Sequential
from keras.layers.core import Dense, Dropout, Activation
from keras.callbacks import EarlyStopping
from matplotlib import pyplot
from sklearn.feature_selection import f_classif
from sklearn.pipeline import Pipeline
from sklearn.preprocessing import StandardScaler
from sklearn.impute import SimpleImputer
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis
from sklearn.model_selection import GridSearchCV
from sklearn.linear_model import SGDClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn.ensemble import RandomForestClassifier
# Pandas options
pd.set_option('display.max_colwidth', 1000, 'display.max_rows', None, 'display.max_columns', None)
# Plotting options
%matplotlib inline
mpl.style.use('ggplot')
sns.set(style='whitegrid')
```

4 Model execution

Keras and TensorFlow libraries are used to execute the artificial neural network. Insights of data has been presented using the matplot library.

4.1 Libraries for Artificial neural network (ANN) model

As shown above (section 3.4), Keras and TensorFlow libraries are used to run the ANN model.

4.2 Libraries for classification model run

As shown above (section 3.4), Tensorflow and keras librariese are used to run the logistic regression, k-nearest neighbour and random forest classifer models. To present the insights of data matplot lib are used. Sklearn libraries are used to evaluate the models metrices.

5 Settings done for accelerating Computation time

This section will explain about how the drive storage is used and GPU setting is made from the google Collaboratory.

5.1 Drive Storage

Drive storage of 36 GB is used to store lending club data. Drive storage takes less time in data uploading over the google colab. Figure 4 shows the utilization of google drive for this project.

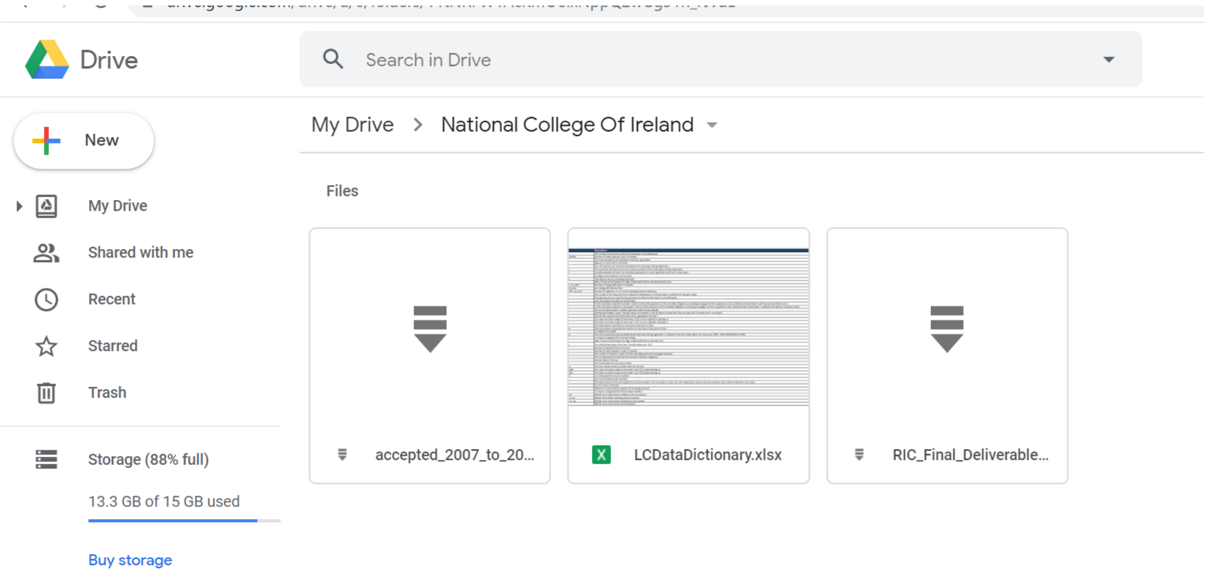
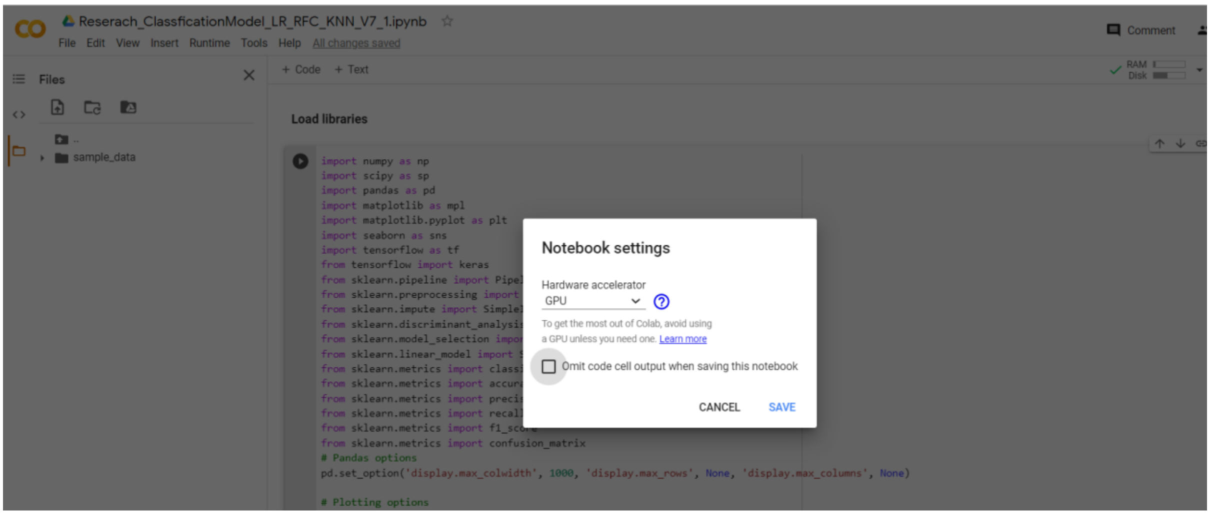


Figure 4: drive storage for the research study

5.2 GPU

To execute code faster and less time consumption GPU was used as run time envoirnment for the study.



Reserach_ClassficationModel_LR_RFC_KNN_V7_1.ipynb
File Edit View Insert Runtime Tools Help All changes saved
Comment
Files
sample_data
+ Code + Text
RAM
Disk

Load libraries

```
import numpy as np
import scipy as sp
import pandas as pd
import matplotlib as mpl
import matplotlib.pyplot as plt
import seaborn as sns
import tensorflow as tf
from tensorflow import keras
from sklearn.pipeline import Pipel
from sklearn.preprocessing import
from sklearn.impute import Simple
from sklearn.discriminant_analysis
from sklearn.model_selection impor
from sklearn.linear_model import S
from sklearn.metrics import classi
from sklearn.metrics import accura
from sklearn.metrics import precis
from sklearn.metrics import recall
from sklearn.metrics import f1_score
from sklearn.metrics import confusion_matrix
# Pandas options
pd.set_option('display.max_colwidth', 1000, 'display.max_rows', None, 'display.max_columns', None)

# Plotting options
```

Notebook settings

Hardware accelerator: GPU

To get the most out of Colab, avoid using a GPU unless you need one. Learn more

□ Omit code cell output when saving this notebook

Figure 5: GPU is used as run time envoirnment for the study

6 Other Software used

Draw.io online tool is used to design the project KDD process and research architecture design. Figure 6 and Figure 7 explains the related design documentations.

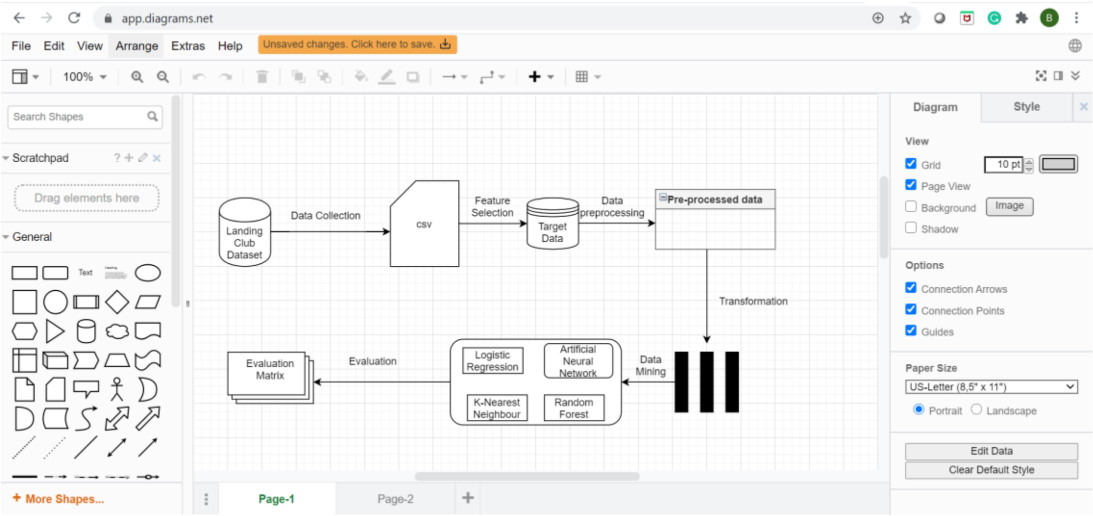
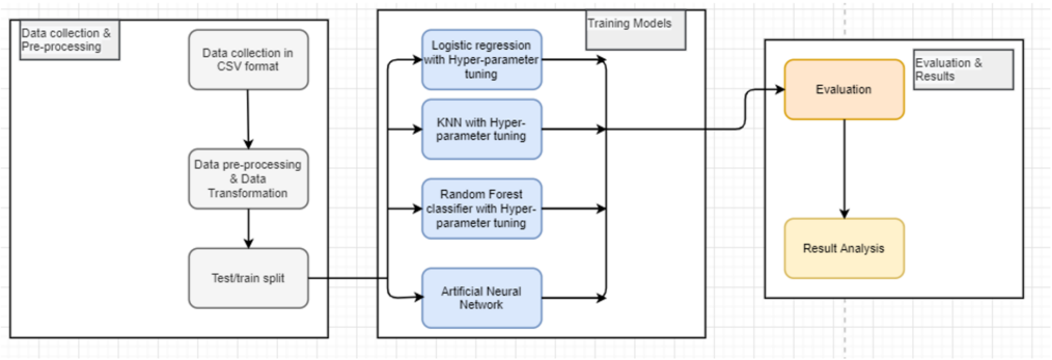


Figure 6: Architecture desgin



Data collection & Pre-processing
Data collection in CSV format
Data pre-processing & Data Transformation
Test/train split
Training Models
Logistic regression with Hyper-parameter tuning
KNN with Hyper-parameter tuning
Random Forest classifier with Hyper-parameter tuning
Artificial Neural Network
Evaluation & Results
Evaluation
Result Analysis

Figure 7: KDD process

References

https://keras.io/guides/sequential_model/

https://matplotlib.org/

https://scikit-learn.org/stable/

https://www.lendingclub.com/info/demand-and-credit-profile.action?source=post_page

https://colab.research.google.com/notebooks/intro.ipynb

https://app.diagrams.net/