CERTIFICADO PROFESIONAL EN DATA SCIENCE
Final Project Report

Student Name: Jon P. Hernández Nevárez

Dataset: Check Credit
File names: "*check_credit.txt*" → Data
"*Credit check database - information.txt*" → Data Description

Phase 1 – Data understanding

1. Variable association:

Variable description	Variable name
1. Age in years (numeric)	Age
2. Sex (text: male, female)	Sex
3. Job (numeric: 0 - unskilled and non-resident, 1 - unskilled and resident, 2 - skilled, 3 - highly skilled)	Job
4. Housing (text: own, rent, or free)	Housing
5. Saving accounts (text - little, moderate, quite rich, rich)	Saving accounts
6. Checking account (numeric, in DM - Deutsch Mark)	Checking account
7. Credit amount (numeric, in DM)	Credit_amount
8. Duration (numeric, in month)	Duration
9. Purpose (text: car, furniture/equipment, radio/TV, domestic appliances, repairs, education, business, vacation/others)	Purpose
10. Class or outcome (how's the credit)	class

2. The number of instances or samples are 1,000.
3. The file is using ';' as a column separator.
4. The file has a few '?' characters that represent missing values.

Phase 2 – Data Wrangling
The program in Python to clean my data is below:

```
f_input = open("check_credit.txt",'r')
f_output = open("check_credit.txt-clean-data.csv", 'w')
for line in f_input:
        if '?' not in line:
                new_output_line = line.replace(";" , ",")
        f_output.write(new_output_line)
f_input.close()
f_output.close()
print("End processing")
```

Phase 3 – Model Planning

1. Compared ML Algorithms:

a) Logistic Regression (functions.Logi)
b) Naïve Bayes (bayes.Naive)
c) Random Forest (tress. Rando)
d) K-Nearest Neighbor (lazy.IKb)
e) Support Vector Machine (functions.SMO)

2. Table with the results from the comparison:

Dataset	(2) functions.Logi |	(1) bayes.Naive	(3) lazy.IBk '-	(4) functions.S	(5) trees.Rando
check_credit.txt-clean-da(100)	73.58(3.60) |	71.70(3.06) *	66.98(3.40) *	72.85(3.29) *	73.23(3.23)
	(v/ /*) |	(0/0/1)	(0/0/1)	(0/0/1)	(0/1/0)

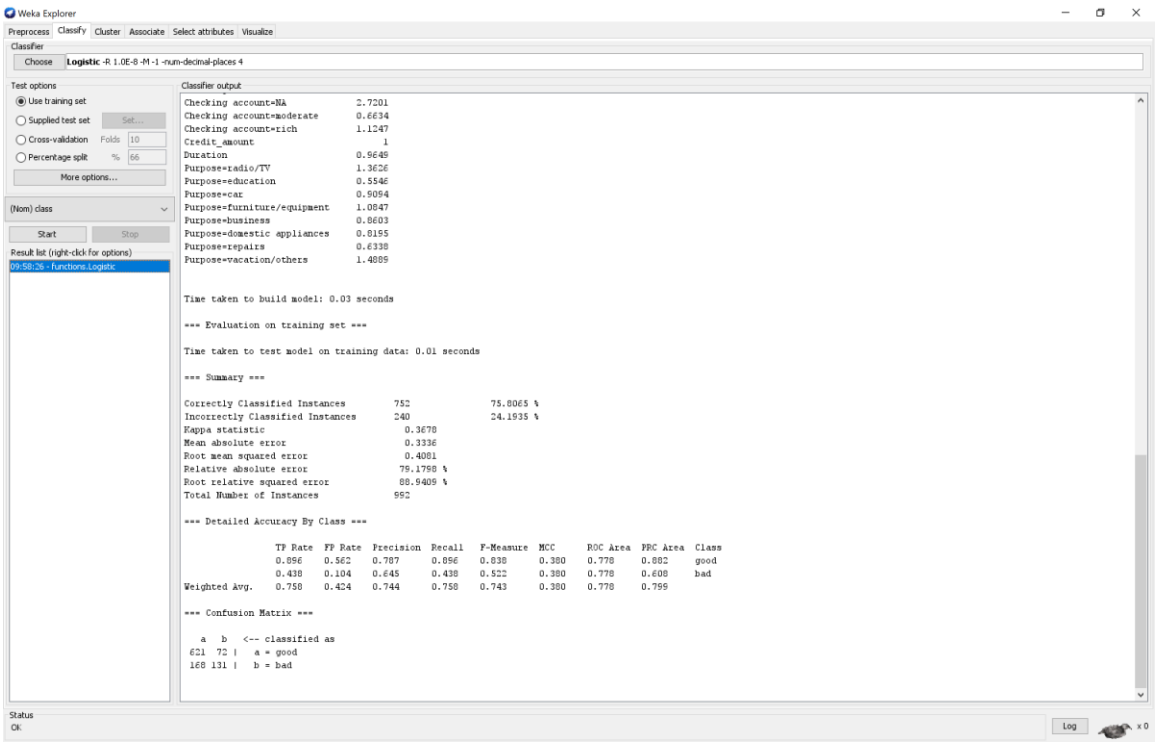
3. Discussion:

The best algorithm to build the model is Logistic Regression, because it has the best performance (i.e. greatest accuracy result).

Phase 4 – Model Building

1. Best Algorithm: Logistic Regression Model for Check Credit Dataset

2. Evaluation on training set:



Phase 5 – Communication Results

1. Metrics of the Model Evaluation (Taken from Model Building)

Metrics	Value
Precision	0.744
Recall	0.758
F-Measure	0.743
ROC Area	0.778

2. Final Statement (taken from the Model Planning)

Based on the accuracy of the compared algorithms, we used the Logistic Regression as the most accurate model (i.e. Accuracy: 73.58 with a Standard deviation of ±3.60). Then it can be stated that this model will make predictions with an accuracy of 73.58% (± 2 x 3.60 = ±7.2%).

In other words, the model will predict if the person requesting a loan has a good credit status with an accuracy between 66.38 % and 80.78%.