Faculty of Information Technology

Final Exam, 2021/2022

Course (Code) Name: **INSY 216 INTRO TO BIG DATA ANALYTIC**

Instructor: **Ishimwe M. Prince** Exam Duration: 2hrs

Group: **All**  Date: …th/06/2022

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

INSTRUCTIONS:

* **Answer all questions**
* **Put all your answers in the booklet**
* **Exam is closed books**

Music is a cultural right, it is used in the promotion and protection of other human rights. It can help in the healing process, dismantling walls and boundaries, reconciliation, and education. Around the world, music is used as a vehicle for social change and bringing communities together. From this background, the Rwandan Ministry of youth and culture conducted a survey in the city of Kigali, aiming to find out the most loved artist by district, based on f the respondent gender, age, and the district where he live. The collected information will be used by different stakeholders before their awareness campaign, to find out who is the best song artist fit for the audience they want to reach in the City of Kigali.

After data collection, a dataset was created for data analysis and modeling. As a data scientist, you are hired to do the following:

**I. To clean the dataset (10 Marks)**

1. By replacing the empty cells with the **median where** the axis = 0. **(5 Marks)**
2. By dealing with the duplicate : **(5 Marks)**
   1. Primarily the duplicate must be highlighted
   2. Then deleted them

**II. To Create a model** **(30 Marks)**

1. Find the training and testing accuracy percentage based on the Random forest, Support Vector Machine, Logistic Regression and Decision Tree Algorithms. **(10 Marks)**
2. Create a persisting model with the algorithm that give the best learning accuracy. **(10 Marks)**
3. Create the **UI** for the persisting model **(10 Marks)**

***God Bless you!***