DFFML: Adding New Machine Learning Models

About Me		
Name	*Yash Lamba*	
	GitHub	*@yashlamba*
	Gitter	*@yashlamba*
University	*Cluster Innovation Centre, University of Delhi*	
	Program	*B.Tech. (Information Technology & Mathematical Innovations)*
	Year	*I*
	Expected Graduation	*2022*
Contact	**Mail**	
	Phone	
	LinkedIn	*https://www.linkedin.com/in/yl1/*
Time Zone	*GMT +5:30 (IST)*	

Code Contributions

Since I finalized DFFML as the organisation I would like to work for this summer, I have been working on a project that was an idea for GSoC'19 and unexpectedly ended up completing most of it. The details are as follows:

-> Issue (Project Description): https://github.com/intel/dffml/issues/15
The task was to add support for various compression modules in DFFML.

-> Pull Requests (Related to Project [#15](https://github.com/intel/dffml/issues/15)):

1. Added Support for Gzip module (MERGED):
https://github.com/intel/dffml/pull/30
It was my first PR for the project, my mentor guided me really well on how to improve and simultaneously write tests for my contribution.

2. Added Support for bz2 module (MERGED):
https://github.com/intel/dffml/pull/32
Completed with the bz2 module including support and testing

3. Added Support for lzma and xz modules (MERGED):
https://github.com/intel/dffml/pull/37
Completed with the lzma module including support and testing. Adding support for xz wasn't planned but as lzma module supports xz files, added that too.

4. Documentation for the Source module (AWAITING CHANGES):
https://github.com/intel/dffml/pull/40
As the project was completed, I offered completing the documentation of the project, got the necessary guidance from my mentor and completed the same.

-> Pull Requests (Out of the project):

1. Fixed typos and grammatical errors in documentation (MERGED):
https://github.com/intel/dffml/pull/26
This was my first PR to DFFML, fixed the errors while understanding the application and code.

Project Information

1. Sub-Organisation: DFFML

2. Project Abstract:

Adding new machine learning models to DFFML which are one of the key elements of the application.

3. Detailed Description:

DFFML provides APIs for dataset generation and storage, and model definition using any machine learning framework, from high level down to low level use is supported. As the goal of DFFML is to build a community driven library of plugins for dataset generation and model definition, so that developers and researchers easily plug and play various pieces of data with various model implementations or generate datasets using the implemented features to increase the accuracy of output.
For this, DFFML needs to implement large number of machine learning models as well as various features. I would like to work on adding new models to DFFML during my summers. I have selected the models which are commonly used and can make DFFML appeal to a larger audience. I have planned to add the below listed Models/Algorithms to DFFML.

1. **Model 1**: Ordinary Least Square Regression (OLSR)
2. **Model 2**: Logistic Regression
3. **Model 3**: k-Nearest Neighbour (kNN)
4. **Model 4**: Naive Bayes

This is in no way an exhaustive list. I have curated a further list of models that I will like to implement to DFFML but given the time frame, I have limited my goals.

Further, I am open to swap models from the above list with models suggested by mentors as I am in the learning phase and can easily switch to something new.

Link to my curated list of models:
https://docs.google.com/document/d/1rxtTc-SA3oUWC9-2mjZreHS9vOBi_DsrcOmPpM5VCGU/edit?usp=sharing

Weekly timeline

Pre-GSoC (April 9 - 7 May):

- As DFFML is at an early stage, I would communicate with my mentors about the expectations and what exactly is the plan to move forward.
- Work on making myself further familiar with DFFML by being thorough with the codebase and working on issues.
- Finish up with my work on source compression
- Brush up all the necessary topics for the chosen models

Community Bonding (May 7 - May 26):

- Further continue the process of learning machine learning models
- Work on issues
- Communicate with my mentors and other selected students about their projects and how we can be of help to each other
- Try implementing one of the easier models with guidance from the mentor such that the GSoC code period becomes more fluid

Week 1 (May 27 - June 2):

- Finish the learning part of Model-1
- Start implementing the model on DFFML

Week 2 (June 3 - June 9):

- Wrap the model around DFFML
- Test the model with various datasets

Week 3 (June 10 - June 16):

- Start with learning part of Model-2
- After the learning process (as it might take time), start implementing it on DFFML

Week 4 (June 17 - June 23):

- Wrap the model around DFFML
- Test the model on various datasets

Week 5 (June 24 - June 30):

- I would like this to be a spare week. Complete the previous models if not complete and start preparing for the next models

Week 6 (July 1 - July 7):

- Finish the learning part of Model-3
- Start implementing the model on DFFML

Week 7 (July 8 - July 14):

- Wrap the model around DFFML
- Test the model with various datasets

Week 8 (July 15 - July 21):

- Finish the learning part of Model-4
- Start implementing the model on DFFML

Week 9 (July 22 - July 28):

- Wrap the model around DFFML
- Test the model on various datasets

Week 10 (July 29 - August 4):

- Finish the remaining work on the implemented models
- Start with testing and documentation

Week 11 (August 5 - August 11):

- Finish writing tests
- Discuss and finalize how the documentation should be with mentors

Week 12 (August 12 - August 18):

- Work on documentation
- Curate a list of datasets for the implemented models

Final week (August 19):

- Finish any remaining work
- Discuss future plans to work with DFFML after GSoC

Stretch Goals

When I am done implementing above algorithms, I will take suggestions from mentors for implementing more algorithms. Moreover, I can work on writing features for DFFML which are a crucial part of the application too. I have an idea to implement OpenCV features in DFFML which could be finalized after a discussion with the mentors.

Other Commitments

➢ End Semester Examination (8th May - 16th May):
  - Time Committed to GSoC: 3-4 hours/day
  - As the community bonding period will be on, I will make time for any meeting or conversations that are to be done.

I don't have any other commitments or summer plans during GSoC. Incase something comes up, I will make sure to inform my mentor a week before and limit my unavailability to not more 2 days.

Are you applying for other projects in SoC?

No, I am only applying for this project.

Further Contribution

I have selected DFFML as my organisation because the project really captures my interest that is machine learning. I have a lot of ideas to work on for DFFML even after GSoC ends and would love to stay a part of the community.