Sir Syed University of Engineering & Technology
Department of Software Engineering FYP Proposal Form

1. Project Credentials

a. Project Number: ___________________
(For FYP Member use only)

b. Project Title: Data Pulse

c. Group Members

S. No.	Group Member Name	Group Member Roll Number	Email (SSUET email address)
1	Abdul Moiz Chishti (GL)	2020F-SE-022	Se20f-022@ssuet.edu.pk
2	Shaheer Khan Qureshi	2020F-SE-003	Se20f-003@ssuet.edu.pk
3	Syed Abdul Aleem	2020F-SE-042	Se20f-042@ssuet.edu.pk

d. Project Status

🟧 Research Based Project

□ Product Based / Service Based Project

e. Project Category (Write down the name if not available)

Category Name	√ or X	Category Name	√ or X
Artificial Intelligence and Big Data	X	Cloud Computing and Cyber Security	
Augmented and Virtual Reality		Game Development	
E-Health		Graphics Animation	
E-Commerce		Nano Technology	
Internet of Things (IoT)		Shared Economy	
Block Chain		Other	

f. Project related to SDGs.

SDG Name	**√ or X**	**SDG Name**	**√ or X**
No Poverty		Zero Hunger	
Good Health and Well-Being		Quality Education	
Gender Equality		Clean Water and Sanitation	
Affordable and Clean Energy		Decent Work and Economic Growth	X
Industry, Innovation, and Infrastructure	X	Reduce Inequalities	
Sustainable Cities and Communities		Responsible Consumption and Production	
Climate Action		Life Below Water	
Life On Land		Peace, Justice, and Strong Institutions	
Partnerships for the Goals			

2. Project Background, Scope and Introduction

Background:

In today's data-driven world, the abundance of information poses a challenge: understanding and utilizing this data efficiently. Automated data profiling and machine learning offer potent solutions to extract valuable insights. However, the process of selecting the right machine learning model can be daunting due to the multitude of available algorithms and the need for in-depth analysis. To streamline this process, our project aims to develop an automated system that performs comprehensive data profiling, compares various machine learning models, and empowers users to download the trained model best suited for their data.

Scope:

The project's scope encompasses the development of an end-to-end platform that automates the data profiling process. It will employ advanced techniques to explore and analyze datasets. The platform will then provide a comparison of multiple machine learning models, evaluating their performance metrics based on the specific dataset characteristics. The system will generate a detailed report outlining the comparison results and provide the option to download the trained model in a deployable format, ensuring ease of integration into various applications. By using and implementing this project on various techniques we hope to bring better results in the end .

Introduction:

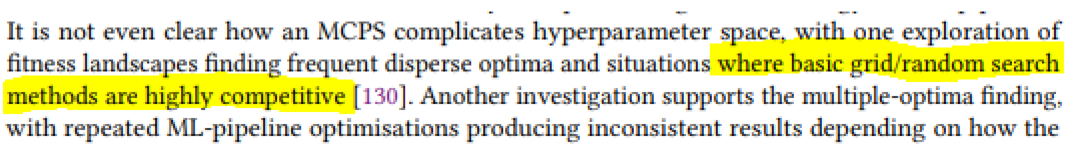
Welcome to our cutting-edge project focused on revolutionizing the way data is analyzed and machine learning models are selected! In a data-rich era, extracting meaningful insights has become a pivotal challenge. Our project aims to address this by introducing an automated system that combines the power of data profiling and machine learning model selection, ultimately empowering users to harness the potential of their datasets effectively.

By leveraging state-of-the-art techniques in data profiling, our platform will delve into the intricacies of your data and essential statistical insights. Subsequently, our system will embark on an extensive comparison of diverse machine learning models. Through a systematic evaluation process, it will determine the most suitable algorithms based on your data's characteristics.

Moreover, to facilitate practical utilization, our platform will allow users to effortlessly download the trained model in a format ready for deployment. This seamless integration ensures that the selected model can be readily applied across various applications, maximizing its usability and impact.

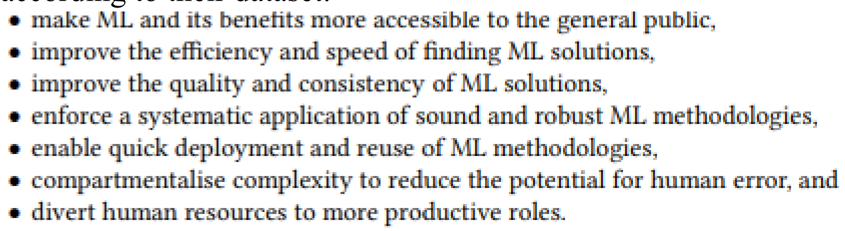
Join us on this transformative journey as we simplify data analysis and model selection, making the power of machine learning accessible to all.

3. Similar Projects and Literature Review:



It is not even clear how an MCPS complicates hyperparameter space, with one exploration of fitness landscapes finding frequent disperse optima and situations where basic grid/random search methods are highly competitive [130]. Another investigation supports the multiple-optima finding, with repeated ML-pipeline optimisations producing inconsistent results depending on how the

According to the source gathered from https://arxiv.org/pdf/2012.12600.pdf , we came to know that there are techniques named as grid and random but are highly competitive and difficult to implement. Instead, we would use other low-code methods which tends to find more accurate models much more efficiently. The users can download the most suited Machine Learning Model according to their dataset.



- make ML and its benefits more accessible to the general public,
- improve the efficiency and speed of finding ML solutions,
- improve the quality and consistency of ML solutions,
- enforce a systematic application of sound and robust ML methodologies,
- enable quick deployment and reuse of ML methodologies,
- compartmentalise complexity to reduce the potential for human error, and
- divert human resources to more productive roles.

By reviewing this literature we came up with an idea to build an efficient and accurate Model selection and Data profiling tool which would be easy to use for all .

4. Problem Statement

In today's data-driven landscape, organizations and individuals encounter the formidable challenge of efficiently harnessing vast datasets for meaningful insights. The complexity lies not only in understanding and profiling this data but also in selecting the most appropriate machine learning models to derive valuable predictions or classifications. The absence of a streamlined process for automated data profiling, comprehensive model comparison, and accessible model retrieval hampers the effective utilization of machine learning in diverse applications.

This project addresses the critical issues surrounding data analysis and machine learning model selection. Specifically, it aims to solve the following challenges:

1. **Data Profiling Complexity**:
   Understanding intricate datasets poses a significant hurdle due to the diversity and volume of information. Automated techniques are needed to explore and analyze data comprehensively.

2. **Model Selection Ambiguity:**
   With a myriad of machine learning algorithms available, selecting the most suitable model for a specific dataset becomes daunting. A systematic approach is required to compare and evaluate multiple models, considering the unique characteristics and requirements of the data.

3. **Lack of Accessibility:**
   Once a suitable model is identified, the process of obtaining and integrating the trained model into applications should be straightforward. The absence of an accessible means to download the trained model in a deployable format limits the practical application of machine learning solutions.

This project aims to bridge these gaps by developing an automated system that performs in-depth data profiling, conducts a thorough comparison of multiple machine learning models, and allows users to easily download the selected and trained model. By addressing these challenges, the project aims to empower users to derive actionable insights and seamlessly integrate machine learning solutions into their workflows.

5. Features

1. Automated Data Profiling:

- **Exploratory Data Analysis:** Conducts comprehensive exploration to unveil statistical characteristics within the dataset.
- **Data Visualization:** Generates visual representations (charts, graphs) to aid in understanding data distributions and relationships.
- **Statistical Summaries:** Provides detailed statistical summaries, including mean, median, standard deviation, etc., for each feature or attribute.

2. Machine Learning Model Comparison:

- **Multiple Model Support:** Allows comparison across various machine learning algorithms, including regression, classification etc.
- **Performance Metrics Evaluation:** Measures and compares model performance using standard metrics like accuracy, precision, recall, F1-score, and ROC curves.

3. Model Training and Selection:

- **Automated Training:** Training of multiple models.
- **Cross-Validation:** Implements cross-validation techniques to ensure robustness and reliability in model selection.
- **Optimal Model Identification:** Recommends the most suitable model based on performance evaluation and dataset characteristics.

4. User Interface and Reports:

- **Interactive Dashboard:** Provides an intuitive user interface for easy navigation and control over the profiling and model comparison process.
- **Comprehensive Reports:** Generates detailed reports summarizing data profiling insights and model comparison results for user understanding and decision-making.

5. Model Download and Deployment:

- **Downloadable Trained Models:** Allows users to download the selected and trained model.

These features collectively form a robust and user-friendly system, empowering users to efficiently profile their data, compare machine learning models, and seamlessly integrate the selected model into their applications.

6. Expected Tools and Technology Requirements

1. Programming Languages:

- **Python:** Widely used for its extensive libraries in data analysis (Pandas, NumPy), machine learning (Scikit-learn, TensorFlow, PyTorch), and visualization (Matplotlib, Seaborn).

2. Data Profiling and Analysis:

- Pandas: For data manipulation, cleaning, and exploratory data analysis.
- NumPy: Efficient handling of numerical operations and arrays for data analysis.
- Dask: For parallel computing and scaling with large datasets.
- SQLAlchemy: Interaction with databases for profiling data stored in SQL.

3. Machine Learning Model Comparison:

- Scikit-learn: Provides a vast array of machine learning algorithms for model comparison and evaluation.
- TensorFlow or PyTorch: For more complex neural network-based models and deep learning comparisons.
- XGBoost, LightGBM, CatBoost: Popular gradient boosting libraries for comparative analysis with tree-based models.

4. Visualization and Reporting:

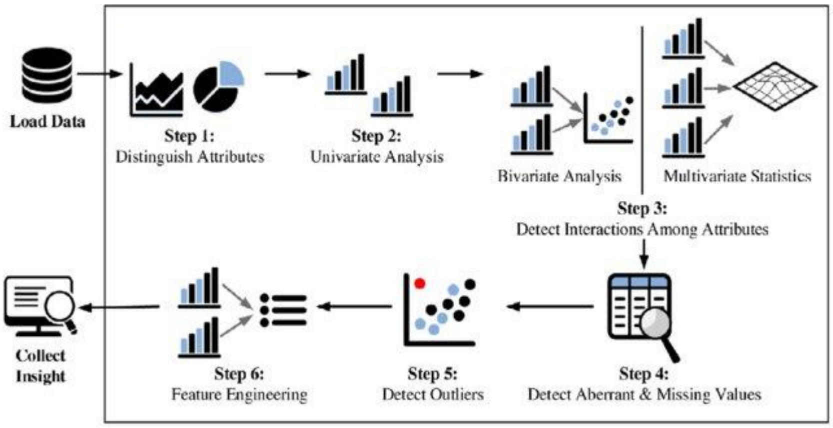
- Matplotlib and Seaborn: For creating static visualizations.
- Plotly or Bokeh: Interactive visualization libraries to enhance user interaction and understanding.
- Jupyter Notebooks or Streamlit: Interactive interfaces for presenting reports and results.

5. User Interface and Interactivity:

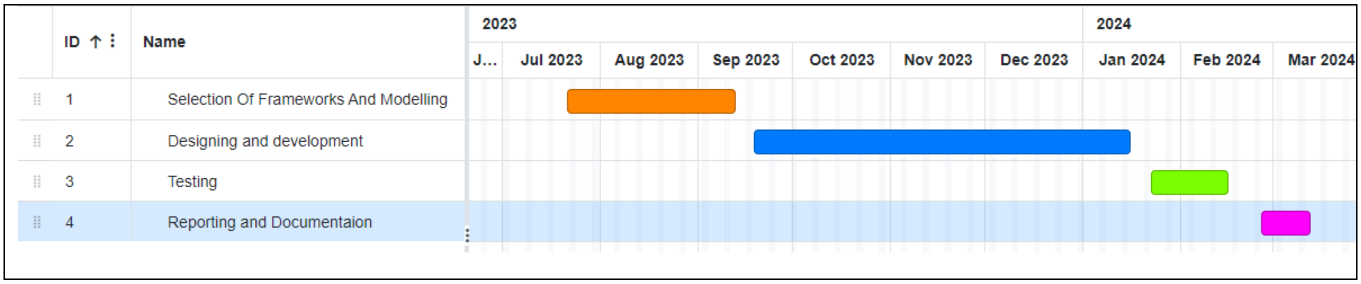
- Streamlit or Dash: Frameworks for creating interactive web applications for user interaction and displaying results.
- HTML/CSS/JavaScript: For customizing and enhancing the user interface of the application.

These tools and technologies provide a strong foundation for building a comprehensive system that automates data profiling, compares machine learning models, and facilitates the retrieval of trained models.

7. Design and Development Methodology



8. Project Planning



ID ↑	Name
1	Selection Of Frameworks And Modelling
2	Designing and development
3	Testing
4	Reporting and Documentaion

Abdul Moiz Chishti will oversee the entire development process, ensuring that the project stays on track, milestones are met, and resources are effectively managed and will coordinate with team members.

Shaheer Khan Qureshi and Syed Abdul Aleem will be responsible for creating an intuitive and visually appealing user interface as well as help to meet the project needs by collaborating their efforts for making the development process easier as per their expertise.

Abdul Moiz Chishti will be responsible for implementing the server-side logic of the application. He will develop the application's core functionalities, including data processing, analysis algorithms and will also ensure the application's scalability, and performance.

For the testing period, All the members will collaborate to ensure its performance and results.

9. Letters/Recommendations

Collaboration Letter is attached from a Logistics Company named Raptor Global logistics (Pvt) Ltd.

10. References

https://towardsdatascience.com/automating-scientific-data-analysis-part-1-c9979cd0817e
https://scripts.iucr.org/cgi-bin/paper?a56909
https://www.researchgate.net/publication/263671317_Automated_data_analysis
https://www.stitchdata.com/resources/automated-data-analytics/
https://arxiv.org/pdf/2012.12600.pdf

11. Appendices

Expected FYP Supervisors

1. Miss Kiran Hidayat
2. Sir Hassan Zaki
3. Dr. Danish Jamil

FYP Committee Comments (For FYP Use Only)

Decision	√ or X	Remarks
Project Accepted		
Project Accepted with Modifications		
Project Needs Major Revision		
Project Disapproved		

Other Comments and Suggestions.