MPSSDEGB
कौशल से रोजगार, समृद्धि अपार

FITT

SANKALP
Ministry of Skill Development & Entrepreneurship

Capstone Project Kickoff

INSTRUCTOR NAME: DATE:

DAY NO. 18

SECTION: B1M7L19T1

Contents

- Capstone Project Guidelines Presentation Understanding expectations and evaluation criteria.
- Team formation and initial planning
- Assigning roles and setting milestones.
- Mentor-Mentee meetings to discuss project direction.
- Setting up a project repository and documentation standards.

ML Project Developmnet::Methodology

- **Define the Problem:**
- **Collect and Explore Data (Data visualization):**
- **Data Preprocessing and Feature Engineering:**
- **Split Data into Training and Testing Sets:**
- **Choose a Model:(ML algorithm)**
- **Train the ML learning algorithm:**
- **Evaluate Model Performance-metrics include accuracy, precision, recall, F1 score,**
- **Visualize the model performance**

ML Project Developmnet::Methodology

- **Define the Problem:**
- **State the context, challenges, and objectives**
- **E.g. Weather forecasting**

- Weather forecast prediction plays a crucial role in various sectors, including agriculture, transportation, and emergency management, as it helps in making informed decisions to mitigate potential risks and optimize resource utilization.
- The current weather forecasting systems often rely on traditional numerical models that struggle to accurately predict local and short-term weather conditions. This limitation poses significant challenges for various sectors, including agriculture, transportation, and emergency management.
- The need for more precise and reliable weather forecasts has prompted the exploration of machine learning (ML) approaches.
- However, the development of an efficient ML-based weather forecast prediction system requires addressing issues such as data quality, model complexity, and interpretability.

ML Project Developmnet::Methodology

- **Collect and Explore Data (Data visualization):**
  - **Collect the data from the repositories like UCI, KEEL, Kaggle, and other sources (Ref. Data repositories slide)**

- **Data Visualization/ Exploration of data:**
  - **Enhances Understanding:** Transforms raw data into visual formats, aiding in comprehension and identifying patterns.
  - **Facilitates Decision-Making:** Enables quick and informed decision-making by presenting key insights visually.
  - **Identifying Trends and Patterns:** Visualizations help in spotting trends, outliers, and correlations that may go unnoticed in raw data.

ML Project Developmnet:: Data Repositories

- **UCI Machine learning Repository (600+ Datasets):** https://archive.ics.uci.edu/datasets
- **Kaggle Data sets (10K+ Datasets):**
  - https://www.kaggle.com/datasets?tags=12107-Computer+Science
- **Knowledge Extraction Evolutionary Learning (KEEL) Data sets (600+ Datasets)**
  - https://sci2s.ugr.es/keel/datasets.php
- **Other Resources:** https://towardsdatascience.com/top-sources-for-machine-learning-datasets-bb6d0dc3378b

ML Project Developmnet::Methodology

- **Data Preprocessing and Feature Engineering: (Ref. Day 8 slides)**
- **Perform the following as applicable on the data collected:**
  - Data Cleaning
  - Data Transform
  - Data Reduction
  - Categorical Encoding
  - Scaling

ML Project Developmnet::Methodology

- **Split Data into Training and Testing Sets:**
  - Split the given data into Training, Validation and Testing data sets. E.g 60:20:20 ratio or 80:10:10 ratio, and so on
  - While splitting, preserve the class proportion in Trg., Testing, Validation sets. E.g. Positive and Negative classes must be distributed evenly in 3 sets.
  - Use k-fold cross validation

ML Project Developmnet::Methodology

- **Choose a Model:(ML algorithm):**
  - Choosing the right machine learning algorithm depends on various factors such as the nature of the problem, the characteristics of the data, and the desired outcome.
  - Ref. Word document for more details.

ML Project Developmnet::Methodology

- **Train the Learning Algorithm: (Ref. Day 9 slides)**
  - Use appropriate libraries of Python and build the models for the training set
  - Use atleast 5 sets of training/testing datasets and take average performance to get reliable estimate. E.g. Use k-fold stratified Cross validation.

ML Project Developmnet::Methodology

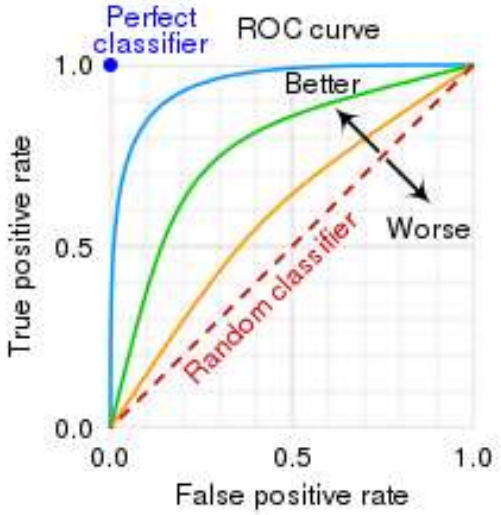
- **Evaluate Model Performance (Ref. Day 9 Slides)**
- Use variety of performance metrics to assess t model performance, atleast
  - Accuracy,
  - Precision, Recall, F1-score
  - MSE (for regression)
  - AUROC (Desirable).
- Compare atleast 5 models for the given task

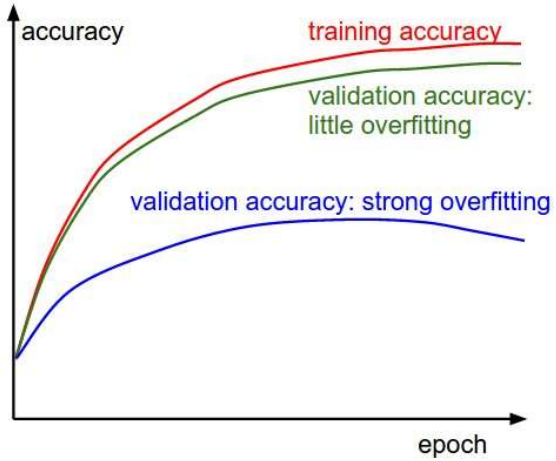
	All Features		Feature Selection		Total Acceleration	
Classifier	Acc	F1	Acc	F1	Acc	F1
LDA	98.29	98.21	97.52	97.41	87.05	86.55
DT	97.14	97.02	97.52	97.43	92.95	92.68
RSVM	61.90	42.88	89.9	88.93	92.38	92.08
LSVM	**98.48**	**98.41**	97.71	97.62	88.38	87.83
1NN	98.10	98.03	97.90	97.82	92.38	92.12
3NN	98.10	98.03	97.90	97.81	91.81	91.56
5NN	98.29	98.22	98.29	98.21	93.52	93.31
7NN	97.71	97.63	**98.48**	**98.41**	92.38	92.11

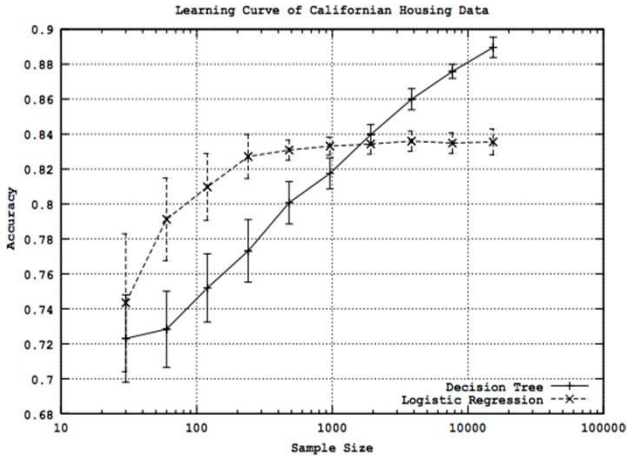
Note: Bold denotes the overall best performance. Underlined results denote the best performance per feature.

ML Project Developmnet::Methodology

- **Visualize the model performance:**
- Use charts, line graphs, Learning curves, ROC curves, Precision-Recall curves







CapStone Project

Homework:
Begin work on the Capstone Project, focusing on research and design.