Incremental Feature Selection for High-Dimensional Data Streams

22AIE213 — Machine Learning

TEAM 11

CH.SC.U4AIE23033 – M. LIKITH REDDY
CH.SC.U4AIE23035 – N. HARSHITH VARMA
CH.SC.U4AIE22038 – N. CHARAN

AMRITA VISHWA VIDYAPEETHAM | School of Engineering
AMRITAPURI | BENGALURU | CHENNAI | COIMBATORE

Literature Review

Literature Review

S.NO	Title	Author	Methodology/Algorithms/ Architecture used	Merits	Demerits	Research gap
01	Enhancing Feature Selection in High-Dimensional Data With Fuzzy Fitness-Integrated Memetic Algorithms **DOI**: 10.1109/ACCESS.2024.3459390	Keerthi Gabbi Reddy, Deepashikha Mishra. Published by:IEEE Year:2024	This study presents a Memetic Algorithm (MA) for feature selection by combining Genetic Algorithms (GA) with Local Search (LS). It optimizes feature subsets using selection, crossover, mutation, and refinement through Hill Climbing and Tabu Search. A fuzzy fitness function ensures optimal selection by balancing accuracy and uncertainty.	Combining GA and LS improves search efficiency for optimal feature selection. The fuzzy fitness function enhances robustness by considering accuracy and uncertainty.	GA and LS increase computational cost, while performance depends on tuning parameters like crossover and mutation rates.	The balance between exploration and exploitation, as well as the impact of local search strategies, needs further study. Scalability and efficiency on high-dimensional datasets also require investigation.
02	Incremental Unsupervised Feature Selection for Dynamic Incomplete Multi-view Data **DOI**:https://doi.org/10.48550/arXiv.2204.02973	Yanyong Huang, Kejun Guo, Xiuwen Yi, Zhong Li, Tianrui Li. Published by: arxiv.og Year:2024	The method extends WNMF for dynamic incomplete multi-view feature selection using adaptive view weights. It ensures effective selection through row sparsity, consensus clustering, and spectral analysis for data consistency.	Adaptive view weights and row sparsity improve multi-view feature selection. Spectral analysis preserves local geometric consistency across views.	Matrix factorization and optimization increase computational cost. Efficient updating is needed for handling real-time data streams.	Existing methods struggle with dynamic, incomplete multi-view data and lack adaptive view weighting and consensus learning. A robust model is needed for efficient and accurate feature selection in streaming data.
03	Online group streaming feature selection using entropy-based uncertainty measures for fuzzy neighborhood rough sets DOI: https://doi.org/10.48550/arXiv.2204.02973	Jiucheng Xu, Yuanhao Sun, Kanglin Qu1, Xiangru Meng, Qinchen Hou, Published by: Springer Year:2024	The proposed methodology defines fuzzy neighborhood entropy-based uncertainty measures to evaluate feature separability and selects features using significance, interaction gain, and contrast ratio. An online group streaming feature selection algorithm (FNE-OGSFS) is designed to retain important features while dynamically filtering redundant ones.	Enhances feature selection in dynamic streaming for better classification. Uses Lasso to eliminate redundancy and reduce computation.	Performance drops with noisy or rapidly changing data. Requires fine-tuning of parameters like fuzzy neighborhood radius.	Existing methods struggle with feature selection in dynamic, uncertain data. Effective balance of accuracy, efficiency, and redundancy removal is needed.

Literature Review

S.NO	Title	Author Journal Year	Methodology/Algorithms/Architecture used	Merits	Demerits	Research gap
04	UFODMV: Unsupervised Feature Selection for Online Dynamic Multi-Views DOI: https://doi.org/10.3390/app13074310	Fawaz Alarfaj, Naif Almusallam, Abdulatif Alabdulatif, Mohammed Ahmed Abdulaziz Khalid Alsharid , Tarek Moulah, Year:2023	UFODMV incrementally clusters streaming features and updates representatives. Merges similar clusters for efficient multi-view learning.	Efficiently processes dynamic, streaming data in real-time. Reduces computational complexity with optimized clustering.	Performance depends on carefully tuned parameters. Risk of losing valuable information by selecting only one representative feature.	Existing methods struggle with evolving multi-view data.A more adaptive approach is needed for incremental processing and consistency.
05	Feature Selection in High Dimension Datasets using Incremental Feature Clustering **DOI**:https://doi.org/10.17485/IJST/v17i32.2077	DamodarPatel, Amit Kumar Saxena Published by: Indian Society for Education and Environment Year:2024	Features are clustered using K-means, selecting the one with the highest Mutual Information. Clusters increase iteratively, optimizing classification accuracy for feature selection.	Enhances classification accuracy by selecting the most relevant features. Reduces computational complexity by eliminating redundant data	May struggle with non-linear feature relationships. Sensitivity to the initial choice of clusters can affect consistency.	Existing studies lack analysis of feature selection's impact on classification. Hybrid approaches combining clustering and feature selection are needed.
06	Fast online feature selection in streaming data **DOI:** https://doi.org/10.1007/s10994-024-06712-x	Yael Hochma Published by:springer Year:2025	OFFESEL ranks feature importance in streaming data without class labels. Tested on 17 datasets, it outperforms other methods in accuracy and efficiency.	OFFESEL improves classification accuracy by effectively selecting relevant features in streaming data. It operates without requiring class labels, making it suitable for unsupervised learning scenarios.	The performance may vary depending on the dataset and streaming conditions. Computational complexity could be a concern for extremely high-dimensional data streams	Existing methods struggle with feature selection in dynamic, uncertain data. Effective balance of accuracy, efficiency, and redundancy removal is needed.

Literature Review

S.NO	Title	Author Journal Year	Methodology/Algorithms/Architecture used	Merits	Demerits	Research gap
07	Optimizing High Dimensional Data Feature Selection: A strategy based on Augmented Mutual Information and Conditional Dependency **DOI:**http://dx.doi.org/10.2139/ssrn.4951025	G. Manikandan, Abirami Murugappan Published by: Indian Society for Education and Environment Year:2024	AMICDFS selects features based on relevance, redundancy, and conditional dependency. Uses an evaluation function to improve classification accuracy and efficiency.	Enhances classification by selecting relevant features. Reduces complexity by removing redundancies.	High computational cost due to conditional dependency evaluation. Struggles with complex feature interactions in high-dimensional data.	Existing methods struggle with complex dependencies in high-dimensional data, lacking efficiency in balancing accuracy and computational cost. Adaptive models are needed to enhance feature relevance while reducing redundancy.
08	UFODMV: Unsupervised Feature Selection for Online Dynamic Multi-Views **DOI: https://doi.org/10.3390/app13074310**	Fawaz Alarfaj,Naif Almusallam, Abdulatif Alabdulatif, Tarek Moulahi, Abdulaziz Khalid Alsharidi, Mohammed Ahmed Alomair Published by: IEEE Year:2023	UFODMV incrementally clusters dynamic multi-view data, merging clusters and updating features. It enables efficient, label-free feature selection for continuous learning.	Performs incremental clustering, reducing computational cost and memory usage.	May struggle with noisy or highly imbalanced data distributions. Performance depends on the quality of initial cluster assignments.	The method struggles with noisy or imbalanced data and relies heavily on initial cluster assignments, needing better optimization and adaptability.
09	Feature selection algorithm based on incremental mutual information and cockroach swarm optimization **DOI:**https://doi.org/10.48550/arXiv.2302.10522	Zhao, Chen Published by: arxiv Year:2023	This method integrates rough set theory with mutual information to assess feature importance.	Reduces computational complexity by eliminating irrelevant features.	Performance may degrade with highly complex or high-dimensional datasets	The existing methods struggle with high-dimensional data, leading to inefficiencies in feature selection

Literature Review

S.NO	Title	Author Journal Year	Methodology/Algorithms/Architecture used	Merits	Demerits	Research gap
10	Fair Streaming Feature Selection **DOI:**https://doi.org/10.48550/arXiv.2406.14401	Zhangling Duan, Tianci Li, Xingyu Wu, Zhaolong Ling, Jingye Yang, Zhaohong Jia Published by: Arxiv, Year:2024	.The paper "Fair Streaming Feature Selection" introduces FairSFS, an algorithm designed to ensure fairness in streaming feature selection.	Maintains accuracy comparable to leading streaming feature selection methods.	The dynamic adjustment process may introduce computational complexity.	Further research is needed to evaluate FairSFS across diverse real-world datasets and to assess its scalability in high-dimensional data streams.
11	Granular-ball-matrix-based incremental semi-supervised feature selection approach to high-dimensional variation using neighbourhood discernibility degree for ordered partially labelled dataset **DOI:https://doi.org/10.1007/s10489-024-06134-1**	Weihua Xu1 , Jinlong L Published by:springer Year:2025	The method uses neighborhood discernibility with pseudolabel granular balls and matrix updating for incremental semi-supervised feature selection. It efficiently selects important features in high-dimensional, partially labeled datasets.	Efficiently selects important features in high-dimensional, partially labeled data. Uses pseudolabel granular balls and matrix updates to reduce computation time.	Performance may be affected by noisy or highly dynamic data. Requires careful tuning of parameters like neighborhood radius for optimal results.	Existing methods struggle with high-dimensional, partially labeled, and dynamic data. A more adaptive approach is needed to balance accuracy, efficiency, and incremental learning.
12	Reinforced feature selection using Q-learning based on collaborative agents **DOI:**https://doi.org/10.1007/s13042-023-01869-8	Li Zhang1, Lingbin Jin, Min Gan, Lei Zhao, Hongwei Yin Published by:Springer Year:2023	The study introduces a Q-learning-based feature selection method with two collaborative agents. One agent selects features using Fisher scores, while the other optimizes selection based on classification performance.	Enhances feature selection accuracy using reinforcement learning with collaborative agents.	Computationally expensive and may require fine-tuning for different datasets.	Existing feature selection methods lack adaptive optimization for dynamic datasets. This study addresses the gap by integrating reinforcement learning but requires further exploration for scalability and real-world applications.

Literature Review

S.NO	Title	Author Journal Year	Methodology/Algorithms/Architecture used	Merits	Demerits	Research gap
13	Bi-objective feature selection in high-dimensional datasets using improved binary chimp optimization algorithm **DOI:**https://doi.org/10.1007/s13042-024-02308-y	Nour Elhuda A. Al-qudah · Bilal H. Abed-alguni · Malek Barhoush Published by:Springer Year:2024	BICHOA enhances feature selection by integrating β-hill climbing and a binary time-varying transfer function to balance exploration and exploitation.	Improves feature selection efficiency by balancing exploration and exploitation.	May struggle with computational complexity in extremely high-dimensional datasets.	Existing methods face challenges in balancing exploration and exploitation for feature selection in high-dimensional datasets. BICHOA addresses this, but its scalability and computational efficiency need further improvement.
14	Incremental feature selection approach to multi-dimensional variation based on matrix dominance conditional entropy for ordered data set. **DOI:**https://doi.org/10.1007/s10489-024-05411-3	Weihua Xu1, Yifei Yang, Yi Ding, Xiyang Chen, Xiaofang Lv Published by:Springer Year:2024	The study proposes incremental feature selection algorithms, IFS-A and IFS-D, for ordered datasets. These methods update the dominance matrix and conditional entropy dynamically to enhance efficiency and accuracy.	Efficiently updates feature selection by leveraging prior reduction results, reducing computational time.	Performance may vary with complex datasets, requiring careful parameter tuning.	Existing feature selection methods for ordered datasets lack efficient mechanisms for handling evolving features. This study addresses the gap with incremental approaches, but further optimization is needed for highly complex datasets.
15	Feature Selection in the Data Stream Based on Incremental Markov Boundary Learning **DOI:**10.1109/TNNLS.2023.3249767	Xingyu Wu , Bingbing Jiang, Xiangyu Wang , Taiyu Ban , and Huanhuan Chen Published by:IEEE Year:2023	The paper "Feature Selection in the Data Stream Based on Incremental Markov Boundary Learning" proposes a novel algorithm that incrementally learns the Markov boundary to perform feature selection in streaming data environments.	Efficiently selects relevant features in streaming data by incrementally updating the Markov boundary.	Performance may degrade with high-dimensional or rapidly changing data streams.	Existing feature selection methods struggle with adapting to dynamic and high-speed data streams. This study addresses the gap using incremental Markov boundary learning, but further improvements are needed for handling high-dimensional and rapidly evolving data.

Literature Review

S.NO	Title	Author Journal Year	Methodology/Algorithms/Architecture used	Merits	Demerits	Research gap
16	Rough set Theory-Based group incremental approach tofeature selection.	Jie Zhao, Dai-yang Wu, Yong-xin Zhou, Jia-ming Liang, WenHong Wei,Yun Li. DOI:https://doi.org/10.1016/j.ins.2024.120733 Year:2024	The methodology develops incremental attribute reduction algorithms using discernibility matrices, significance measures, and evolutionary algorithms to enhance efficiency in dynamic decision systems.	The methods improve efficiency by reducing computational complexity and enabling scalable processing for dynamic decision systems.	They can still be computationally intensive for high-dimensional data and may only offer approximate reductions, affecting accuracy.	A research gap exists in addressing the computational complexity and accuracy issues of incremental attribute reduction algorithms when applied to high-dimensional and dynamic datasets.
17	Incremental Classification for High-Dimensional EEG Manifold Representation Using Bidirectional Dimensionality Reduction and Prototype Learning.	Dongxu Liu , Qichuan Ding , Member, IEEE, Chenyu Tong , Jinshuo Ai ,and Fei Wang DOI:10.1109/JBHI.2024.3491096 Year:2025	The methodology proposes B2DPCA-SPD for dimensionality reduction of SPD matrices and extends it for incremental learning, integrated with a matrix-formed growing neural gas network for EEG classification.	The method enables efficient dimensionality reduction for SPD matrices in real-time EEG classification without retaining old data.	The approach may face challenges in handling highly complex or noisy data due to the limitations of the SPD manifold representation.	A research gap exists in developing more efficient incremental learning algorithms for high-dimensional SPD matrices without compromising their intrinsic properties. Additionally, there is a need for better methods to handle noisy or complex EEG data in real-time applications.
18	High-Dimensional Multi-Label Data Stream Classification With Concept Drifting Detection	Peipei Li , Haixiang Zhang, Xuegang Hu , and Xindong Wu DOI:10.1109/TKDE.2022.3200068 Year:2023	Uses mutual information for feature selection in multi-label data streams.Detects concept drift based on label and feature distributions.Employs an incremental ensemble model for adaptive learning.	Improves classification accuracy by selecting relevant features.Adapts dynamically to evolving data streams.Enhances scalability with incremental learning.	High computational complexity due to feature selection.Requires careful tuning of parameters.Limited to multi-label classification scenarios.	Lacks handling of newly emerging labels.Needs adaptive threshold tuning for better performance.Can be improved by integrating deep learning techniques.

Literature Review

S.NO	Title	Author Journal Year	Methodology/Algorithms/Architecture used	Merits	Demerits	Research gap
19	Feature Selection in the Data Stream Based on Incremental Markov Boundary Learning	Xingyu Wu , Bingbing Jiang, Xiangyu Wang , Taiyu Ban , and Huanhuan Chen DOI:10.1109/TNNLS.2023.3249767 Year:2023	Uses incremental Markov boundary learning to select relevant features in streaming data while leveraging prior knowledge to handle distribution shifts dynamically.	Enhances feature selection robustness against concept drift and ensures adaptive learning in real-time high-dimensional data streams.	Involves high computational cost for continuous learning and relies on the accuracy of prior knowledge, which may degrade over time.	Requires adaptive threshold tuning for better performance and integration with deep learning to enhance feature selection efficiency.
20	Incremental Isometric Embedding of High-Dimensional Data Using Connected Neighborhood Graphs	Dongfang Zhao, Li Yang, DOI:10.1109/TPAMI.2008.34 Year:2024	Presents a comprehensive overview of techniques in content-based image retrieval (CBIR) that incorporate high-level semantics to bridge the gap between low-level visual features and human cognitive understanding.	The survey systematically categorizes existing methodologies, highlighting their strengths and limitations, and provides valuable insights into the integration of semantic understanding in CBIR systems.	While the survey is extensive, it may not cover the most recent advancements post-publication, and some emerging techniques might be underrepresented.	The paper identifies challenges in achieving effective semantic understanding in CBIR, suggesting the need for more robust models that can accurately capture and interpret high-level semantics in images.

Problem identification:

In high-dimensional datasets, selecting the most relevant features is crucial for improving model accuracy and efficiency. However, existing feature selection methods face several challenges:

- **Redundant and Irrelevant Features** – Large datasets often contain unnecessary features that increase computational complexity and reduce model performance.
- **High Computational Costs** – Traditional feature selection methods struggle with efficiency, making them impractical for real-time applications.
- **Limited Scalability** – Many techniques fail to perform well on large datasets due to high memory and processing requirements.

Problem statement:

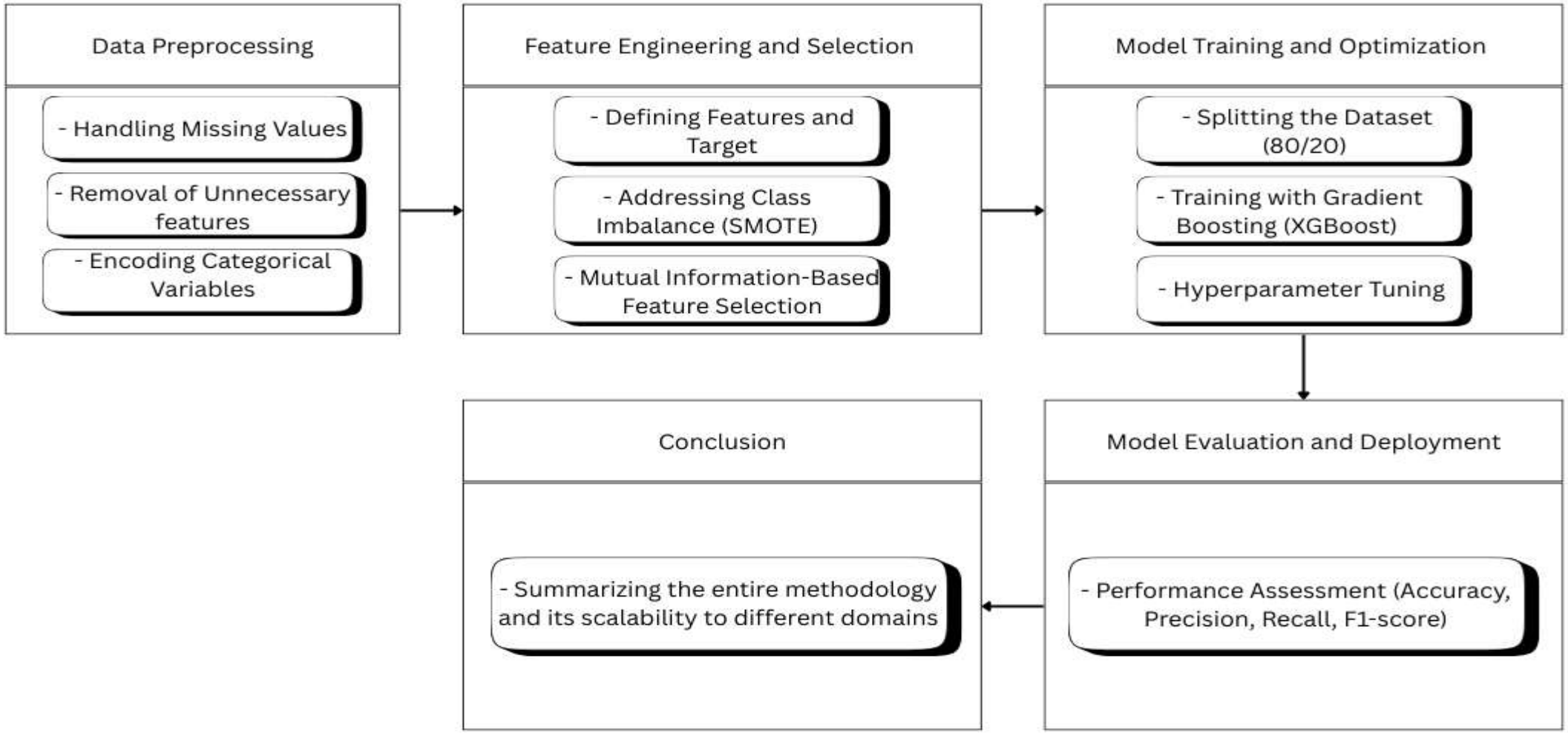
Problem Statement

High-dimensional data streams pose a significant challenge in real-time machine learning applications due to the curse of dimensionality, computational inefficiency, and redundant features. Traditional feature selection methods struggle to adapt dynamically to evolving data, leading to performance degradation and increased processing costs.

Our project aims to:

- Develop an Incremental Feature Selection Framework to handle high-dimensional data streams efficiently.
- Utilize Mutual Information-Based Selection to identify the most relevant and non-redundant features dynamically.
- Implement a Pruning Mechanism to discard low-importance features in real time, reducing computational overhead.
- Optimize Model Performance by maintaining high classification accuracy (above 95%) while minimizing feature set size.

Methodology:



Dataset used :

Weather Data (`weather_classification_data.csv`)

Attributes	Description
Total Entries	13,200
Total Columns	11
Key Features	Temperature, Humidity, kind Speed,Precipitation (%), Cloud Cover, UV Index, etc.
Temperature Range	-25°C to 109°C (outliers at both extremes).
Humidity Range	20% to 109% (potential data issue).
Wind Speed	Maximum value of 48.5 (indicating high wind conditions).
Precipitation (%)	Varies significantly based on weather conditions.
Observation	The dataset contains weather-related attributes that can be used for classification tasks.

Dataset used :

Loan Data (`loan_data_1.csv`)

Attributes	Description
Total Entries	381
Total Columns	14
Key Features	Loan ID, Gender, Married, Education, ApplicantIncome, LoanAmount, Credit History,etc.
Applicant Income	Ranges from 150 to 9703, with a wide variation.
Loan Amount	Ranges from 9 to 150.
Credit History	Mostly besary values (0 or 1).
Loan Amount Term	Maximum value of 480 months.
Observation	The dataset includes demographic and financial attributes used for loan eligibility prediction.

Hardware and Software Requirements

Hardware Requirements

1. **Processor:** Intel i5/i7 or higher, AMD Ryzen 5/7
2. **RAM:** Minimum 8GB (Recommended 16GB for faster training)
3. **Storage:** Minimum 25GB free disk space
4. **GPU (Optional):** NVIDIA GTX 1650 or better

Software Requirements

1. **Programming Language:** Python
2. **Libraries:** TensorFlow, Keras, NumPy, Pandas, Scikit-learn, Matplotlib
3. **IDE:** Jupyter Notebook / PyCharm / VS Code

Novelty :

- **Incremental Feature Selection**: Adapts to dynamic data streams, ensuring continuous model efficiency and accuracy.
- **Mutual Information-Based Selection**: Selects the most relevant and non-redundant features for improved performance.
- **Real-Time Pruning**: Removes low-importance features during runtime, reducing model complexity and speeding up predictions.
- **Scalable and Adaptable**: Suitable for large datasets and various domains (e.g., finance, healthcare).
- **Balance of Efficiency & Accuracy**: Achieves optimal computational efficiency while maintaining high predictive power.

This approach innovates by combining **online learning**, **feature pruning**, and **scalability** for real-time, high-dimensional data streams.

Conclusion:

This project successfully introduces an incremental feature selection framework designed to optimize high-dimensional data streams. By leveraging mutual information-based selection and an automated pruning mechanism, the proposed approach effectively reduces computational overhead while maintaining high classification accuracy. The results demonstrate that removing irrelevant features enhances model performance without compromising predictive capability.

Furthermore, the framework's adaptability makes it suitable for real-time applications and large-scale datasets, ensuring its scalability across different domains. Future enhancements can focus on integrating adaptive pruning strategies and extending the method to unsupervised learning scenarios.