NTAD Dam Hazard Classification Predictive Model

Devin Park

October 2025

Contents

1 Project Goal

The primary objective of this project is to construct a predictive model for assessing the hazard potential of dams using the NTAD_Dams dataset. The model aims to classify dams according to their hazard status by analyzing a combination of environmental, structural, and locational attributes. This study employs interpretable machine learning algorithms, including OneR, Naive Bayes, J48, and BayesNet, to evaluate and compare their predictive performance in identifying high risk dams. By determining the most influential factors contributing to hazard classification, this project seeks to provide data driven insights that can inform public safety decisions, guide hazard mitigation efforts, and support agencies such as FEMA and the United States Army Corps of Engineers in the prioritization of dam inspections and maintenance planning.

2 Description of Dataset

The dataset used in this study is a representation of the National Inventory of Dams, which is maintained and published by the United States Army Corps of Engineers in cooperation with the Association of State Dam Safety Officials, as well as state, territorial, and federal agencies (U.S. Department of Transportation, 2025). It is also part of the United States Department of Transportation Bureau of Transportation Statistics National Transportation Atlas Database. The dataset documents more than ninety thousand dams across the United States and its territories and serves as a comprehensive resource for understanding dam characteristics and hazard potential. The dataset classifies dams according to hazard potential (High, Significant, Low, Undetermined), which reflects the probable consequences of failure, not the probability of failure itself. In the dataset, each hazard potential category is assigned an id: High is 4, Signifcant is 3, Low is 2, and Undetermined is 1. The class for our machine learning models become **hazardId**.

Official Attributes and Definitions

The following list enumerates the official attribute names and their definitions as provided in the accompanying data dictionary https://doi.org/10.21949/1529016

Attribute	Definition
OBJECTID	Internal feature number.
id	Unique identifier.
federalId	The unique identifier for each dam record. For most dams, federalID is the NID ID of the dam prior to the NID data transmittal by the submitting agency.
name	The official name of the dam. For dams that do not have an official name, one is assigned by the agency.
latitude	Dam latitude as a single value, in decimal degrees.
longitude	Dam longitude as a single value, in decimal degrees.
hazardId	Category indicating potential hazard to the downstream area if a failure were to occur.
hazard	Hazard potential classification as a text value.
city	Name of the nearest city to the dam.
county	County name where the dam is located.
state	Two letter postal abbreviation for the state where the dam is located.
nidHeight	Maximum retaining height of the dam, measured in feet.
nidStorage	Maximum storage at the dam, measured in acre feet.
nidSurfaceArea	Surface area of the reservoir at normal storage, measured in acres.

Attribute	Definition
nidDrainageArea	Size of the area that drains into the reservoir, measured in square miles.
nidCrestLength	Length of the dam, measured along the top of the dam, in feet.
nidCrestElevation	Elevation of the top of the dam, measured in feet above mean sea level.
nidType	Types of materials used to construct the dam.
nidOwnerType	Type of ownership.
nidOwnerName	Owner name.
nidPrimaryPurposeId	Code for the primary purpose of the dam.
nidPrimaryPurpose	Description of the primary purpose of the dam.
nidOtherPurposeId	Code for the other purposes of the dam.
nidOtherPurpose	Description of the other purposes of the dam.
nidYearCompleted	Year when construction of the dam was completed.
nidYearModified	Year when the dam was last modified.
nidHydrologicUnit	Hydrologic Unit Code.
nidRiver	Name of the principal river or stream on which the dam is built.
nidNearestCity	Name of the nearest city or community.
nidCongressionalDistrict	Congressional district in which the dam is located.
nidLongitude	Longitude in decimal degrees at the dam location.
nidLatitude	Latitude in decimal degrees at the dam location.
nidEmergencyActionPlanId	Code value to indicate whether the dam has an Emergency Action Plan.
nidEmergencyActionPlan	Text description of whether the dam has an Emergency Action Plan.
nidInspectionFrequencyId	Code to indicate how often the dam is inspected.
nidInspectionFrequency	Description of how often the dam is inspected.
nidInspectionDate	Date of the most recent inspection.
nidRegulatoryId	Code for the regulatory agency responsible for dam safety oversight.
nidRegulatory	Text description of the regulatory agency responsible for dam safety oversight.
nidFederalAgencyId	Code for the federal agency involved with the dam.
nidFederalAgency	Name of the federal agency involved with the dam.
nidMaxStorage	Maximum storage at the dam, in acre feet.
nidNormalStorage	Normal storage at the dam, in acre feet.
nidDrainageAreaMi2	Drainage area in square miles.
nidSurfaceAreaAcres	Surface area of reservoir at normal storage, in acres.
nidCrestElevationFt	Crest elevation, in feet.
nidCrestLengthFt	Crest length, in feet.
nidDamHeightFt	Dam height, in feet.
nidMaxDischargeCfs	Maximum discharge in cubic feet per second.
nidFoundationType	Foundation type for the dam.
nidSpillwayType	Spillway type for the dam.
nidOutletWorks	Indicates if the dam has outlet works.
nidConditionAssessment	Overall condition assessment rating.
nidConditionAssessmentDate	Date of the most recent condition assessment.
nidConditionAssessmentId	Code for the condition assessment rating.
nidCountyFips	County FIPS code.
nidStateKey	Two letter abbreviation of the state.

Attribute	Definition
nidStateName	State name.
nidNation	Country code.
nidZipcode	Postal code.
nidHuc2	Hydrologic Unit Code two digit region.
nidHuc4	Hydrologic Unit Code four digit subregion.
nidHuc6	Hydrologic Unit Code six digit basin.
nidHuc8	Hydrologic Unit Code eight digit subbasin.
nidFemaRegion	FEMA region where the dam is located.
nidFemaCommunity	Name of the community participant in the National Flood Insurance Program that is local to the dam.
nidAiannh	Name of recognized American Indian, Alaska Native, or Native Hawaiian community areas where applicable.
nidEapLastExerciseDate	Date of the most recent exercise of the Emergency Action Plan.
nidEapLastExerciseType	Type of the most recent Emergency Action Plan exercise.
nidEapVerified	Indicates if the Emergency Action Plan is verified.
nidEapNextDueDate	Date the next Emergency Action Plan exercise is due.
nidEapSchedule	Frequency schedule for Emergency Action Plan exercises.
nidEapNotes	Notes related to Emergency Action Plan.
nidInspectionFrequencyNotes	Notes related to inspection frequency.
nidOwnerTypeId	Code for owner type.
nidRegulatoryAgencyId	Code for the regulatory agency.
nidPrimaryPurposeCode	Code for the primary purpose.
nidOtherPurposeCode	Code for other purposes.
nidOutletWorksId	Code for the presence of outlet works.
nidSpillwayTypeId	Code for the spillway type.
nidFoundationTypeId	Code for the foundation type.
nidConditionAssessmentCode	Code for the overall condition assessment.
nidEmergencyActionPlanCode	Code for Emergency Action Plan status.
nidInspectionFrequencyCode	Code for inspection frequency.
nidHazardCode	Code for hazard potential classification.
nidHazardText	Text for hazard potential classification.
nidSourceAgency	The source agency that submitted the data.
nidSubmittingAgency	The agency that transmitted data to the NID.
nidRecordStatus	Indicates if the record is current or archived.
nidDataLastUpdated	Date of the most recent data update.
nidUrl	Link to the official NID entry for the dam.
privateDamId	Code to indicate whether a dam is a private dam.
politicalPartyId	Code for political party that currently holds the Congressional District seat the dam is located.
huc2	Hydrologic Unit Code (HUC) two digit region.
huc4	Hydrologic Unit Code (HUC) four digit subregion.
huc6	Hydrologic Unit Code (HUC) six digit basin.
huc8	Hydrologic Unit Code (HUC) eight digit subbasin.
zipcode	Post Office Zip Codes.
nation	Code for country where dam is located.
stateKey	Two letter abbreviation of the state where dam is located.
femaRegion	Federal Emergency Management Agency (FEMA) Region where dam is located.
femaCommunity	Name of community participant in the National Flood Insurance Program (NFIP) local to the dam.

Attribute	Definition
aiannh	Name of recognized American Indian, Alaska Native, or Native Hawaiian community areas where applicable.

3 Dataset Preprocessing

All data preprocessing was done with python.

3.1 Data Injectivity Test

In this section, the dataset is examined to perform a *data injectivity test* across all attributes. The purpose of this test is to identify attributes that contain redundant or functionally dependent information. Data injectivity refers to the property that the values of one attribute uniquely determine the values of another. Formally, for two attributes a_1 and a_2, the attribute a_2 is said to be *injective with respect to* a_1 if and only if

$$\forall \text{ rows } r_i, r_j \quad r_i[a_j] = r_j[a_j] \implies r_i[a_i] = r_j[a_i].$$

This condition ensures that if two rows share the same value in attribute a_2, they must also share the same value in attribute a_1.

To conduct the data injectivity test, each attribute a_1 in the NTAD_Dams dataset is compared pairwise with all other attributes $a_2, a_3, \ldots, a_n$. For every pair (a_1, a_k), the analysis determines whether the values in a_k injectively map to those in a_1. If such a mapping exists, a_k is considered injective with respect to a_1, implying that one of the two attributes may be redundant or derivable from the other.

Identifying injective relationships serves several key purposes:

1. **Redundancy detection:** Reveals attributes that encode the same information in alternate forms, such as numerical identifiers and their corresponding textual labels.
2. **Functional dependence analysis:** Highlights attributes whose values are completely determined by other attributes, indicating dependency structures within the dataset.
3. **Dimensionality reduction:** Supports the removal of superfluous attributes without losing informational content, thereby simplifying the dataset for downstream modeling.

For example, if both `longitude` and `x` yield identical groupings of records, then `longitude` is injective with respect to `x`. This indicates that the two attributes convey the same underlying information. Detecting such injective pairs is therefore an important preprocessing step that ensures data minimality, reduces redundancy, and improves the interpretability of the predictive model.

3.1.1 Results of the Data Injectivity Test

This section summarizes attribute pairs that exhibit injective or bijective relationships, along with the resulting keep or remove decisions taken to reduce redundancy while preserving informational content.

Bijective relationships

- **0 (OBJECTID) and 1 (id) and 2 (federalId):** All three are mutually injective and therefore bijective.
  - *Decision:* Keep 0 (OBJECTID). Remove 1 (id) and 2 (federalId).
  - *Rationale:* These fields encode the same unique identifier so retaining one prevents loss of information.

One way injective relationships

- **99 (x) injective with respect to 5 (longitude), and 100 (y) injective with respect to 4 (latitude):** If two records share the same value in `x` then they share the same value in `longitude`, and similarly for `y` and `latitude`. The converse does not hold. This indicates that `x` and `y` are lower resolution or derived encodings of the primary coordinates.
  - *Decision:* Remove 99 (x) and 100 (y). Keep 5 (longitude) and 4 (latitude).
  - *Rationale:* The latitude and longitude attributes admit more distinct values and therefore carry at least as much information.
- **Hydrologic Unit Codes 89 (huc2), 90 (huc4), 91 (huc6), and 92 (huc8):** In practice, a given value of `huc8` uniquely determines the corresponding values of `huc6`, `huc4`, and `huc2`. Thus equality on `huc8` implies equality on the coarser codes. Your empirical test found consistent injectivity in this direction.
  - *Decision:* Keep 89 (huc2). Remove 90 (huc4), 91 (huc6), and 92 (huc8) for the nominal feature set.
  - *Rationale:* `huc4`, `huc6`, and `huc8` have very high cardinality which can complicate nominal modeling without regularization or target encoding. Retaining `huc2` preserves basin level signal with manageable cardinality.

Attributes with no detected injective ties

- **71 (secondaryLengthOfLocks)** and **73 (secondaryWidthOfLocks):** No clear injective relationships to other attributes were detected.
- **88 (politicalPartyId):** No injective connection to safety or structural attributes.
- **94 (nation):** No injective connection observed.

3.2 Deleting Irregular Rows

- *Row 39881* was deleted because the values along the row did not match up the attributes

3.3 Working with Weka

This project used both weka and python. For the dataset to be imported to weka, all entries with quotations and special characters (`\n` and `\r`) had to replaced with whitespace.

3.4 Unreasonable Attributes

Several attributes in the NTAD_Dams dataset were deemed *unreasonable* for inclusion in the predictive modeling process. These attributes were removed because they exhibit one or more of the following characteristics:

1. Extremely high nominal cardinality, often functioning as unique identifiers rather than informative predictors.
2. Irregular or incomplete categorical coverage, meaning that new dams may contain values not represented in the training dataset (out-of-vocabulary risk).
3. Administrative or textual metadata unrelated to structural, locational, or hazard-relevant properties.

Removed Attributes and Justification

- **OBJECTID:** Nominal attribute with 92,522 distinct values. Serves purely as a unique identifier with no predictive value.
- **Name:** Nominal attribute with 77,595 distinct values. Each dam name is unique, making it unsuitable for learning generalizable patterns.
- **ownerNames:** Nominal attribute with 53,818 distinct values. High variability and inconsistent naming conventions; unlikely to generalize to unseen ownership entities.
- **nidId:** Nominal attribute with 91,776 distinct values. Duplicates the role of `OBJECTID` and offers no additional information.
- **designerNames:** Nominal attribute with 6,545 distinct values. Sparse and unstandardized, often containing non-repeating free-text entries.
- **sourceAgency:** Nominal attribute indicating the submitting organization. Removed because the set of possible agencies is incomplete; a new submission source would not be represented in the training list.
- **stateFedId:** Nominal attribute with 62,050 distinct values. Serves as an administrative identifier, not a predictive feature.
- **County:** Nominal attribute with partial categorical coverage. If a new dam is located in a county not represented in the dataset, the model cannot assign a valid category, leading to generalization errors.
- **countyState:** Nominal attribute combining county and state. Suffers from the same generalization issue when unseen geographic combinations appear.
- **City:** Nominal attribute with irregular coverage. Cities absent from the training data would produce unseen categorical values at inference time.
- **riverName:** Nominal attribute representing river or stream names. High cardinality and inconsistent naming patterns (e.g., abbreviations, alternative spellings) make it unstable for modeling. New rivers would also be unrepresented.
- **congDist:** Nominal attribute for congressional districts. The mapping changes periodically due to redistricting; unseen districts would not map cleanly to the existing training categories.
- **stateRegulatoryAgency:** Nominal attribute with irregular text entries. Agencies not listed in the dataset would appear as unknown classes, making the attribute unreliable.
- **Zipcode:** Nominal attribute with 16,794 distinct values. Sparse, inconsistent, and region-specific; also prone to out-of-range inputs for new locations.
- **dateUpdated, inspectionDate, conditionAssessDate, yearCompleteId, yearCompletedId:** Nominal or date-like text attributes with nonstandard formats and missing values. New dams or future updates would produce values not observed in the training data.
- **inspectionFrequency:** Not suitable as a predictor because the inspection frequency depends on administrative policy, dam age, and regulatory jurisdiction rather than physical or hazard-related factors.
- **websiteUrl:** Nominal attribute containing web links. Not predictive and subject to missing or malformed URLs.
- **usaceDivision, usaceDistrict, femaCommunity:** Nominal administrative fields. Incomplete coverage and inconsistent naming conventions; new districts or communities would produce unrecognized categories.

- **Nation:** Removed because all records belong to the United States. Other detected values were irregular or erroneous.
- **stateKey:** Redundant with the existing `state` attribute.
- **outletGateTypes:** Deleted because of the abundent number of irregular values not listed in documentation. e.g `Slide (slice gate)6`, `Valve1`, Roller2 were never mentioned in documentation

Rationale for Exclusion

These attributes were excluded to prevent high-cardinality nominal features from overfitting the model and to avoid *out-of-vocabulary errors*, where a categorical value encountered during inference does not exist in the training set. Removing such fields improves model generalization, simplifies feature encoding, and ensures that retained attributes represent interpretable, stable, and domain-relevant information.

3.5 Missing Values (Column-Wise Analysis)

Attributes where 70% or more instances do not have said attribute are removed as high rates of missingness prevent us from reliably replacing those missing values.

- **otherNames:** 72% of entries missing.
- **formerNames:** 89% of entries missing.
- **otherStructureId:** 99% of entries missing.
- **fedOwnerIds:** 97% of entries missing.
- **fedFundingIds:** 85% of entries missing.
- **fedDesignIds:** 71% of entries missing.
- **fedConstructionIds:** 85% of entries missing.
- **fedRegulatoryIds:** 95% of entries missing.
- **fedInspectionIds:** 82% of entries missing.
- **fedOperationIds:** 97% of entries missing.
- **fedOtherIds:** 99% of entries missing.
- **yearModified:** 93% of entries missing.
- **secondaryLengthOfLocks:** 100% of entries missing.
- **secondaryWidthOfLocks:** 100% of entries missing.
- **eapLastRevDate:** 81% of entries missing.
- **operationalStatusId:** 81% of entries missing.
- **operationalStatusDate:** 82% of entries missing.
- **lastEapExcerDate:** 99% of entries missing.
- **politicalPartyId:** 100% of entries missing.
- **aiannh:** 94% of entries missing.

3.6 Missing Values (Row-Wise Analysis)

In addition to column-wise missingness, a row-level completeness check was performed to identify individual records with excessive missing data. Any record containing missing values for 70% or more were removed.

- **Row 39880:** Contained missing values in over 70% of all attributes.

3.7 Changing Datatypes

3.7.1 Converted String Attributes

Some of the attributes when imported to Weka was identified as the string datatype. These attributes had to be converted to either numeric or nominal to be able to use attribute selection algorithms.

- **Latitude:** Converted from `string` to `numeric`.
- **Longitude:** Converted from `string` to `numeric`.
- **primaryPurposeId:** Converted from `string` to `nominal`.
- **nidHeight:** Converted from `string` to `numeric`.
- **eapId:** Converted from `string` to `nominal`.
- **ownerTypeIds:** Converted from `string` to `nominal`.
- **nonFederalDamOnFederalId:** Converted from `string` to `nominal`.
- **Distance:** Converted from `string` to `numeric`.
- **fedRegulateId:** Converted from `string` to `nominal`.
- **jurisdictionAuthorityId:** Converted from `string` to `nominal`.

3.7.2 One-Hot Encoding

Some attributes had possible values separated by semi colons e.g a value of foundationTypeIds could be 1;2 which means it has category 1 and 2. For these kinds of attributes, one-hot encoding has to be done.

- **ownerTypeIds:** Encoded across 12 categories $[1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12]$.
- **purposeIds:** Encoded across 6 categories $[1, 2, 3, 4, 5, 6]$.
- **damTypeIds:** Encoded across 12 categories $[1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12]$.
- **coreTypeIds:** Encoded across 6 categories $[1, 2, 3, 4, 5, 6]$.
- **foundationTypeIds:** Encoded across 4 categories $[1, 2, 3, 4]$.

As a result of encoding, we will delete the parent attribute and replace it with many attributes which represent the encoding e.g attribute *purposeIds* of just 1 would be *purposeIds__1*

3.8 Dealing with Missing Values (Class)

All instances had a value for the class *hazardId*. Therefore, no instance had to be removed due to missing class attribute.

3.9 Dealing with Missing Values (Continuous, Nominal, Discrete)

- Missing value for **continuous data** was replaced using *mean*
- Missing values for **nominal data** was replaced using *mode*
- Missing values for **discrete data** was replaced using *median*

3.10 Obvious Attributes

Some attributes were removed from the dataset because they were found to be closely related to the target variable `hazardId`. Although these attributes are not identical to `hazardId`, they share strong conceptual and statistical relationships with it. Retaining them could introduce partial data leakage, leading the model to rely on correlated administrative indicators rather than independent structural or environmental factors.

- **eapId:** Represents the Emergency Action Plan (EAP) classification for each dam. Since EAPs are typically implemented or updated in proportion to the dam's hazard level, this attribute tends to correlate strongly with `hazardId`. Its inclusion could cause the model to infer hazard indirectly through administrative preparedness rather than physical risk indicators.
- **conditionAssessId:** Encodes the dam's latest condition assessment category. While not identical to hazard classification, this attribute reflects similar evaluative criteria—dams with poorer condition ratings often exhibit higher hazard potential. Retaining this attribute could therefore inflate the model's predictive accuracy by exploiting overlapping information.

Rationale for Removal

Although neither attribute is a direct duplicate of `hazardId`, both exhibit strong conceptual and empirical relationships with it. Removing them ensures that the predictive model focuses on independent explanatory variables, such as structural, environmental, and locational attributes, rather than correlated administrative assessments. This step reduces redundancy, improves model generalization, and prevents subtle forms of data leakage.

3.11 Data Discretization

Numeric data needed to be discretized in order to use certain classification algorithms. Discretization was done using the *equal frequency binning* method. The number of bins was determined based on how many distinct values the numeric attribute had:

- `latitude` had **86,308** unique values and was binned into **1,000** bins.
- `longitude` had **88,035** unique values and was binned into **1,000** bins.
- `nidHeight` had **384** unique values and was binned into **50** bins.
- `separateStructuresCount` had **15** unique values and was binned into **10** bins.
- `distance` had **458** unique values and was binned into **50** bins.
- `damHeight` had **373** unique values and was binned into **50** bins.
- `hydraulicHeight` had **325** unique values and was binned into **50** bins.
- `structuralHeight` had **346** unique values and was binned into **50** bins.
- `damLength` had **4,267** unique values and was binned into **100** bins.
- `volume` had **15,120** unique values and was binned into **100** bins.

- `nidStorage` had **11,248** unique values and was binned into **100** bins.
- `maxStorage` had **11,045** unique values and was binned into **100** bins.
- `normalStorage` had **7,690** unique values and was binned into **100** bins.
- `surfaceArea` had **4,672** unique values and was binned into **100** bins.
- `drainageArea` had **4,880** unique values and was binned into **100** bins.
- `maxDischarge` had **8,049** unique values and was binned into **100** bins.
- `spillwayWidth` had **1,455** unique values and was binned into **100** bins.
- `numberOfLocks` had **11** unique values and was binned into **10** bins.
- `lengthOfLocks` had **38** unique values and was binned into **10** bins.
- `widthOfLocks` had **22** unique values and was binned into **10** bins.

4 Attribute Selection

Attribute selection aims to reduce the number of input variables while retaining the most relevant information for prediction. By removing redundant or irrelevant attributes, the resulting models are simpler, faster, and often more accurate. In this project, four evaluation methods were applied in Weka: *CfsSubsetEval*, *CorrelationAttributeEval*, *GainRatioAttributeEval*, and *InfoGainAttributeEval*. Each method ranks or selects attributes based on different statistical criteria.

4.1 CfsSubsetEval

CfsSubsetEval (Correlation-based Feature Selection) evaluates subsets of attributes rather than individual ones. It selects groups of features that are highly correlated with the class but have low intercorrelation with each other. This helps eliminate redundant attributes that provide overlapping information, improving the model's generalization and reducing overfitting.

From the Weka outputs in Appendix A.1, we get the chosen attributes to be:

- primaryPurposeId
- ownerTypeIds__6
- primaryOwnerTypeId
- purposeIds__8
- purposeIds__10
- state
- foundationTypeIds__1
- damHeight
- nidHeightId
- volume
- nidStorage
- normalStorage
- maxDischarge
- huc2
- femaRegion

4.2 CorrelationAttributeEval

CorrelationAttributeEval evaluates each attribute individually by measuring its correlation with the target class. A high correlation indicates that the attribute is strongly predictive of the class label. Attributes with low correlation values contribute less useful information and can be removed to simplify the model.

Setting the threshold to 0.1 from the Weka outputs in Appendix A.2, we get:

- purposeIds__10
- ownerTypeIds__1
- foundationTypeIds__1
- purposeIds__8
- stateRegulatedId
- primaryOwnerTypeId
- purposeIds__5
- fedRegulatedId
- ownerTypeIds__5
- permittingAuthorityId
- enforcementAuthorityId
- femaRegion
- foundationTypeIds__3

4.3 GainRatioAttributeEval

GainRatioAttributeEval is based on information theory and evaluates attributes using the Gain Ratio metric. It measures how much information about the class is gained by knowing the value of an attribute, normalized to prevent bias toward attributes with many distinct values. This method is particularly effective for categorical data and is commonly used in decision tree algorithms.

Setting the threshold to 0.08 from the Weka outputs in Appendix A.3, we get:

- ownerTypeIds__6
- foundationTypeIds__2
- damTypeIds__4
- widthOfLocks
- lengthOfLocks
- damTypeIds__1
- isAssociatedStructureId
- foundationTypeIds__1
- coreTypeIds__1
- damTypeIds__7
- purposeIds__3

4.4 InfoGainAttributeEval

InfoGainAttributeEval (Information Gain) evaluates each attribute individually by calculating how much it reduces the uncertainty (entropy) of the class. Attributes that provide greater information gain are considered more useful for predicting the target variable. Unlike Gain Ratio, it does not normalize the result, so attributes with many unique values may receive higher scores.

Setting the threshold to 0.08 from the Weka outputs in Appendix A.4, we get:

- state
- longitude
- huc2
- nidStorage
- maxStorage
- femaRegion
- latitude
- damHeight
- normalStorage
- nidHeight
- primaryPurposeId
- hydraulicHeight
- nidHeightId
- structuralHeight
- volume
- maxDischarge

4.5 SelfSelected

- primaryPurposeId
- damHeight
- state
- surfaceArea
- damLength
- volume

5 Model Selection

5.1 OneR

The *OneR* (One Rule) algorithm builds a simple rule-based model that uses only a single attribute to make predictions. It evaluates each attribute individually and selects the one that yields the lowest error rate. Although simple, OneR often provides a good baseline and highlights which individual feature is most predictive.

5.2 Naive Bayes

Naive Bayes is a probabilistic classifier based on Bayes' Theorem, assuming that all attributes are independent given the class label. It calculates the probability of each class for a given instance and predicts the class with the highest probability. It is efficient and works well even with relatively small datasets.

5.3 J48

J48 is Weka's implementation of the C4.5 decision tree algorithm. It recursively splits data based on attributes that provide the highest information gain, forming a tree where each path represents a classification rule. It handles both categorical and numerical data and includes pruning to prevent overfitting.

5.4 BayesNet

BayesNet (Bayesian Network) is a graphical probabilistic model that represents dependencies between attributes using a directed acyclic graph. Unlike Naive Bayes, it learns conditional dependencies among variables, allowing it to capture more complex relationships in the data while still producing probabilistic predictions.

5.5 Models

The following are all the models used for this study in the format of
`[Classification model] + [Attribute Selection Method]`

- OneR
  - OneR + CfsSubsetEval
  - OneR + CorrelationAttributeEval
  - OneR + GainRatioAttributeEval
  - OneR + InfoGainAttributeEval
  - OneR + SelfSelected
- NaiveBayes
  - NaiveBayes + CfsSubsetEval
  - NaiveBayes + CorrelationAttributeEval
  - NaiveBayes + GainRatioAttributeEval

  - NaiveBayes + InfoGainAttributeEval
  - NaiveBayes + SelfSelected
- J48
  - J48 + CfsSubsetEval
  - J48 + CorrelationAttributeEval
  - J48 + GainRatioAttributeEval
  - J48 + InfoGainAttributeEval
  - J48 + SelfSelected
- BayesNet
  - BayesNet + CfsSubsetEval
  - BayesNet + CorrelationAttributeEval
  - BayesNet + GainRatioAttributeEval
  - BayesNet + InfoGainAttributeEval
  - BayesNet + SelfSelected

The summary results of each model from weka can be found in Appendix B.

6 Results

Aggregating results from Appendix B, we get the following tables.

Table 2: Training Accuracy Comparison Across Models

Classifier	CfsSubsetEval	Correlation	GainRatio	InfoGain	SelfSelected
BayesNet	65.17	65.15	67.47	65.47	68.04
J48	78.62	71.71	67.53	76.79	74.84
NaiveBayes	65.19	65.15	67.47	65.26	68.08
OneR	70.46	66.17	66.17	70.46	70.46

Table 3: Testing Accuracy Comparison Across Models

Classifier	CfsSubsetEval	Correlation	GainRatio	InfoGain	SelfSelected
BayesNet	65.13	64.65	66.88	65.03	67.90
J48	75.75	70.37	66.95	74.15	72.97
NaiveBayes	65.15	64.65	66.88	64.95	67.97
OneR	70.54	65.87	65.87	70.54	70.54

Table 4: Testing TP Rate Comparison Across Models

Classifier	CfsSubsetEval	Correlation	GainRatio	InfoGain	SelfSelected
BayesNet	0.651	0.647	0.669	0.650	0.679
J48	0.757	0.704	0.669	0.742	0.730
NaiveBayes	0.651	0.647	0.669	0.650	0.680
OneR	0.705	0.659	0.659	0.705	0.705

Table 5: Testing FP Rate Comparison Across Models

Classifier	CfsSubsetEval	Correlation	GainRatio	InfoGain	SelfSelected
BayesNet	0.190	0.333	0.543	0.188	0.252
J48	0.305	0.382	0.544	0.330	0.358
NaiveBayes	0.190	0.333	0.543	0.187	0.252
OneR	0.394	0.593	0.593	0.394	0.394

Table 6: Testing ROC Area Comparison Across Models

Classifier	CfsSubsetEval	Correlation	GainRatio	InfoGain	SelfSelected
BayesNet	0.835	0.757	0.567	0.832	0.824
J48	0.841	0.787	0.567	0.819	0.810
NaiveBayes	0.835	0.757	0.567	0.832	0.824
OneR	0.656	0.533	0.533	0.656	0.656

6.1 Overfitting and Underfitting

Analyzing Table 2 and Table 3, we do not see any clear signs of overfitting. Generally, it appears to be the case that models with higher training data also have higher testing accuracy, which is a sign against overfitting. In addition, none of the training accuracies are signifcantly high (over 90%). Looking at the relatively low accuracies (<= 70%) of models using *BayesNet*, *NaiveBayes* , and *OneR* classifiers, we see some signs of underfitting. The only classifier whose models' are performing well in terms of accuracy seems to be *J48* classifier models.

6.2 TP,FP Rates, and ROC

From Table 4 we see that the best performing model is *J48+CfsSubsetEval*. In addition, we also see that *J48* performed better across all attribute selection algorithms in contrast to other classification methods. The high performence of *J48* is further supported by Table 6 as models with *J48* has some of the highest performences (except for *J48+GainRatio*). However, one downside of *J48* we see in Table 5 is that *J48* models seem to have higher FPR rates when compared to *BayesNet* and *NaiveBayes*. Looking further into the performence of *J48+CfsSubsetEval*, our best model yet, in Appendix B.3.1, we see that the FPR rate of `val_1` in both training and testing were the lowest. In addition, observing the confusion matrix for `val_2`, the attribute with the highest FPR, we see that most instances with label `val_2` are being classified correctly or being classified as `val_3` or `val_4`. Referring back to the definition of *hazardId* in Section 2, this means `val_2`, the lowest hazard potential, is mostly being categorized as low hazard potential or of higher hazard potential. Since low hazards are being predicted mostly as low hazards or higher hazards, the high FPR rate is not a significant detriment to the model, as it is better to be cautious by predicting as higher than predicting as lower.

7 Discussion and Conclusion

In conclusion, the best-performing model was *J48+CfsSubsetEval*. This model achieved the highest training and testing accuracy, along with strong true positive (TPR) and ROC values. Although it exhibited a relatively high false positive rate (FPR), its tendency to overpredict higher hazard levels is acceptable for this context, as it is safer to flag potential hazards than to overlook them. Overall, the model balanced accuracy, generalization, and interpretability effectively, making it well-suited for identifying hazard risk levels.

Working independently on this project has taught me how to organize and manage a complete machine learning workflow from data preprocessing to model evaluation. I became more comfortable

using Weka and learned how different feature selection methods and classification algorithms affect model performance. I also gained practical experience interpreting machine learning metrics such as confusion matrices, precision, recall, TP and FP rates, and ROC curves. Understanding these metrics helped me evaluate not just how accurate a model was, but how reliable and sensitive it was to detecting the correct hazard categories. Through this project, I learned how to analyze results critically and make informed decisions about model quality and suitability for real-world applications.

References

U.S. Department of Transportation, B. o. T. S. (2025). *Dams – dataset (national transportation atlas database, ntad).* https://geodata.bts.gov/datasets/usdot::dams/about. (Accessed: 2025-10-21)

8 Appendix

A Attribute Selection (Weka Outputs)

A.1 CfsSubsetEval

```
=== Attribute Selection on all input data ===

Search Method:
       Best first.
       Start set: no attributes
       Search direction: forward
       Stale search after 5 node expansions
       Total number of subsets evaluated: 1399
       Merit of best subset found: 0.139

Attribute Subset Evaluator (supervised, Class (nominal): 80 hazardId):
       CFS Subset Evaluator
       Including locally predictive attributes

Selected attributes: 3,10,11,22,24,27,56,60,63,65,66,68,71,78,79 : 15
                    primaryPurposeId
                    ownerTypeIds__6
                    primaryOwnerTypeId
                    purposeIds__8
                    purposeIds__10
                    state
                    foundationTypeIds__1
                    damHeight
                    nidHeightId
                    volume
                    nidStorage
                    normalStorage
                    maxDischarge
                    huc2
                    femaRegion
```

A.2 CorrelationAttributeEval

```
=== Attribute Selection on all input data ===

Search Method:
       Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 80 hazardId):
```

```
       Correlation Ranking Filter
Ranked attributes:
 0.19948   24 purposeIds__10
 0.16154    5 ownerTypeIds__1
 0.14381   56 foundationTypeIds__1
 0.14223   22 purposeIds__8
 0.14018   29 stateRegulatedId
 0.13219   11 primaryOwnerTypeId
 0.124     19 purposeIds__5
 0.10939   30 fedRegulatedId
 0.10586    9 ownerTypeIds__5
 0.10508   32 permittingAuthorityId
 0.10507   34 enforcementAuthorityId
 0.10274   79 femaRegion
 0.1015    58 foundationTypeIds__3
 0.09736   78 huc2
 0.09731   31 jurisdictionAuthorityId
 0.09545   55 coreTypeIds__6
 0.09543   37 primaryDamTypeId
 0.09248    6 ownerTypeIds__2
 0.08889   63 nidHeightId
 0.08697   48 damTypeIds__11
 0.08616   10 ownerTypeIds__6
 0.08494   20 purposeIds__6
 0.08402   44 damTypeIds__7
 0.08381    3 primaryPurposeId
 0.08131   65 volume
 0.07722   13 isAssociatedStructureId
 0.07374   59 foundationTypeIds__4
 0.06817   15 purposeIds__1
 0.0672    51 coreTypeIds__2
 0.06684   61 hydraulicHeight
 0.06545   27 state
 0.06448   52 coreTypeIds__3
 0.06293   72 spillwayTypeId
 0.06242   25 purposeIds__11
 0.06217   33 inspectionAuthorityId
 0.06037   46 damTypeIds__9
 0.05846    7 ownerTypeIds__3
 0.05659   17 purposeIds__3
 0.05441   70 drainageArea
 0.05207   42 damTypeIds__5
 0.05068   62 structuralHeight
 0.04974   38 damTypeIds__1
 0.04892   18 purposeIds__4
 0.04871   60 damHeight
 0.04331   76 widthOfLocks
 0.04245   75 lengthOfLocks
 0.03921   49 damTypeIds__12
 0.03849   35 secretaryAgricultureBuiltId
 0.03769   28 distance
 0.03673   36 nrcsWatershedAuthorizationId
 0.03518   71 maxDischarge
 0.03382   68 normalStorage
 0.03331   69 surfaceArea
 0.03273    4 nidHeight
 0.03147   67 maxStorage
 0.03047   26 purposeIds__12
 0.02958   66 nidStorage
 0.02928    8 ownerTypeIds__4
 0.0287    23 purposeIds__9
 0.02789   40 damTypeIds__3
 0.02514   41 damTypeIds__4
 0.02245   21 purposeIds__7
 0.02201   73 spillwayWidth
 0.0207    54 coreTypeIds__5
 0.01927   43 damTypeIds__6
 0.01925   64 damLength
```

```
 0.01836   57 foundationTypeIds__2
 0.01527   45 damTypeIds__8
 0.01524   39 damTypeIds__2
 0.01482   50 coreTypeIds__1
 0.01126   16 purposeIds__2
 0.01056    2 longitude
 0.00993   53 coreTypeIds__4
 0.00798   47 damTypeIds__10
 0.00639    1 latitude
 0.00533   14 nonFederalDamOnFederalId
 0.0052    77 privateDamId
 0         74 numberOfLocks
 0         12 separateStructuresCount

Selected attributes:
24,5,56,22,29,11,19,30,9,32,
34,79,58,78,31,55,37,6,63,48,
10,20,44,3,65,13,59,15,51,61,27,
52,72,25,33,46,7,17,70,42,62,38,
18,60,76,75,49,35,28,36,71,68,69,
4,67,26,66,8,23,40,41,21,73,54,
43,64,57,45,39,50,16,2,53,47,1,14,77,74,12 : 79
```

A.3 GainRatioAttributeEval

```
=== Attribute Selection on all input data ===

Search Method:
       Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 80 hazardId):
       Gain Ratio feature evaluator

Ranked attributes:
 0.19009   10 ownerTypeIds__6
 0.17827   57 foundationTypeIds__2
 0.14918   41 damTypeIds__4
 0.14019   76 widthOfLocks
 0.13088   75 lengthOfLocks
 0.09654   38 damTypeIds__1
 0.09366   13 isAssociatedStructureId
 0.0924    56 foundationTypeIds__1
 0.0921    50 coreTypeIds__1
 0.09185   44 damTypeIds__7
 0.08104   17 purposeIds__3
 0.07924    6 ownerTypeIds__2
 0.07435   24 purposeIds__10
 0.06903   63 nidHeightId
 0.06611   39 damTypeIds__2
 0.06589   27 state
 0.06566   30 fedRegulatedId
 0.06542   22 purposeIds__8
 0.06002   20 purposeIds__6
 0.05485   51 coreTypeIds__2
 0.05312   78 huc2
 0.05251   25 purposeIds__11
 0.05188   79 femaRegion
 0.0462    11 primaryOwnerTypeId
 0.0456    15 purposeIds__1
 0.04241   48 damTypeIds__11
 0.04043    3 primaryPurposeId
 0.03963   42 damTypeIds__5
 0.03665    5 ownerTypeIds__1
 0.03459   65 volume
 0.03214    7 ownerTypeIds__3
 0.03021    9 ownerTypeIds__5
```

```
 0.03017   37 primaryDamTypeId
 0.02763   61 hydraulicHeight
 0.02736   40 damTypeIds__3
 0.02731   60 damHeight
 0.02703   62 structuralHeight
 0.026     66 nidStorage
 0.02499    2 longitude
 0.0248    58 foundationTypeIds__3
 0.02458   29 stateRegulatedId
 0.02412   67 maxStorage
 0.024     71 maxDischarge
 0.02362   54 coreTypeIds__5
 0.02222    4 nidHeight
 0.02208   53 coreTypeIds__4
 0.02067   46 damTypeIds__9
 0.01995   49 damTypeIds__12
 0.01954   68 normalStorage
 0.01922   23 purposeIds__9
 0.01904   35 secretaryAgricultureBuiltId
 0.0187    73 spillwayWidth
 0.01777   19 purposeIds__5
 0.01763   55 coreTypeIds__6
 0.01752   52 coreTypeIds__3
 0.01663   72 spillwayTypeId
 0.01588   36 nrcsWatershedAuthorizationId
 0.01561   47 damTypeIds__10
 0.01545   34 enforcementAuthorityId
 0.01544   32 permittingAuthorityId
 0.01478   70 drainageArea
 0.01453   69 surfaceArea
 0.01424   28 distance
 0.01422   31 jurisdictionAuthorityId
 0.01386   59 foundationTypeIds__4
 0.01381   43 damTypeIds__6
 0.01373    1 latitude
 0.01206   64 damLength
 0.01014    8 ownerTypeIds__4
 0.00886   18 purposeIds__4
 0.00807   21 purposeIds__7
 0.00745   33 inspectionAuthorityId
 0.00329   77 privateDamId
 0.00328   14 nonFederalDamOnFederalId
 0.00291   26 purposeIds__12
 0.00195   45 damTypeIds__8
 0.00192   16 purposeIds__2
 0         74 numberOfLocks
 0         12 separateStructuresCount

Selected attributes:
10,57,41,76,75,38,13,56,50,
44,17,6,24,63,39,27,30,22,20,
51,78,25,79,11,15,48,3,42,5,
65,7,9,37,61,40,60,62,66,2,58,
29,67,71,54,4,53,46,49,68,23,35,
73,19,55,52,72,36,47,34,32,70,
69,28,31,59,43,1,64,8,18,21,33,
77,14,26,45,16,74,12 : 79
```

A.4 InfoGainAttributeEval

```
=== Attribute Selection on all input data ===

Search Method:
       Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 80 hazardId):
```

```
       Information Gain Ranking Filter

Ranked attributes:
 0.3348     27 state
 0.248928    2 longitude
 0.184329   78 huc2
 0.170173   66 nidStorage
 0.156733   67 maxStorage
 0.137675   79 femaRegion
 0.136745    1 latitude
 0.13655    60 damHeight
 0.1228     68 normalStorage
 0.114066    4 nidHeight
 0.108365    3 primaryPurposeId
 0.10508    61 hydraulicHeight
 0.09709    63 nidHeightId
 0.096398   62 structuralHeight
 0.095574   65 volume
 0.087279   71 maxDischarge
 0.076368   69 surfaceArea
 0.072335   70 drainageArea
 0.071123   11 primaryOwnerTypeId
 0.07078    64 damLength
 0.053457   73 spillwayWidth
 0.048989   24 purposeIds__10
 0.043871   28 distance
 0.034076    5 ownerTypeIds__1
 0.026192   56 foundationTypeIds__1
 0.026134   22 purposeIds__8
 0.024034   10 ownerTypeIds__6
 0.02205     9 ownerTypeIds__5
 0.021433   29 stateRegulatedId
 0.020096   58 foundationTypeIds__3
 0.019958   30 fedRegulatedId
 0.017605   37 primaryDamTypeId
 0.017417   19 purposeIds__5
 0.014951   55 coreTypeIds__6
 0.014276   72 spillwayTypeId
 0.013578   23 purposeIds__9
 0.013071   52 coreTypeIds__3
 0.012559   34 enforcementAuthorityId
 0.012552   32 permittingAuthorityId
 0.011371   31 jurisdictionAuthorityId
 0.011083   36 nrcsWatershedAuthorizationId
 0.010933   20 purposeIds__6
 0.010534   59 foundationTypeIds__4
 0.010465    6 ownerTypeIds__2
 0.010416   35 secretaryAgricultureBuiltId
 0.009453   44 damTypeIds__7
 0.009091   48 damTypeIds__11
 0.007925   13 isAssociatedStructureId
 0.007406    7 ownerTypeIds__3
 0.005864   51 coreTypeIds__2
 0.005611   46 damTypeIds__9
 0.005527   33 inspectionAuthorityId
 0.005417   25 purposeIds__11
 0.005007   15 purposeIds__1
 0.004029   42 damTypeIds__5
 0.003806    8 ownerTypeIds__4
 0.003571   17 purposeIds__3
 0.003157   38 damTypeIds__1
 0.0029     76 widthOfLocks
 0.00288    18 purposeIds__4
 0.002803   75 lengthOfLocks
 0.001718   49 damTypeIds__12
 0.001335   26 purposeIds__12
 0.001287   40 damTypeIds__3
 0.001128   45 damTypeIds__8
```

```
 0.000911   16 purposeIds__2
 0.000885   21 purposeIds__7
 0.000859   41 damTypeIds__4
 0.000639   54 coreTypeIds__5
 0.00047    77 privateDamId
 0.000469   14 nonFederalDamOnFederalId
 0.000466   57 foundationTypeIds__2
 0.000396   43 damTypeIds__6
 0.000357   47 damTypeIds__10
 0.0003     39 damTypeIds__2
 0.000241   50 coreTypeIds__1
 0.000188   53 coreTypeIds__4
 0          12 separateStructuresCount
 0          74 numberOfLocks

Selected attributes:
27,2,78,66,67,79,1,60,68,4,
3,61,63,62,65,71,69,70,11,64,
73,24,28,5,56,22,10,9,29,58,
30,37,19,55,72,23,52,34,32,
31,36,20,59,6,35,44,48,13,7,
51,46,33,25,15,42,8,17,38,76,
18,75,49,26,40,45,16,21,41,54,
77,14,57,43,47,39,50,53,12,74 : 79
```

B Model Selection (Weka Outputs)

B.1 OneR

B.1.1 OneR + CfsSubsetEval

Training:

```
=== Summary ===

Correctly Classified Instances   52152               70.4623 %
Incorrectly Classified Instances 21862               29.5377 %
Kappa statistic                      0.334
Mean absolute error                  0.1477
Root mean squared error              0.3843
Relative absolute error             56.3439 %
Root relative squared error        106.1557 %
Total Number of Instances        74014

=== Detailed Accuracy By Class ===

               TP Rate FP Rate Precision Recall  F-Measure MCC     ROC Area PRC Area Class
               0.279   0.056   0.526     0.279   0.365     0.292   0.612    0.278    val_4
               0.926   0.597   0.745     0.926   0.826     0.401   0.664    0.738    val_2
               0.441   0.011   0.658     0.441   0.528     0.522   0.715    0.315    val_1
               0.245   0.037   0.482     0.245   0.325     0.283   0.604    0.210    val_3
Weighted Avg.  0.705   0.405   0.669     0.705   0.668     0.372   0.650    0.571

=== Confusion Matrix ===

    a     b     c    d   <-- classified as
 3754  8093   345  1246 |    a = val_4
 2236 44757   249  1069 |    b = val_2
   31  1731  1434    58 |    c = val_1
 1121  5532   151  2207 |    d = val_3
```

Testing:

```
=== Summary ===

Correctly Classified Instances   13053               70.5415 %
```

```
Incorrectly Classified Instances  5451             29.4585 %
Kappa statistic                      0.344
Mean absolute error                  0.1473
Root mean squared error              0.3838
Relative absolute error             56.0616 %
Root relative squared error        105.766 %
Total Number of Instances        18504

=== Detailed Accuracy By Class ===

               TP Rate FP Rate Precision Recall  F-Measure MCC     ROC Area PRC Area Class
               0.279   0.057   0.526     0.279   0.365     0.290   0.611    0.279    val_4
               0.926   0.581   0.747     0.926   0.827     0.415   0.672    0.740    val_2
               0.450   0.011   0.638     0.450   0.528     0.519   0.719    0.309    val_1
               0.271   0.039   0.498     0.271   0.351     0.304   0.616    0.227    val_3
Weighted Avg.  0.705   0.394   0.671     0.705   0.670     0.383   0.656    0.573

=== Confusion Matrix ===

    a    b    c    d   <-- classified as
  947 2026   93  328 |    a = val_4
  534 11135  66  288 |    b = val_2
    6  391  340   19 |    c = val_1
  315 1351   34  631 |    d = val_3
```

B.1.2 OneR + CorrelationAttributeEval

Training:

```
=== Summary ===

Correctly Classified Instances   48975             66.1699 %
Incorrectly Classified Instances 25039             33.8301 %
Kappa statistic                      0.0869
Mean absolute error                  0.1692
Root mean squared error              0.4113
Relative absolute error             64.5318 %
Root relative squared error        113.6075 %
Total Number of Instances        74014

=== Detailed Accuracy By Class ===

               TP Rate FP Rate Precision Recall  F-Measure MCC     ROC Area PRC Area Class
               0.137   0.030   0.505     0.137   0.216     0.191   0.554    0.226    val_4
               0.976   0.904   0.670     0.976   0.794     0.157   0.536    0.669    val_2
               0.000   0.000   ?         0.000   ?         ?       0.500    0.044    val_1
               0.000   0.000   ?         0.000   ?         ?       0.500    0.122    val_3
Weighted Avg.  0.662   0.595   ?         0.662   ?         ?       0.533    0.495

=== Confusion Matrix ===

    a    b    c    d   <-- classified as
 1845 11593   0    0 |    a = val_4
 1181 47130   0    0 |    b = val_2
    1  3253   0    0 |    c = val_1
  623  8388   0    0 |    d = val_3
```

Testing:

```
=== Summary ===

Correctly Classified Instances   12188             65.8668 %
Incorrectly Classified Instances  6316             34.1332 %
Kappa statistic                      0.0853
Mean absolute error                  0.1707
Root mean squared error              0.4131
Relative absolute error             64.9578 %
```

```
Root relative squared error        113.849 %
Total Number of Instances        18504

=== Detailed Accuracy By Class ===

               TP Rate FP Rate Precision Recall  F-Measure MCC     ROC Area PRC Area Class
               0.134   0.030   0.501     0.134   0.211     0.186   0.552    0.226    val_4
               0.976   0.904   0.667     0.976   0.792     0.158   0.536    0.666    val_2
               0.000   0.000   ?         0.000   ?         ?       0.500    0.041    val_1
               0.000   0.000   ?         0.000   ?         ?       0.500    0.126    val_3
Weighted Avg.  0.659   0.593   ?         0.659   ?         ?       0.533    0.492

=== Confusion Matrix ===

    a     b     c     d   <-- classified as
  455  2939     0     0 |     a = val_4
  290 11733     0     0 |     b = val_2
    0   756     0     0 |     c = val_1
  164  2167     0     0 |     d = val_3
```

B.1.3 OneR + GainRatioAttributeEval

Training:

```
=== Summary ===

Correctly Classified Instances   48975               66.1699 %
Incorrectly Classified Instances 25039               33.8301 %
Kappa statistic                      0.0869
Mean absolute error                  0.1692
Root mean squared error              0.4113
Relative absolute error             64.5318 %
Root relative squared error        113.6075 %
Total Number of Instances        74014

=== Detailed Accuracy By Class ===

               TP Rate FP Rate Precision Recall  F-Measure MCC     ROC Area PRC Area Class
               0.137   0.030   0.505     0.137   0.216     0.191   0.554    0.226    val_4
               0.976   0.904   0.670     0.976   0.794     0.157   0.536    0.669    val_2
               0.000   0.000   ?         0.000   ?         ?       0.500    0.044    val_1
               0.000   0.000   ?         0.000   ?         ?       0.500    0.122    val_3
Weighted Avg.  0.662   0.595   ?         0.662   ?         ?       0.533    0.495

=== Confusion Matrix ===

    a     b     c     d   <-- classified as
 1845 11593     0     0 |     a = val_4
 1181 47130     0     0 |     b = val_2
    1  3253     0     0 |     c = val_1
  623  8388     0     0 |     d = val_3
```

Testing:

```
=== Summary ===

Correctly Classified Instances   12188               65.8668 %
Incorrectly Classified Instances  6316               34.1332 %
Kappa statistic                      0.0853
Mean absolute error                  0.1707
Root mean squared error              0.4131
Relative absolute error             64.9578 %
Root relative squared error        113.849 %
Total Number of Instances        18504

=== Detailed Accuracy By Class ===
```

```
               TP Rate FP Rate Precision Recall  F-Measure MCC     ROC Area PRC Area Class
               0.134   0.030   0.501     0.134   0.211     0.186   0.552    0.226    val_4
               0.976   0.904   0.667     0.976   0.792     0.158   0.536    0.666    val_2
               0.000   0.000   ?         0.000   ?         ?       0.500    0.041    val_1
               0.000   0.000   ?         0.000   ?         ?       0.500    0.126    val_3
Weighted Avg.  0.659   0.593   ?         0.659   ?         ?       0.533    0.492

=== Confusion Matrix ===

    a     b    c    d   <-- classified as
  455  2939    0    0 |    a = val_4
  290 11733    0    0 |    b = val_2
    0   756    0    0 |    c = val_1
  164  2167    0    0 |    d = val_3
```

B.1.4 OneR + InfoGainAttributeEval

Training:

```
=== Summary ===

Correctly Classified Instances   52152               70.4623 %
Incorrectly Classified Instances 21862               29.5377 %
Kappa statistic                      0.334
Mean absolute error                  0.1477
Root mean squared error              0.3843
Relative absolute error             56.3439 %
Root relative squared error        106.1557 %
Total Number of Instances        74014

=== Detailed Accuracy By Class ===

               TP Rate FP Rate Precision Recall  F-Measure MCC     ROC Area PRC Area Class
               0.279   0.056   0.526     0.279   0.365     0.292   0.612    0.278    val_4
               0.926   0.597   0.745     0.926   0.826     0.401   0.664    0.738    val_2
               0.441   0.011   0.658     0.441   0.528     0.522   0.715    0.315    val_1
               0.245   0.037   0.482     0.245   0.325     0.283   0.604    0.210    val_3
Weighted Avg.  0.705   0.405   0.669     0.705   0.668     0.372   0.650    0.571

=== Confusion Matrix ===

    a     b     c    d   <-- classified as
 3754  8093   345  1246 |    a = val_4
 2236 44757   249  1069 |    b = val_2
   31  1731  1434    58 |    c = val_1
 1121  5532   151  2207 |    d = val_3
```

Testing:

```
=== Summary ===

Correctly Classified Instances   13053               70.5415 %
Incorrectly Classified Instances  5451               29.4585 %
Kappa statistic                      0.344
Mean absolute error                  0.1473
Root mean squared error              0.3838
Relative absolute error             56.0616 %
Root relative squared error        105.766 %
Total Number of Instances        18504

=== Detailed Accuracy By Class ===

               TP Rate FP Rate Precision Recall  F-Measure MCC     ROC Area PRC Area Class
               0.279   0.057   0.526     0.279   0.365     0.290   0.611    0.279    val_4
               0.926   0.581   0.747     0.926   0.827     0.415   0.672    0.740    val_2
               0.450   0.011   0.638     0.450   0.528     0.519   0.719    0.309    val_1
               0.271   0.039   0.498     0.271   0.351     0.304   0.616    0.227    val_3
```

```
Weighted Avg.  0.705   0.394   0.671     0.705   0.670     0.383   0.656    0.573

=== Confusion Matrix ===

    a     b     c     d   <-- classified as
  947  2026    93   328 |    a = val_4
  534 11135    66   288 |    b = val_2
    6   391   340    19 |    c = val_1
  315  1351    34   631 |    d = val_3
```

B.1.5 OneR + SelfSelected

Training:

```
=== Summary ===

Correctly Classified Instances   52152               70.4623 %
Incorrectly Classified Instances 21862               29.5377 %
Kappa statistic                      0.334
Mean absolute error                  0.1477
Root mean squared error              0.3843
Relative absolute error             56.3439 %
Root relative squared error        106.1557 %
Total Number of Instances        74014

=== Detailed Accuracy By Class ===

               TP Rate FP Rate Precision Recall  F-Measure MCC     ROC Area PRC Area Class
               0.279   0.056   0.526     0.279   0.365     0.292   0.612    0.278    val_4
               0.926   0.597   0.745     0.926   0.826     0.401   0.664    0.738    val_2
               0.441   0.011   0.658     0.441   0.528     0.522   0.715    0.315    val_1
               0.245   0.037   0.482     0.245   0.325     0.283   0.604    0.210    val_3
Weighted Avg.  0.705   0.405   0.669     0.705   0.668     0.372   0.650    0.571

=== Confusion Matrix ===

    a     b     c     d   <-- classified as
 3754  8093   345  1246 |    a = val_4
 2236 44757   249  1069 |    b = val_2
   31  1731  1434    58 |    c = val_1
 1121  5532   151  2207 |    d = val_3
```

Testing:

```
=== Summary ===

Correctly Classified Instances   13053               70.5415 %
Incorrectly Classified Instances  5451               29.4585 %
Kappa statistic                      0.344
Mean absolute error                  0.1473
Root mean squared error              0.3838
Relative absolute error             56.0616 %
Root relative squared error        105.766 %
Total Number of Instances        18504

=== Detailed Accuracy By Class ===

               TP Rate FP Rate Precision Recall  F-Measure MCC     ROC Area PRC Area Class
               0.279   0.057   0.526     0.279   0.365     0.290   0.611    0.279    val_4
               0.926   0.581   0.747     0.926   0.827     0.415   0.672    0.740    val_2
               0.450   0.011   0.638     0.450   0.528     0.519   0.719    0.309    val_1
               0.271   0.039   0.498     0.271   0.351     0.304   0.616    0.227    val_3
Weighted Avg.  0.705   0.394   0.671     0.705   0.670     0.383   0.656    0.573

=== Confusion Matrix ===

    a     b     c     d   <-- classified as
```

```
  947  2026   93   328 |    a = val_4
  534 11135   66   288 |    b = val_2
    6   391  340    19 |    c = val_1
  315  1351   34   631 |    d = val_3
```

B.2 NaiveBayes

B.2.1 NaiveBayes + CfsSubsetEval

Training:

```
=== Summary ===

Correctly Classified Instances   48246               65.185 %
Incorrectly Classified Instances 25768               34.815 %
Kappa statistic                      0.4093
Mean absolute error                  0.1798
Root mean squared error              0.3651
Relative absolute error             68.5788 %
Root relative squared error        100.8519 %
Total Number of Instances        74014

=== Detailed Accuracy By Class ===

               TP Rate FP Rate Precision Recall  F-Measure MCC     ROC Area PRC Area Class
               0.521   0.087   0.570     0.521   0.544     0.450   0.849    0.603    val_4
               0.708   0.238   0.849     0.708   0.772     0.450   0.831    0.892    val_2
               0.835   0.101   0.275     0.835   0.414     0.443   0.944    0.625    val_1
               0.479   0.111   0.373     0.479   0.419     0.331   0.810    0.436    val_3
Weighted Avg.  0.652   0.189   0.715     0.652   0.672     0.435   0.837    0.772

=== Confusion Matrix ===

    a     b     c     d   <-- classified as
 6999  3231   667  2541 |    a = val_4
 3395 34217 6140  4559 |    b = val_2
   76   320  2718   140 |    c = val_1
 1800  2556   343  4312 |    d = val_3
```

Testing:

```
=== Summary ===

Correctly Classified Instances   12055               65.1481 %
Incorrectly Classified Instances  6449               34.8519 %
Kappa statistic                      0.4083
Mean absolute error                  0.1802
Root mean squared error              0.3647
Relative absolute error             68.5688 %
Root relative squared error        100.4985 %
Total Number of Instances        18504

=== Detailed Accuracy By Class ===

               TP Rate FP Rate Precision Recall  F-Measure MCC     ROC Area PRC Area Class
               0.509   0.087   0.568     0.509   0.537     0.441   0.844    0.596    val_4
               0.712   0.240   0.846     0.712   0.773     0.452   0.831    0.890    val_2
               0.813   0.096   0.264     0.813   0.399     0.428   0.941    0.618    val_1
               0.496   0.116   0.382     0.496   0.432     0.341   0.808    0.445    val_3
Weighted Avg.  0.651   0.190   0.713     0.651   0.672     0.435   0.835    0.769

=== Confusion Matrix ===

   a    b    c    d   <-- classified as
1727  825  164  678 |   a = val_4
 838 8557 1466 1162 |   b = val_2
  17   95  615   29 |   c = val_1
```

```
  457  636   82 1156 |   d = val_3
```

B.2.2 NaiveBayes + CorrelationAttributeEval

Training:

```
=== Summary ===

Correctly Classified Instances   48223               65.1539 %
Incorrectly Classified Instances 25791               34.8461 %
Kappa statistic                      0.3012
Mean absolute error                  0.2138
Root mean squared error              0.3478
Relative absolute error             81.5668 %
Root relative squared error         96.0607 %
Total Number of Instances        74014

=== Detailed Accuracy By Class ===

               TP Rate FP Rate Precision Recall  F-Measure MCC     ROC Area PRC Area Class
               0.476   0.151   0.412     0.476   0.441     0.308   0.744    0.426    val_4
               0.826   0.464   0.770     0.826   0.797     0.376   0.769    0.844    val_2
               0.419   0.045   0.299     0.419   0.349     0.318   0.867    0.253    val_1
               0.063   0.024   0.269     0.063   0.102     0.077   0.694    0.239    val_3
Weighted Avg.  0.652   0.335   0.623     0.652   0.628     0.325   0.759    0.669

=== Confusion Matrix ===

    a     b     c     d   <-- classified as
 6390  5790   668   590 |    a = val_4
 6264 39904  1229   914 |    b = val_2
  140  1715  1362    37 |    c = val_1
 2715  4426  1303   567 |    d = val_3
```

Testing:

```
=== Summary ===

Correctly Classified Instances   11963               64.6509 %
Incorrectly Classified Instances  6541               35.3491 %
Kappa statistic                      0.2966
Mean absolute error                  0.2162
Root mean squared error              0.3498
Relative absolute error             82.3004 %
Root relative squared error         96.3998 %
Total Number of Instances        18504

=== Detailed Accuracy By Class ===

               TP Rate FP Rate Precision Recall  F-Measure MCC     ROC Area PRC Area Class
               0.473   0.151   0.413     0.473   0.441     0.306   0.743    0.426    val_4
               0.823   0.463   0.767     0.823   0.794     0.374   0.768    0.843    val_2
               0.431   0.049   0.273     0.431   0.335     0.308   0.865    0.242    val_1
               0.060   0.025   0.259     0.060   0.097     0.069   0.686    0.243    val_3
Weighted Avg.  0.647   0.333   0.618     0.647   0.623     0.320   0.757    0.667

=== Confusion Matrix ===

   a    b    c    d   <-- classified as
1604 1465  186  139 |   a = val_4
1551 9894  331  247 |   b = val_2
  25  393  326   12 |   c = val_1
 701 1141  350  139 |   d = val_3
```

B.2.3 NaiveBayes + GainRatioAttributeEval

Training:

```
=== Summary ===

Correctly Classified Instances   49939               67.4724 %
Incorrectly Classified Instances 24075               32.5276 %
Kappa statistic                      0.1716
Mean absolute error                  0.2469
Root mean squared error              0.3523
Relative absolute error             94.1852 %
Root relative squared error         97.3264 %
Total Number of Instances        74014

=== Detailed Accuracy By Class ===

               TP Rate FP Rate Precision Recall  F-Measure MCC     ROC Area PRC Area Class
               0.176   0.034   0.535     0.176   0.265     0.231   0.578    0.260    val_4
               0.960   0.816   0.689     0.960   0.802     0.240   0.572    0.688    val_2
               0.197   0.007   0.554     0.197   0.291     0.314   0.625    0.148    val_1
               0.061   0.008   0.509     0.061   0.109     0.144   0.541    0.160    val_3
Weighted Avg.  0.675   0.540   0.633     0.675   0.598     0.230   0.572    0.522

=== Confusion Matrix ===

    a     b     c    d   <-- classified as
 2363 10675  119   281 |    a = val_4
 1383 46384  302   242 |    b = val_2
    2  2603  642     7 |    c = val_1
  670  7696   95   550 |    d = val_3
```

Testing:

```
=== Summary ===

Correctly Classified Instances   12375               66.8774 %
Incorrectly Classified Instances  6129               33.1226 %
Kappa statistic                      0.1606
Mean absolute error                  0.2482
Root mean squared error              0.3542
Relative absolute error             94.4802 %
Root relative squared error         97.6164 %
Total Number of Instances        18504

=== Detailed Accuracy By Class ===

               TP Rate FP Rate Precision Recall  F-Measure MCC     ROC Area PRC Area Class
               0.164   0.034   0.523     0.164   0.249     0.216   0.571    0.253    val_4
               0.959   0.824   0.683     0.959   0.798     0.226   0.568    0.682    val_2
               0.181   0.008   0.481     0.181   0.263     0.278   0.616    0.124    val_1
               0.067   0.008   0.542     0.067   0.119     0.158   0.543    0.171    val_3
Weighted Avg.  0.669   0.543   0.628     0.669   0.590     0.218   0.567    0.516

=== Confusion Matrix ===

    a     b     c    d   <-- classified as
  556  2750   23    65 |    a = val_4
  331 11526  102    64 |    b = val_2
    0   616  137     3 |    c = val_1
  177  1975   23   156 |    d = val_3
```

B.2.4 NaiveBayes + InfoGainAttributeEval

Training:

```
=== Summary ===
```

```
Correctly Classified Instances   48302               65.2606 %
Incorrectly Classified Instances 25712               34.7394 %
Kappa statistic                      0.4128
Mean absolute error                  0.1769
Root mean squared error              0.3819
Relative absolute error             67.481 %
Root relative squared error        105.4853 %
Total Number of Instances        74014

=== Detailed Accuracy By Class ===

               TP Rate FP Rate Precision Recall  F-Measure MCC     ROC Area PRC Area Class
               0.513   0.086   0.570     0.513   0.540     0.445   0.857    0.605    val_4
               0.708   0.228   0.854     0.708   0.774     0.459   0.833    0.899    val_2
               0.837   0.104   0.271     0.837   0.409     0.439   0.948    0.615    val_1
               0.496   0.112   0.379     0.496   0.430     0.343   0.818    0.445    val_3
Weighted Avg.  0.653   0.183   0.719     0.653   0.674     0.441   0.841    0.778

=== Confusion Matrix ===

    a    b    c    d   <-- classified as
 6891 3093  838 2616 |    a = val_4
 3439 34215 6094 4563 |    b = val_2
   93  304 2724  133 |    c = val_1
 1667 2461  411 4472 |    d = val_3
```

Testing:

```
=== Summary ===

Correctly Classified Instances   12019               64.9535 %
Incorrectly Classified Instances  6485               35.0465 %
Kappa statistic                      0.4067
Mean absolute error                  0.179
Root mean squared error              0.3836
Relative absolute error             68.1207 %
Root relative squared error        105.7037 %
Total Number of Instances        18504

=== Detailed Accuracy By Class ===

               TP Rate FP Rate Precision Recall  F-Measure MCC     ROC Area PRC Area Class
               0.494   0.084   0.569     0.494   0.529     0.433   0.846    0.588    val_4
               0.711   0.236   0.848     0.711   0.774     0.455   0.827    0.891    val_2
               0.815   0.101   0.255     0.815   0.389     0.419   0.940    0.581    val_1
               0.505   0.117   0.384     0.505   0.436     0.346   0.803    0.439    val_3
Weighted Avg.  0.650   0.187   0.714     0.650   0.671     0.436   0.832    0.766

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
1675 798 220 701 |   a = val_4
 836 8551 1485 1151 |  b = val_2
  15  87 616  38 |   c = val_1
 418 643  93 1177 |   d = val_3
```

B.2.5 NaiveBayes + SelfSelected

Training:

```
=== Summary ===

Correctly Classified Instances   50387               68.0777 %
Incorrectly Classified Instances 23627               31.9223 %
Kappa statistic                      0.413
Mean absolute error                  0.1838
```

```
Root mean squared error              0.3312
Relative absolute error             70.137 %
Root relative squared error         91.4758 %
Total Number of Instances        74014

=== Detailed Accuracy By Class ===

               TP Rate FP Rate Precision Recall  F-Measure MCC     ROC Area PRC Area Class
               0.511   0.085   0.571     0.511   0.539     0.445   0.851    0.599    val_4
               0.781   0.332   0.815     0.781   0.798     0.441   0.817    0.888    val_2
               0.746   0.060   0.363     0.746   0.488     0.490   0.946    0.561    val_1
               0.375   0.087   0.374     0.375   0.374     0.287   0.803    0.396    val_3
Weighted Avg.  0.681   0.245   0.697     0.681   0.686     0.425   0.827    0.761

=== Confusion Matrix ===

    a     b     c     d   <-- classified as
 6868  4288   406  1876 |    a = val_4
 3360 37711 3625  3615 |    b = val_2
   65   583  2427   179 |    c = val_1
 1740  3664   226  3381 |    d = val_3
```

Testing:

```
=== Summary ===

Correctly Classified Instances   12577              67.9691 %
Incorrectly Classified Instances  5927              32.0309 %
Kappa statistic                      0.4084
Mean absolute error                  0.185
Root mean squared error              0.3315
Relative absolute error             70.4294 %
Root relative squared error         91.3497 %
Total Number of Instances        18504

=== Detailed Accuracy By Class ===

               TP Rate FP Rate Precision Recall  F-Measure MCC     ROC Area PRC Area Class
               0.502   0.085   0.571     0.502   0.534     0.439   0.847    0.596    val_4
               0.786   0.343   0.810     0.786   0.798     0.438   0.815    0.884    val_2
               0.713   0.056   0.351     0.713   0.471     0.472   0.939    0.548    val_1
               0.380   0.088   0.383     0.380   0.381     0.293   0.801    0.403    val_3
Weighted Avg.  0.680   0.252   0.693     0.680   0.683     0.421   0.824    0.757

=== Confusion Matrix ===

   a    b    c    d   <-- classified as
1705 1111  88  490 |   a = val_4
 820 9448 855  900 |   b = val_2
  16  165 539   36 |   c = val_1
 447  947  52  885 |   d = val_3
```

B.3 J48

B.3.1 J48 + CfsSubsetEval

Training:

```
=== Summary ===

Correctly Classified Instances   58193              78.6243 %
Incorrectly Classified Instances 15821              21.3757 %
Kappa statistic                      0.543
Mean absolute error                  0.1575
Root mean squared error              0.2806
Relative absolute error             60.0923 %
Root relative squared error         77.5201 %
```

```
Total Number of Instances        74014

=== Detailed Accuracy By Class ===

               TP Rate FP Rate Precision Recall  F-Measure MCC     ROC Area PRC Area Class
               0.581   0.056   0.696     0.581   0.633     0.564   0.887    0.702    val_4
               0.944   0.414   0.811     0.944   0.872     0.591   0.874    0.914    val_2
               0.525   0.006   0.805     0.525   0.636     0.638   0.966    0.623    val_1
               0.340   0.021   0.694     0.340   0.456     0.441   0.856    0.557    val_3
Weighted Avg.  0.786   0.283   0.775     0.786   0.768     0.570   0.879    0.819

=== Confusion Matrix ===

    a     b     c    d   <-- classified as
 7803  4932   140  563 |    a = val_4
 1756 45620   179  756 |    b = val_2
   83  1431  1708   32 |    c = val_1
 1568  4287    94 3062 |    d = val_3
```

Testing:

```
=== Summary ===

Correctly Classified Instances   14016               75.7458 %
Incorrectly Classified Instances  4488               24.2542 %
Kappa statistic                      0.4845
Mean absolute error                  0.1699
Root mean squared error              0.3002
Relative absolute error             64.6669 %
Root relative squared error         82.735 %
Total Number of Instances        18504

=== Detailed Accuracy By Class ===

               TP Rate FP Rate Precision Recall  F-Measure MCC     ROC Area PRC Area Class
               0.524   0.068   0.634     0.524   0.574     0.492   0.837    0.591    val_4
               0.927   0.444   0.795     0.927   0.856     0.539   0.845    0.880    val_2
               0.508   0.006   0.785     0.508   0.617     0.620   0.946    0.587    val_1
               0.302   0.029   0.596     0.302   0.400     0.370   0.797    0.436    val_3
Weighted Avg.  0.757   0.305   0.740     0.757   0.737     0.512   0.841    0.759

=== Confusion Matrix ===

    a     b    c    d   <-- classified as
 1779  1380   32  203 |    a = val_4
  567 11150   48  258 |    b = val_2
   16   340  384   16 |    c = val_1
  446  1157   25  703 |    d = val_3
```

B.3.2 J48 + CorrelationAttributeEval

Training:

```
=== Summary ===

Correctly Classified Instances   53072               71.7054 %
Incorrectly Classified Instances 20942               28.2946 %
Kappa statistic                      0.3712
Mean absolute error                  0.2031
Root mean squared error              0.3187
Relative absolute error             77.4869 %
Root relative squared error         88.0276 %
Total Number of Instances        74014

=== Detailed Accuracy By Class ===

               TP Rate FP Rate Precision Recall  F-Measure MCC     ROC Area PRC Area Class
```

```
               0.395   0.057   0.607     0.395   0.479     0.404   0.803    0.530    val_4
               0.929   0.549   0.761     0.929   0.837     0.451   0.801    0.856    val_2
               0.332   0.013   0.549     0.332   0.414     0.407   0.897    0.378    val_1
               0.197   0.039   0.414     0.197   0.267     0.222   0.736    0.295    val_3
Weighted Avg.  0.717   0.374   0.681     0.717   0.684     0.413   0.797    0.707

=== Confusion Matrix ===

    a    b    c    d   <-- classified as
 5312 6988  254  884 |    a = val_4
 1899 44903 308 1201 |    b = val_2
   75 1667 1079  433 |    c = val_1
 1466 5444  323 1778 |    d = val_3
```

Testing:

```
=== Summary ===

Correctly Classified Instances   13022               70.374 %
Incorrectly Classified Instances  5482               29.626 %
Kappa statistic                      0.3455
Mean absolute error                  0.2069
Root mean squared error              0.3239
Relative absolute error             78.734 %
Root relative squared error         89.2693 %
Total Number of Instances        18504

=== Detailed Accuracy By Class ===

               TP Rate FP Rate Precision Recall  F-Measure MCC     ROC Area PRC Area Class
               0.365   0.060   0.576     0.365   0.447     0.368   0.792    0.502    val_4
               0.922   0.562   0.753     0.922   0.829     0.426   0.792    0.849    val_2
               0.312   0.013   0.499     0.312   0.384     0.375   0.882    0.346    val_1
               0.199   0.043   0.402     0.199   0.266     0.214   0.727    0.281    val_3
Weighted Avg.  0.704   0.382   0.666     0.704   0.670     0.386   0.787    0.693

=== Confusion Matrix ===

    a    b    c    d   <-- classified as
 1239 1836   71  248 |    a = val_4
  520 11084  92  327 |    b = val_2
   17  388  236  115 |    c = val_1
  374 1420   74  463 |    d = val_3
```

B.3.3 J48 + GainRatioAttributeEval

Training:

```
=== Summary ===

Correctly Classified Instances   49985               67.5345 %
Incorrectly Classified Instances 24029               32.4655 %
Kappa statistic                      0.1722
Mean absolute error                  0.2477
Root mean squared error              0.3519
Relative absolute error             94.5042 %
Root relative squared error         97.2144 %
Total Number of Instances        74014

=== Detailed Accuracy By Class ===

               TP Rate FP Rate Precision Recall  F-Measure MCC     ROC Area PRC Area Class
               0.178   0.034   0.536     0.178   0.268     0.233   0.579    0.259    val_4
               0.961   0.817   0.689     0.961   0.802     0.240   0.572    0.688    val_2
               0.197   0.007   0.553     0.197   0.291     0.314   0.625    0.148    val_1
               0.058   0.007   0.547     0.058   0.106     0.149   0.542    0.160    val_3
Weighted Avg.  0.675   0.541   0.638     0.675   0.598     0.231   0.572    0.522
```

```
=== Confusion Matrix ===

    a     b    c    d   <-- classified as
 2397 10688  120  233 |    a = val_4
 1390 46419  303  199 |    b = val_2
    3  2604  642    5 |    c = val_1
  681  7708   95  527 |    d = val_3
```

Testing:

```
=== Summary ===

Correctly Classified Instances   12388               66.9477 %
Incorrectly Classified Instances  6116               33.0523 %
Kappa statistic                      0.161
Mean absolute error                  0.2491
Root mean squared error              0.3539
Relative absolute error             94.8142 %
Root relative squared error         97.5402 %
Total Number of Instances        18504

=== Detailed Accuracy By Class ===

               TP Rate FP Rate Precision Recall  F-Measure MCC     ROC Area PRC Area Class
               0.168   0.034   0.522     0.168   0.254     0.219   0.571    0.253    val_4
               0.959   0.826   0.683     0.959   0.798     0.225   0.568    0.683    val_2
               0.181   0.008   0.481     0.181   0.263     0.278   0.617    0.124    val_1
               0.064   0.006   0.616     0.064   0.116     0.170   0.544    0.171    val_3
Weighted Avg.  0.669   0.544   0.637     0.669   0.590     0.219   0.567    0.516

=== Confusion Matrix ===

    a     b    c    d   <-- classified as
  570  2754   23   47 |    a = val_4
  345 11532  102   44 |    b = val_2
    1   616  137    2 |    c = val_1
  175  1984   23  149 |    d = val_3
```

B.3.4 J48 + InfoGainAttributeEval

Training:

```
=== Summary ===

Correctly Classified Instances   56834               76.7882 %
Incorrectly Classified Instances 17180               23.2118 %
Kappa statistic                      0.4957
Mean absolute error                  0.1693
Root mean squared error              0.291
Relative absolute error             64.5964 %
Root relative squared error         80.3728 %
Total Number of Instances        74014

=== Detailed Accuracy By Class ===

               TP Rate FP Rate Precision Recall  F-Measure MCC     ROC Area PRC Area Class
               0.516   0.058   0.662     0.516   0.580     0.506   0.859    0.645    val_4
               0.943   0.461   0.793     0.943   0.862     0.550   0.857    0.901    val_2
               0.499   0.011   0.682     0.499   0.576     0.567   0.954    0.513    val_1
               0.301   0.016   0.724     0.301   0.425     0.426   0.847    0.541    val_3
Weighted Avg.  0.768   0.314   0.756     0.768   0.745     0.528   0.861    0.793

=== Confusion Matrix ===

    a     b    c    d   <-- classified as
 6930  5677  349  482 |    a = val_4
```

```
 1971 45567  250   523 |    b = val_2
   38  1565 1623    28 |    c = val_1
 1523  4617  157  2714 |    d = val_3
```

Testing:

```
=== Summary ===

Correctly Classified Instances   13721               74.1515 %
Incorrectly Classified Instances  4783               25.8485 %
Kappa statistic                      0.4428
Mean absolute error                  0.1813
Root mean squared error              0.3096
Relative absolute error             69.0109 %
Root relative squared error         85.3116 %
Total Number of Instances        18504

=== Detailed Accuracy By Class ===

               TP Rate FP Rate Precision Recall  F-Measure MCC     ROC Area PRC Area Class
               0.460   0.070   0.598     0.460   0.520     0.434   0.804    0.538    val_4
               0.929   0.483   0.781     0.929   0.849     0.508   0.824    0.860    val_2
               0.491   0.012   0.644     0.491   0.557     0.546   0.935    0.467    val_1
               0.266   0.025   0.608     0.266   0.370     0.351   0.778    0.425    val_3
Weighted Avg.  0.742   0.330   0.720     0.742   0.716     0.476   0.819    0.730

=== Confusion Matrix ===

    a    b    c    d   <-- classified as
 1562 1549   96  187 |    a = val_4
  584 11169  70  200 |    b = val_2
   11  362  371   12 |    c = val_1
  456 1217   39  619 |    d = val_3
```

B.3.5 J48 + SelfSelected

Training:

```
=== Summary ===

Correctly Classified Instances   55390               74.8372 %
Incorrectly Classified Instances 18624               25.1628 %
Kappa statistic                      0.4447
Mean absolute error                  0.1813
Root mean squared error              0.3011
Relative absolute error             69.182 %
Root relative squared error         83.1767 %
Total Number of Instances        74014

=== Detailed Accuracy By Class ===

               TP Rate FP Rate Precision Recall  F-Measure MCC     ROC Area PRC Area Class
               0.473   0.062   0.630     0.473   0.540     0.462   0.838    0.609    val_4
               0.937   0.512   0.775     0.937   0.848     0.496   0.829    0.879    val_2
               0.459   0.007   0.748     0.459   0.569     0.572   0.953    0.570    val_1
               0.253   0.019   0.649     0.253   0.364     0.360   0.825    0.483    val_3
Weighted Avg.  0.748   0.348   0.732     0.748   0.721     0.477   0.836    0.768

=== Confusion Matrix ===

    a    b    c    d   <-- classified as
 6358 6406  170  504 |    a = val_4
 2152 45258 213  688 |    b = val_2
   90 1631 1493   40 |    c = val_1
 1489 5122  119 2281 |    d = val_3
```

Testing:

```
=== Summary ===

Correctly Classified Instances   13502               72.968 %
Incorrectly Classified Instances  5002               27.032 %
Kappa statistic                      0.4077
Mean absolute error                  0.1892
Root mean squared error              0.3127
Relative absolute error             72.0293 %
Root relative squared error         86.1892 %
Total Number of Instances        18504

=== Detailed Accuracy By Class ===

               TP Rate FP Rate Precision Recall  F-Measure MCC     ROC Area PRC Area Class
               0.437   0.071   0.580     0.437   0.499     0.411   0.803    0.539    val_4
               0.927   0.526   0.766     0.927   0.838     0.466   0.810    0.856    val_2
               0.438   0.007   0.720     0.438   0.544     0.547   0.927    0.499    val_1
               0.235   0.024   0.583     0.235   0.335     0.319   0.782    0.415    val_3
Weighted Avg.  0.730   0.358   0.707     0.730   0.701     0.441   0.810    0.728

=== Confusion Matrix ===

    a    b    c    d   <-- classified as
 1483 1702   42  167 |    a = val_4
  611 11140  57  215 |    b = val_2
   36  379  331   10 |    c = val_1
  425 1328   30  548 |    d = val_3
```

B.4 BayesNet

B.4.1 BayesNet + CfsSubsetEval

Training:

```
=== Summary ===

Correctly Classified Instances   48232               65.166 %
Incorrectly Classified Instances 25782               34.834 %
Kappa statistic                      0.4092
Mean absolute error                  0.1797
Root mean squared error              0.3654
Relative absolute error             68.5691 %
Root relative squared error        100.9214 %
Total Number of Instances        74014

=== Detailed Accuracy By Class ===

               TP Rate FP Rate Precision Recall  F-Measure MCC     ROC Area PRC Area Class
               0.521   0.087   0.570     0.521   0.544     0.450   0.849    0.603    val_4
               0.708   0.237   0.849     0.708   0.772     0.450   0.831    0.892    val_2
               0.835   0.101   0.275     0.835   0.414     0.443   0.945    0.625    val_1
               0.479   0.112   0.373     0.479   0.419     0.331   0.810    0.436    val_3
Weighted Avg.  0.652   0.189   0.715     0.652   0.672     0.435   0.837    0.772

=== Confusion Matrix ===

    a    b    c    d   <-- classified as
 6999 3225  670 2544 |    a = val_4
 3396 34200 6143 4572 |    b = val_2
   77  319 2717  141 |    c = val_1
 1799 2559  337 4316 |    d = val_3
```

Testing:

```
=== Summary ===
```

```
Correctly Classified Instances   12051               65.1265 %
Incorrectly Classified Instances  6453               34.8735 %
Kappa statistic                      0.4082
Mean absolute error                  0.1801
Root mean squared error              0.3649
Relative absolute error             68.559 %
Root relative squared error        100.5698 %
Total Number of Instances        18504

=== Detailed Accuracy By Class ===

               TP Rate FP Rate Precision Recall  F-Measure MCC     ROC Area PRC Area Class
               0.509   0.087   0.569     0.509   0.537     0.441   0.844    0.596    val_4
               0.711   0.240   0.846     0.711   0.773     0.451   0.831    0.890    val_2
               0.813   0.096   0.264     0.813   0.399     0.428   0.942    0.618    val_1
               0.497   0.116   0.382     0.497   0.432     0.342   0.808    0.445    val_3
Weighted Avg.  0.651   0.190   0.713     0.651   0.671     0.435   0.835    0.769

=== Confusion Matrix ===

   a    b    c    d   <-- classified as
 1727  825  162  680 |   a = val_4
  839 8550 1468 1166 |   b = val_2
   16   95  615   30 |   c = val_1
  455  636   81 1159 |   d = val_3
```

B.4.2 BayesNet + CorrelationAttributeEval

Training:

```
=== Summary ===

Correctly Classified Instances   48223               65.1539 %
Incorrectly Classified Instances 25791               34.8461 %
Kappa statistic                      0.3012
Mean absolute error                  0.2138
Root mean squared error              0.3478
Relative absolute error             81.5651 %
Root relative squared error         96.0679 %
Total Number of Instances        74014

=== Detailed Accuracy By Class ===

               TP Rate FP Rate Precision Recall  F-Measure MCC     ROC Area PRC Area Class
               0.476   0.151   0.412     0.476   0.441     0.308   0.744    0.426    val_4
               0.826   0.464   0.770     0.826   0.797     0.376   0.769    0.844    val_2
               0.419   0.045   0.299     0.419   0.349     0.318   0.867    0.253    val_1
               0.063   0.024   0.269     0.063   0.102     0.077   0.694    0.239    val_3
Weighted Avg.  0.652   0.335   0.623     0.652   0.628     0.325   0.759    0.669

=== Confusion Matrix ===

    a     b     c     d   <-- classified as
 6390  5790   668   590 |    a = val_4
 6264 39904 1229   914 |    b = val_2
  140  1715  1362    37 |    c = val_1
 2715  4426  1303   567 |    d = val_3
```

Testing:

```
=== Summary ===

Correctly Classified Instances   11963               64.6509 %
Incorrectly Classified Instances  6541               35.3491 %
Kappa statistic                      0.2966
Mean absolute error                  0.2162
Root mean squared error              0.3498
```

```
Relative absolute error             82.2989 %
Root relative squared error         96.4073 %
Total Number of Instances        18504

=== Detailed Accuracy By Class ===

               TP Rate FP Rate Precision Recall  F-Measure MCC     ROC Area PRC Area Class
               0.473   0.151   0.413     0.473   0.441     0.306   0.743    0.426    val_4
               0.823   0.463   0.767     0.823   0.794     0.374   0.768    0.843    val_2
               0.431   0.049   0.273     0.431   0.335     0.308   0.865    0.242    val_1
               0.060   0.025   0.259     0.060   0.097     0.069   0.686    0.243    val_3
Weighted Avg.  0.647   0.333   0.618     0.647   0.623     0.320   0.757    0.667

=== Confusion Matrix ===

   a    b    c    d   <-- classified as
 1604 1465 186  139 |   a = val_4
 1551 9894 331  247 |   b = val_2
   25  393  326   12 |   c = val_1
  701 1141 350  139 |   d = val_3
```

B.4.3 BayesNet + GainRatioAttributeEval

Training:

```
=== Summary ===

Correctly Classified Instances   49940             67.4737 %
Incorrectly Classified Instances 24074             32.5263 %
Kappa statistic                      0.1717
Mean absolute error                  0.2469
Root mean squared error              0.3523
Relative absolute error             94.1863 %
Root relative squared error         97.3252 %
Total Number of Instances        74014

=== Detailed Accuracy By Class ===

               TP Rate FP Rate Precision Recall  F-Measure MCC     ROC Area PRC Area Class
               0.176   0.034   0.535     0.176   0.265     0.231   0.578    0.260    val_4
               0.960   0.816   0.689     0.960   0.802     0.240   0.572    0.688    val_2
               0.197   0.007   0.555     0.197   0.291     0.314   0.625    0.148    val_1
               0.061   0.008   0.509     0.061   0.109     0.144   0.541    0.160    val_3
Weighted Avg.  0.675   0.540   0.633     0.675   0.598     0.230   0.572    0.522

=== Confusion Matrix ===

    a     b    c    d   <-- classified as
 2364 10675  118  281 |    a = val_4
 1383 46384  302  242 |    b = val_2
    2  2603  642    7 |    c = val_1
  670  7696   95  550 |    d = val_3
```

Testing:

```
=== Summary ===

Correctly Classified Instances   12375             66.8774 %
Incorrectly Classified Instances  6129             33.1226 %
Kappa statistic                      0.1606
Mean absolute error                  0.2482
Root mean squared error              0.3542
Relative absolute error             94.4823 %
Root relative squared error         97.6153 %
Total Number of Instances        18504

=== Detailed Accuracy By Class ===
```

```
               TP Rate FP Rate Precision Recall  F-Measure MCC     ROC Area PRC Area Class
               0.164   0.034   0.523     0.164   0.249     0.216   0.571    0.253    val_4
               0.959   0.824   0.683     0.959   0.798     0.226   0.568    0.682    val_2
               0.181   0.008   0.481     0.181   0.263     0.278   0.617    0.124    val_1
               0.067   0.008   0.542     0.067   0.119     0.158   0.543    0.171    val_3
Weighted Avg.  0.669   0.543   0.628     0.669   0.590     0.218   0.567    0.516

=== Confusion Matrix ===

    a     b    c    d   <-- classified as
  556  2750   23   65 |    a = val_4
  331 11526  102   64 |    b = val_2
    0   616  137    3 |    c = val_1
  177  1975   23  156 |    d = val_3
```

B.4.4 BayesNet + InfoGainAttributeEval

Training:

```
=== Summary ===

Correctly Classified Instances   48460               65.4741 %
Incorrectly Classified Instances 25554               34.5259 %
Kappa statistic                      0.4151
Mean absolute error                  0.176
Root mean squared error              0.3814
Relative absolute error             67.1494 %
Root relative squared error        105.346 %
Total Number of Instances        74014

=== Detailed Accuracy By Class ===

               TP Rate FP Rate Precision Recall  F-Measure MCC     ROC Area PRC Area Class
               0.514   0.085   0.572     0.514   0.541     0.447   0.858    0.606    val_4
               0.711   0.229   0.854     0.711   0.776     0.460   0.834    0.899    val_2
               0.837   0.101   0.276     0.837   0.415     0.444   0.950    0.618    val_1
               0.499   0.113   0.379     0.499   0.431     0.344   0.819    0.446    val_3
Weighted Avg.  0.655   0.183   0.719     0.655   0.675     0.443   0.842    0.778

=== Confusion Matrix ===

    a     b     c    d   <-- classified as
 6909  3105   795  2629 |    a = val_4
 3423 34335  5966  4587 |    b = val_2
   92   305  2722   135 |    c = val_1
 1658  2474   385  4494 |    d = val_3
```

Testing:

```
=== Summary ===

Correctly Classified Instances   12034               65.0346 %
Incorrectly Classified Instances  6470               34.9654 %
Kappa statistic                      0.4071
Mean absolute error                  0.1782
Root mean squared error              0.3832
Relative absolute error             67.8358 %
Root relative squared error        105.6172 %
Total Number of Instances        18504

=== Detailed Accuracy By Class ===

               TP Rate FP Rate Precision Recall  F-Measure MCC     ROC Area PRC Area Class
               0.494   0.084   0.568     0.494   0.529     0.433   0.846    0.588    val_4
               0.713   0.237   0.848     0.713   0.774     0.456   0.827    0.891    val_2
               0.811   0.099   0.259     0.811   0.392     0.422   0.941    0.582    val_1
```

```
                0.505   0.118   0.382     0.505   0.435     0.345   0.803    0.439    val_3
Weighted Avg.   0.650   0.188   0.714     0.650   0.671     0.436   0.832    0.766

=== Confusion Matrix ===

   a    b    c    d   <-- classified as
1677  802  212  703 |   a = val_4
 839 8568 1455 1161 |   b = val_2
  15   89  613   39 |   c = val_1
 419  646   90 1176 |   d = val_3
```

B.4.5 BayesNet + SelfSelected

Training:

```
=== Summary ===

Correctly Classified Instances   50360               68.0412 %
Incorrectly Classified Instances 23654               31.9588 %
Kappa statistic                      0.4125
Mean absolute error                  0.1838
Root mean squared error              0.3314
Relative absolute error             70.1271 %
Root relative squared error         91.5526 %
Total Number of Instances        74014

=== Detailed Accuracy By Class ===

               TP Rate FP Rate Precision Recall  F-Measure MCC     ROC Area PRC Area Class
               0.511   0.085   0.571     0.511   0.539     0.445   0.851    0.599    val_4
               0.780   0.332   0.815     0.780   0.797     0.441   0.817    0.888    val_2
               0.744   0.060   0.361     0.744   0.486     0.488   0.947    0.560    val_1
               0.375   0.088   0.373     0.375   0.374     0.287   0.803    0.396    val_3
Weighted Avg.  0.680   0.245   0.697     0.680   0.685     0.425   0.827    0.761

=== Confusion Matrix ===

    a     b     c     d   <-- classified as
 6862  4284   409  1883 |    a = val_4
 3355 37696 3636 3624 |    b = val_2
   67   585  2420   182 |    c = val_1
 1738  3661   230  3382 |    d = val_3
```

Testing:

```
=== Summary ===

Correctly Classified Instances   12564               67.8988 %
Incorrectly Classified Instances  5940               32.1012 %
Kappa statistic                      0.4074
Mean absolute error                  0.185
Root mean squared error              0.3317
Relative absolute error             70.4192 %
Root relative squared error         91.4251 %
Total Number of Instances        18504

=== Detailed Accuracy By Class ===

               TP Rate FP Rate Precision Recall  F-Measure MCC     ROC Area PRC Area Class
               0.502   0.085   0.570     0.502   0.534     0.438   0.847    0.596    val_4
               0.785   0.343   0.809     0.785   0.797     0.436   0.815    0.884    val_2
               0.712   0.056   0.350     0.712   0.469     0.470   0.940    0.547    val_1
               0.380   0.088   0.382     0.380   0.381     0.292   0.801    0.403    val_3
Weighted Avg.  0.679   0.252   0.693     0.679   0.683     0.420   0.824    0.757

=== Confusion Matrix ===
```

```
   a    b    c    d   <-- classified as
1703 1113  88  490 |   a = val_4
 821 9437 860  905 |   b = val_2
  16  166 538   36 |   c = val_1
 447  946  52  886 |   d = val_3
```