Leveraging 'ensemble' Machine Learning models to drive automation of specification generation and data mapping in clinical trial data submission

Venugopal Mallarapu, eClinical Solutions, Mansfield, MA, USA
Nagendra Karamala, eClinical Solutions, Bengaluru, KA, India
Nathan Johnson, eClinical Solutions, Mansfield, MA, USA

ABSTRACT

The submission of clinical trial data to regulatory authorities is a complex and time-consuming process, often requiring extensive manual effort to map data to standardized formats. Leveraging an ensemble of machine learning models for automated specification generation and data mapping can streamline this process. By analyzing existing study protocols, statistical analysis plans, and case report forms, these ensemble models can generate data specifications and mappings with high accuracy and minimal human intervention. The ensemble approach combines the strengths of multiple models, reducing errors and improving consistency across studies. This automation accelerates the data submission timeline while ensuring regulatory compliance. Furthermore, automated mapping facilitates the reuse of standardized data elements, promoting data interoperability and enabling cross-study analyses. As clinical trials become increasingly complex with larger volumes of diverse data sources, an ensemble of automated models provides a scalable and robust solution to enhance efficiency, data quality, and regulatory compliance in submissions.

INTRODUCTION

In the realm of clinical trials, the accurate and efficient submission of trial data is paramount. The process of specification generation and data mapping, however, remains a complex and labor-intensive task, often prone to human error. With the advent of advanced machine learning techniques, there is a burgeoning opportunity to streamline and automate these processes, thereby enhancing both accuracy and efficiency.

This paper explores the potential of leveraging ensemble machine learning models to drive the automation of specification generation and data mapping in clinical trial data submissions. Ensemble models, which combine the predictions of multiple base learners to improve overall performance, offer a robust solution to the challenges faced in this domain. By integrating various machine learning algorithms, ensemble models can effectively handle the diverse and intricate nature of clinical trial data.

The primary objective of this paper is to demonstrate how ensemble machine learning models can be utilized to automate the specification generation and data mapping processes, reducing the manual effort required and minimizing the risk of errors. Through a series of experiments and case studies, we aim to highlight the efficacy of these models in real-world clinical trial settings.

CHALLENGES IN CLINICAL TRIAL DATA SUBMISSION

Clinical trial data submission involves several challenges that can impact the efficiency and accuracy of the process. Here are some of the key challenges:

1. Data Quality and Integrity: Ensuring consistent data quality across multiple sites and datasets is a significant challenge. Variations in data collection methods, discrepancies in data entry, and interpretation errors can compromise the integrity of the trial data1.
2. Regulatory Compliance: Adhering to regulatory standards and guidelines, such as Good Clinical Practice (GCP) and International Council for Harmonization (ICH) guidelines, is crucial. Ensuring compliance across different jurisdictions and evolving regulatory landscapes adds complexity1.
3. Timeliness and Efficiency: Meeting project timelines and ensuring timely data submission is essential. Delays in data delivery can lead to costly setbacks and jeopardize study outcomes1.
4. Data Privacy and Security: Safeguarding patient data against unauthorized access is a top priority. Implementing robust data privacy and access control policies is necessary to protect sensitive information while ensuring accessibility for authorized users1.

5. Interoperability and Standardization: The lack of standardized data formats and interoperability between different systems can hinder the seamless exchange and integration of clinical trial data2.
6. Complexity of Multicenter Studies: Multicenter studies require consent from all investigators at each center, which can be logistically challenging. Additionally, managing data collected in multiple languages and ensuring its quality adds to the complexity2.
7. Technological Limitations: Remote access platforms with software limitations and internet requirements can impede data sharing and analysis. On-site data analysis may be necessary when data cannot be moved2.

Addressing these challenges requires a combination of robust quality control processes, adherence to regulatory standards, effective communication and collaboration, and the use of advanced technologies to streamline data management and submission processes.

BENEFITS OF AUTOMATION

Automation can significantly mitigate the challenges in clinical trial data submission by enhancing efficiency, accuracy, and compliance. Here are some ways automation can help:

1. Improving Data Quality and Integrity: Automated data validation and cleaning processes can identify and correct errors in real-time, ensuring high-quality data. Machine learning algorithms can detect anomalies and inconsistencies that might be missed by manual checks.
2. Ensuring Regulatory Compliance: Automation tools can be programmed to adhere to regulatory standards and guidelines, ensuring that all submissions meet the necessary requirements. This reduces the risk of non-compliance and the associated penalties.
3. Enhancing Timeliness and Efficiency: Automated workflows can streamline data collection, processing, and submission, significantly reducing the time required for these tasks. This helps in meeting project deadlines and accelerates the overall trial timeline.
4. Strengthening Data Privacy and Security: Automated systems can enforce strict access controls and encryption protocols to protect sensitive patient data. This ensures that only authorized personnel can access the data, thereby maintaining confidentiality and security.
5. Facilitating Interoperability and Standardization: Automation can standardize data formats and ensure compatibility between different systems. This enables seamless data exchange and integration, reducing the manual effort required to reconcile disparate data sources.
6. Managing Complexity of Multicenter Studies: Automated systems can handle the logistics of multicenter studies more efficiently by centralizing data collection and management. This ensures consistency across sites and simplifies the coordination of data from multiple locations.
7. Overcoming Technological Limitations: Automation can optimize data transfer and processing, even in environments with limited technological infrastructure. Cloud-based solutions can provide remote access to data, facilitating real-time analysis and decision-making.

By leveraging automation, clinical trial processes can become more streamlined, accurate, and compliant, ultimately leading to faster and more reliable trial outcomes.

DATA MAPPING

Clinical trial data is organized and formatted in various ways based on its intended use. The optimal structure for data collection is not necessarily optimal for review, for tabulation, or for analysis. Thus, while data is being processed it is necessary for it to be changed from one organizational model to another. Data mapping is the process of transforming data from one standard to another. This includes the transformation of data from its raw collection format to a tabulation standard, or transformation from the tabulation standard to an analysis standard. Mapping could also define the transformation of data from one standard to an output format, such as a data review listing.

The process of mapping data typically involves two major components: specification and execution.

- Specification describes the documentation of how mapping is to be performed. This step is usually performed by a programmer or subject matter expert who understands both the source and the target data standards.
- Execution is the carrying out of the mapping defined in the specification, typically by a programmer who translates the mapping instruction into code.

MAPPING SPECIFICATION

The specification document is an important component of the mapping process. Typically this is developed prior to mapping execution and defines how mapping should be performed. It will be used by programmers during mapping execution as a technical guide to the mapping.

A mapping specification at a minimum contains linkages between source data, target data, and the conversions between them. For example, in the case of mapping from raw to SDTM, it would include for each SDTM target variable the corresponding source variable(s) and a description of the derivation.

Table 1: Sample mapping

Target Domain	Target Variable	Source Table	Source Variable(s)	Derivation
AE	AESTDTC	AE1	AESTDAT	Convert datetime variable to ISO8601 date time expression
AE	USUBJID	AE1	PROJECT SUBJECT	Concatenate PROJECT and SUBJECT separated by '-'

SDTM specification

The SDTM mapping specification is an important document that's used when designing the process by which raw data will be converted to SDTM. This document specifies how the raw data is to be converted and is used by the SDTM programmer and testing team.

1. Examine the CRFs and raw data and identify which SDTM domains you need.
2. Against each SDTM domain, note which raw dataset will provide the input data.
3. Against each SDTM domain, list all variables and describe how they are to be programmed.

Mapping Execution

Once a Specification is developed, programmers carry out the mapping activities to transform data from the source to the target format. This can be done using a variety of languages and technologies, including R®, SAS® or SQL.

As part of the execution process, mapping activities undergo some form of validation. This could include independent double programming, code review, output review, and/or manual or programmatic checks of derived values. The validation process also will include formal checks of the resulting data against the intended standard.

APPROACH

The approach to auto-mapping involved the steps outlined (Fig.1) below:

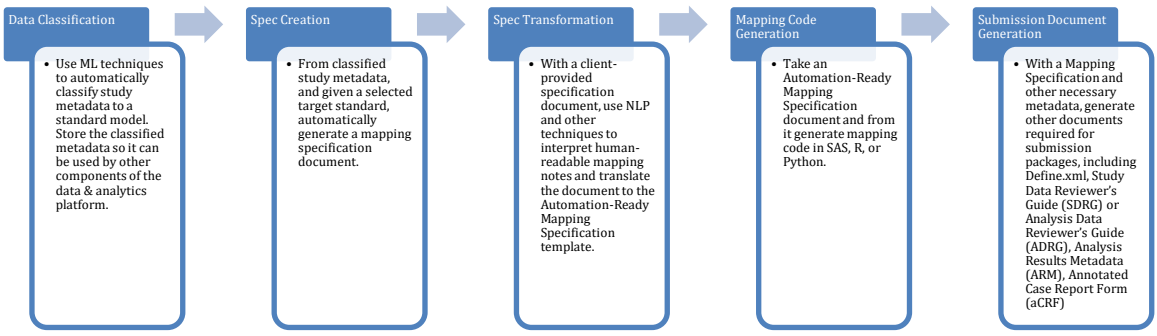

Fig. 1: Approach to Automapping

SOLUTIONS AND SYSTEM WORKFLOW

Our initiative to build the machine learning models involved multiple steps. The diagram below (Fig. 2) depicts the steps:

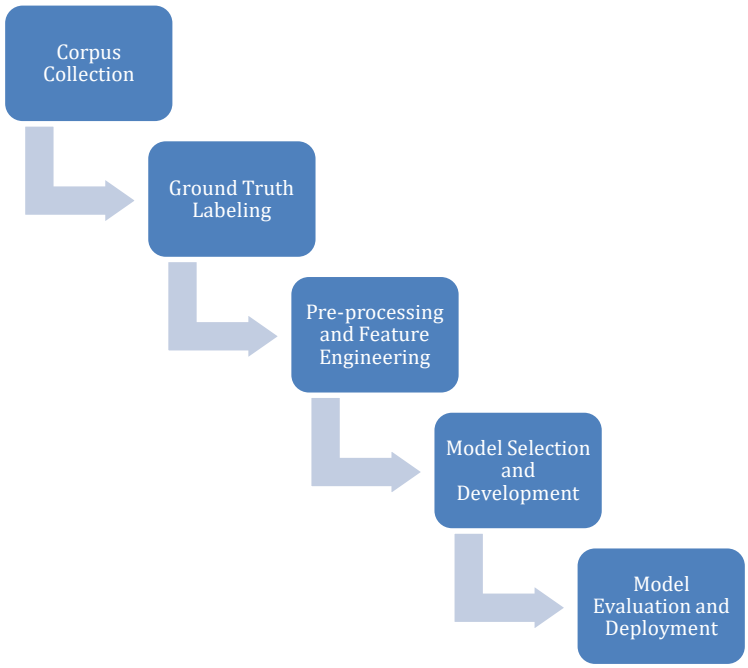

Fig. 2: Steps to build and deploy ML models

CORPUS DATA COLLECTION:

To create the corpus data, we looked at multiple sources that will be leveraged to train, test & validate the machine learning model. The categories are highlighted in the diagram (Fig. 3) below:

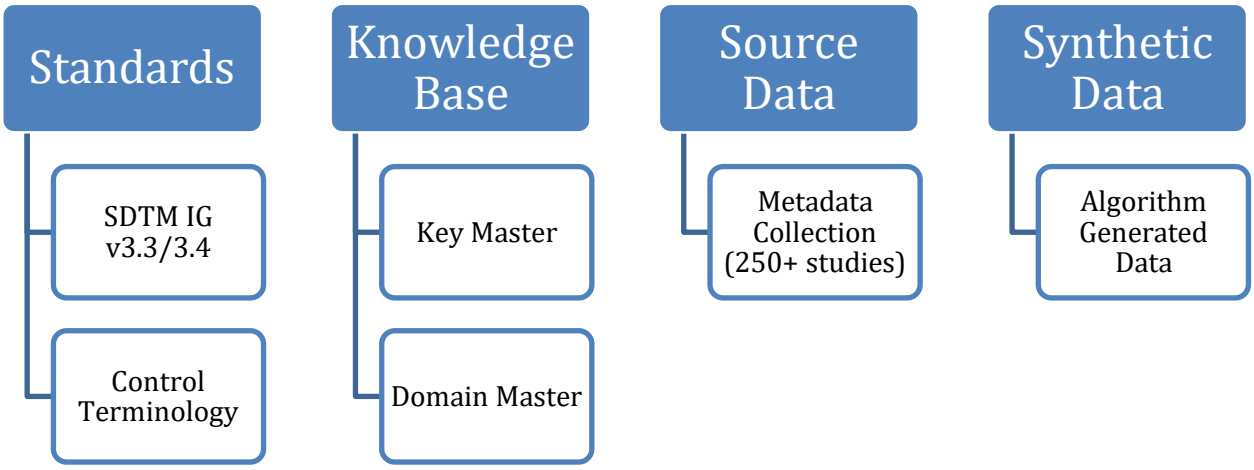

Fig. 3: Sources of training data

STANDARD DATA:

1. SDTM-IG versions 3.3/ 3.4: SDTM provides a standard for organizing and formatting data to streamline processes in collection, management, analysis and reporting. Implementing SDTM supports data aggregation and warehousing; fosters mining and reuse; facilitates sharing; helps perform due diligence and other important data review activities; and improves the regulatory

review and approval process. This is crucial data which is required for generating the Mapping Specification. In GTL task, Target variable (classification variable) is the Target Variable from SDTM-IG.

2. Control Terminology: Controlled Terminology is the set of code lists and valid values used with data items within CDISC-defined datasets. Controlled Terminology provides the values required for submission to FDA and PMDA in CDISC-compliant datasets. The collected information from study CRF, which needs to be standardized using Controlled Terminology based on the Control ID.

KNOWLEDGE BASE

1. Key Master: Knowledge base for unique and standard key information which is used for Rule based algorithms.

2. Domain Master: Knowledge base for DOMAIN mapping information which has been identified and collected based on team experience.

SOURCE META DATA:

1. Metadata collection: Metadata is defined as the information that describes and explains data. We have collected 250+ studies metadata across different Therapeutic Areas and EDC tools. The collected information is source which has been used to generate Mapping Specification (SOURCE TABLE, SOURCE VARIABLE and DESCRIPTION of VARIABLE)

SYNTHETIC DATA PREPERATION:

1. Mock data: We have created algorithm to generative mock data for every table which has been collected from Study Data. The data which is used for validation of the Auto Mapping.

END-TO-END AUTOMATED MAPPING PROCESS

The following diagram (Fig. 4) depicts the end-to-end approach used to generate mapped data sets:

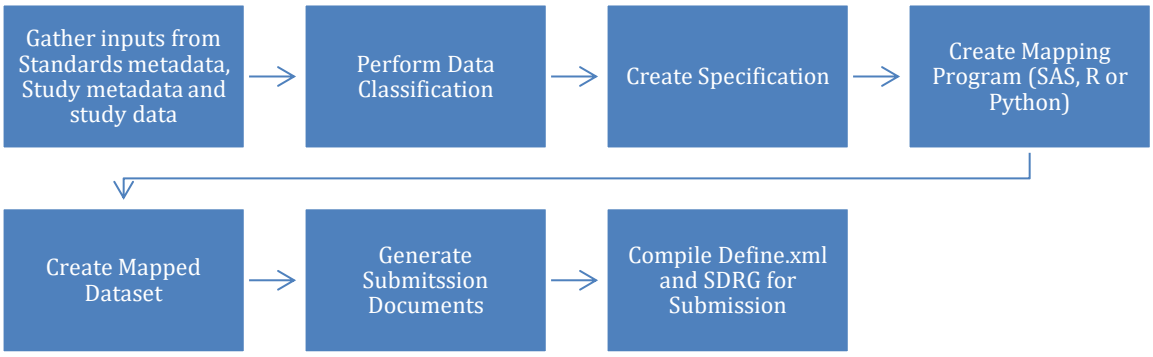

Fig. 4: End-to-End Automated Mapping Process

GROUND TRUTH LABELING:

Ground truth labeling was performed to leverage data for training, testing and validation. Key highlights of the annotation exercise are:

- Total observations - 845990

- Columns: 15
- Variables (Excluding SEQ) :
    - STUDYNAME,
    - DOMAIN,
    - VARIABLE,
    - LABEL,
    - ORIGIN,
    - KEY,
    - ROLE,
    - TYPE,
    - LENGTH,
    - FORMAT,
    - CODELIST,
    - PAGENO,
    - CDISC_NOTES,
    - CORE,
    - MAPPING_NOTES
- DOMAINS: 56
- No. of Resource who has contributed for this task: 2

DATA CLASSIFICATION

Various transformations of data expressed are currently stored in human-readable Mapping Notes in mapping specification documents. The aim of this project is to derive mapping code automatically based on a specification document. To accomplish this, existing Mapping Notes can be categorized into different types of mappings that each would have an automated function to generate code. Any derivation that falls outside of the established standard mapping types would require custom code. The intent is to expand the list and capabilities of these standard functions so an increasing number of mappings can be handled automatically.

Here is the list of all the Ground Truth label of Mapping Types:

1. Direct
2. Coalesce
3. Concatenation
4. Substring
5. Condition
6. DateDiff
7. Datetime
8. Decode
9. Imputation
10. MaxOf
11. MinOf
12. Scan
13. Sequence
14. TextInsert
15. Aggregation
16. Transpose

TARGET VARIABLE:

Review of SDTM Mapping Notes and identify what is the Source Table and Source Variable. Based on the logic and data, Source Variable must be labelled and mapped with Target Variable. Example: AEACN1 – Action taken with Study Medication of <STUDY A> - This must be mapped to AEACN of SDTM Target Variable. As per SDTM IG 3.3, we have 1700+ Target variables are available. Any source data which has been collected from eCRF, based on the study, source should be matched with SDTM target variable.

DESIGNING AND DEPLOYING ENSEMBLE MODEL

The following diagram (Fig. 5) depicts the end-to-end approach to designing and deploying the ensemble model for automapping

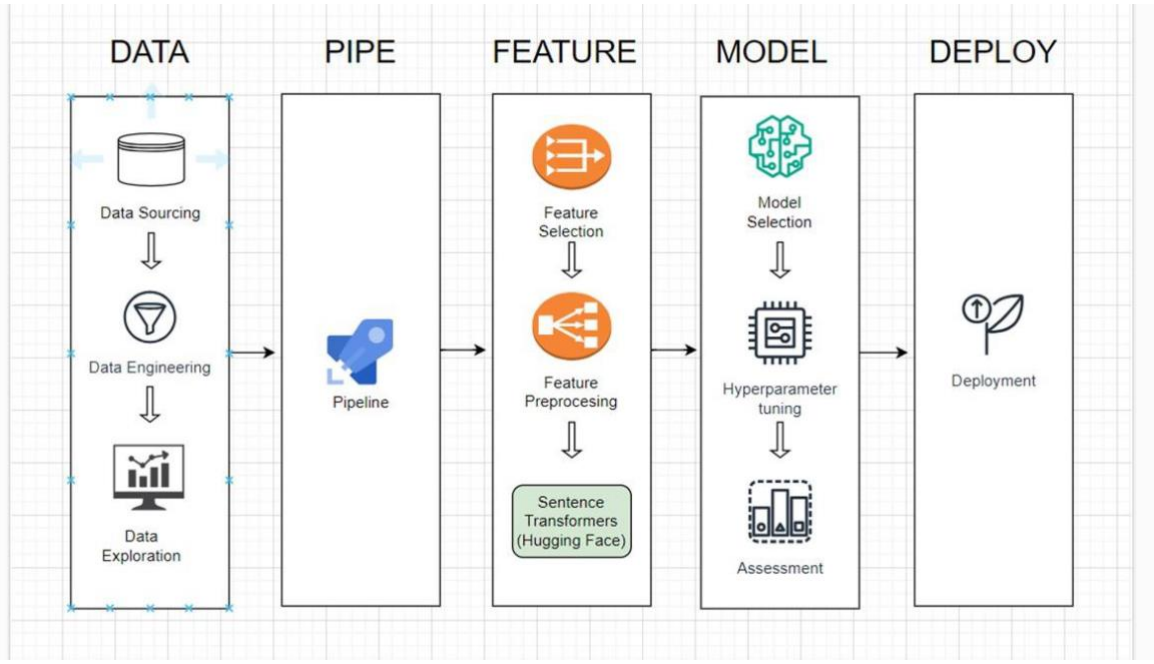

Fig. 5: Approach to designing and deploying ensemble models

BENEFITS OF ENSEMBLE MODELS FOR AUTOMAPPING

Using ensemble machine learning models for automapping clinical trial data offers several benefits that can enhance the efficiency, accuracy, and reliability of the data submission process. Here are some key advantages:

1. **Improved Accuracy**: Ensemble models combine the strengths of multiple algorithms, leading to more accurate predictions and mappings. This reduces the likelihood of errors compared to using a single model.
2. **Robustness**: By aggregating the outputs of various models, ensemble methods can handle diverse and complex datasets more effectively. This robustness ensures consistent performance across different types of clinical trial data.
3. **Reduced Overfitting**: Ensemble models, such as bagging and boosting, help mitigate overfitting by averaging out the biases of individual models. This leads to better generalization on unseen data, which is crucial for reliable data mapping.
4. **Enhanced Predictive Power**: Combining multiple models often results in superior predictive performance. This is particularly beneficial in automapping, where precise and reliable data transformations are essential.
5. **Flexibility**: Ensemble methods can integrate different types of machine learning algorithms (e.g., decision trees, neural networks, support vector machines), allowing for a more flexible and comprehensive approach to data mapping.
6. **Scalability**: Ensemble models can be scaled to handle large volumes of data, making them suitable for extensive clinical trial datasets. This scalability ensures that the automapping process remains efficient even as data size increases.
7. **Error Reduction**: By leveraging the diverse strengths of multiple models, ensemble methods can identify and correct errors that might be overlooked by individual models. This leads to higher data quality and integrity.

8. **Adaptability**: Ensemble models can be continuously updated and retrained with new data, ensuring that the automapping process evolves and improves over time. This adaptability is crucial in the dynamic field of clinical trials.

Overall, ensemble machine learning models provide a powerful and reliable solution for automapping clinical trial data, enhancing the overall efficiency and accuracy of the data submission process.

MODELING APPROACH

We have used Ensemble with Majority Voting Classifier with Sentence Transformation. The final prediction is determined by the class label that receives the most votes from the individual models. In case of a tie, the algorithm can employ various tie-breaking strategies, such as selecting the first or last prediction. Some of the optimal parameters which are used for classification include, but not limited to, learning_rate , n_estimators, max_features, max_depth, min_samples_split, random_state etc..

ENSEMBLE MODELS SELECTED:

The list below depicts the models considered for our ensemble combinations

1. Logistic Regression (LR):
   Description: A statistical model used for binary classification tasks. It predicts the probability of a binary outcome (e.g., yes/no, true/false) based on one or more predictor variables.

   Use Case: Commonly used in medical diagnosis, credit scoring, and marketing.

2. Random Forest (RF):
   Description: An ensemble learning method that constructs multiple decision trees during training and outputs the mode of the classes (classification) or mean prediction (regression) of the individual trees.

   Use Case: Used in various applications like fraud detection, disease prediction, and stock market analysis.

3. Support Vector Machine (SVM):
   Description: A supervised learning model that finds the optimal hyperplane which maximizes the margin between different classes in the feature space.

   Use Case: Effective in high-dimensional spaces and used in text classification, image recognition, and bioinformatics.

4. K-nearest Neighbor (KNN):
   Description: A simple, instance-based learning algorithm that classifies a data point based on how its neighbors are classified. It uses the majority vote of its k-nearest neighbors.

   Use Case: Used in recommendation systems, pattern recognition, and intrusion detection.

5. Multi-Layer Perceptron (MLP):
   Description: A type of artificial neural network with one or more hidden layers between the input and output layers. It uses backpropagation for training.

   Use Case: Applied in speech recognition, image classification, and machine translation.

6. Extra Trees (ET):
   Description: An ensemble learning method similar to Random Forest but with more randomness. It builds multiple trees using random splits of all observations and features.

   Use Case: Used in regression and classification tasks where high variance is a concern.

7. Decision Tree (DT):
   Description: A tree-like model used for decision making. It splits the data into subsets based on the value of input features, creating branches until a decision is made.

   Use Case: Used in customer segmentation, risk analysis, and medical diagnosis.

8. Bernoulli-Naïve Bayes:
   Description: A variant of the Naïve Bayes classifier for binary/boolean features. It assumes that the presence of a particular feature is independent of the presence of any other feature.

   Use Case: Commonly used in text classification tasks like spam detection and sentiment analysis.

ENSEMBLE MODEL COMPARISION

The following table (Table 2) summarizes the ensembles used and the corresponding outcomes:

1. Ensemble A: 8 Models (Logistic Regression, Random Forest, Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Multi-Layer Perceptron, Extra Trees, Decision Tree and Bernoulli Naïve Bayes)
2. Ensemble B: 5 Models (Logistic Regression, Random Forest, SVM, KNN, Multi-Layer Perceptron)
3. Ensemble C: 3 Models (Random Forest, SVM and Multi-Layer Perceptron)
4. Ensemble D: 5 Models with 16 features
5. Ensemble E: 4 Models with 5 features (As per F1 Score, this Model has been selected)

	Ensemble A	Ensemble B	Ensemble C	Ensemble D	Ensemble E
Description and estimators	8 Models (LR, RF, SVM, KNN, MLP, ET, DT and BNB)	5 Models (LR, RF, SVM, KNN, MLP)	3 Models (RF, SVM, MLP)	5 Models (LR, RF, SVM, KNN, MLP)	4 Models (LR, RF, SVM, MLP)
Features	5	5	5	16	5
Accuracy	0.862	0.887	0.889	0.875	88.86
F1 Score	0.606	0.662	0.655	0.626	0.787
Execution Time for (50 sample size)	30 to 40 mins	8 mins	5 mins	15 mins	Below 15 secs
Precision	0.638	0.679	0.623	0.680	0.807
Recall	0.603	0.668	0.655	0.660	0.783
Comments	Since this model takes time to load and execution, this has to be eliminated.	This model execution time is 8 mins , score is decent and 5 set of estimators.	This model execution time is 5 mins and score is decent. However it has only 3 estimators.	This model execution time is 15 mins , score is below expectations. So this has to be retired.	Fine tuning the code to execute faster
Next Steps	Eliminated	1. Adding more corpus to improve the scores. 2. Finetune Hyper Parameters 3. Adding more features which are appropriate to model	Eliminated	Eliminated	1. Finetune of Hyper parameters 2. Adding more corpus for including control values

Table 2: Ensemble of models used

CONCLUSION

The automation of specification generation and data mapping in clinical trial data submission represents a significant advancement in the field of clinical research. By leveraging ensemble machine learning models, we can address many of the challenges inherent in this process, such as data quality, regulatory compliance, and efficiency.

Ensemble methods, with their ability to combine the strengths of multiple algorithms, offer a robust and flexible solution that enhances the accuracy and reliability of data mapping. These models not only improve predictive performance but also provide a scalable and adaptable framework that can evolve with the dynamic needs of clinical trials.

The implementation of these advanced techniques can lead to substantial reductions in manual effort and error rates, ultimately accelerating the timeline for clinical trial submissions and ensuring higher data integrity. As the industry continues to embrace digital transformation, the integration of ensemble machine learning models into clinical data management processes will be pivotal in driving innovation and improving outcomes.

In conclusion, the adoption of ensemble machine learning models for automating clinical trial data mapping is a promising step towards more efficient, accurate, and reliable data submissions. This approach not only streamlines the submission process but also sets the stage for future advancements in clinical research, paving the way for more effective and timely medical interventions.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Author Name: Venugopal Mallarapu
Company: eClinical Solutions
Address:603 West St., Mansfield MA, USA
Work Phone: +1 919 599 3012
Email: vmallarapu@eclinicalsol.com
Website: https://www.eclinicalsol.com