

*Presentation01: GOOGLE AI LANDSCAPE*

Assignment Title

**Student Name: KARTHIK THALLAM**

**Student ID: 500188370**

**Course Code: AISC 1005**

**Instructor Name: Alaaaldin Alradhi**

By submitting this assignment, you confirm that you alone have contributed to this submission. Any suspicion of copying or plagiarism in this work will result in an investigation of Academic Misconduct and may result in a “0” on the work, an “0" in the course, or possibly more severe penalties as well as a Disciplinary Notice on your academic record under the Student Code of Academic Conduct, which can be found online at:<https://www.loyalistcollege.com/about-loyalist/policies/aop-216-academic-honesty/>

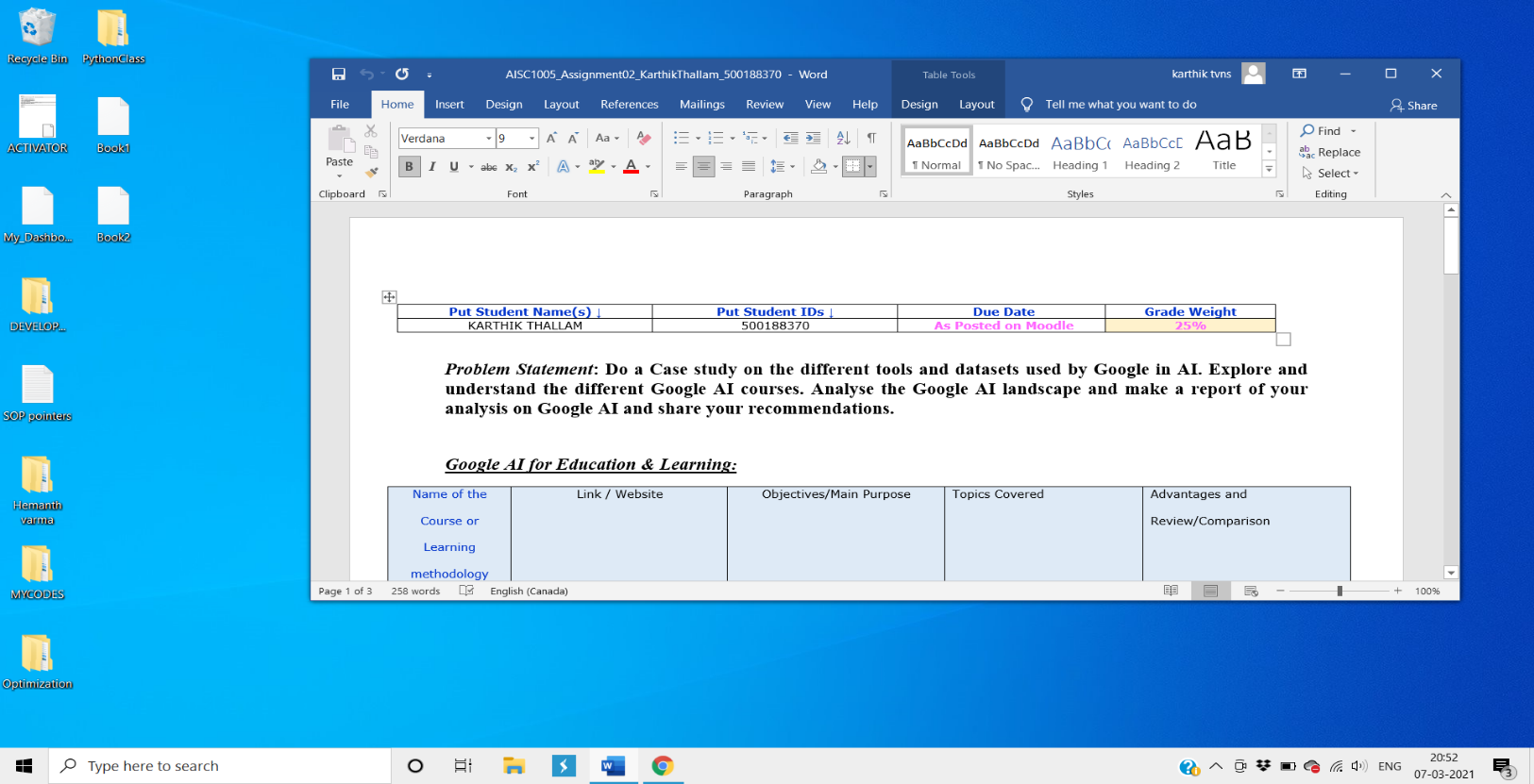
|  |  |
| --- | --- |
| Is this my best effort? |  |
| Has the spelling been checked? |  |
| Is the references section correct? |  |
| Is all cited material referenced? |  |
| Are there any materials that you have not cited? (and should therefore be expunged). |  |
| Is the word length OK? |  |
| Have I checked for spelling and grammar? |  |
| Is your name, student number, instructor’s name and course code on the title page? |  |
| Could someone read my introduction and have a rough idea of the body? |  |
| Could someone read my conclusion and know what they would have learn had they read the body? |  |
| Were any of my arguments unsubstantiated or biased? |  |
| Did I leave the essay for a few hours and then read it again from a more objective perspective? |  |
| Did I check the whole checklist? |  |

**Assignment Submission Checklist:**

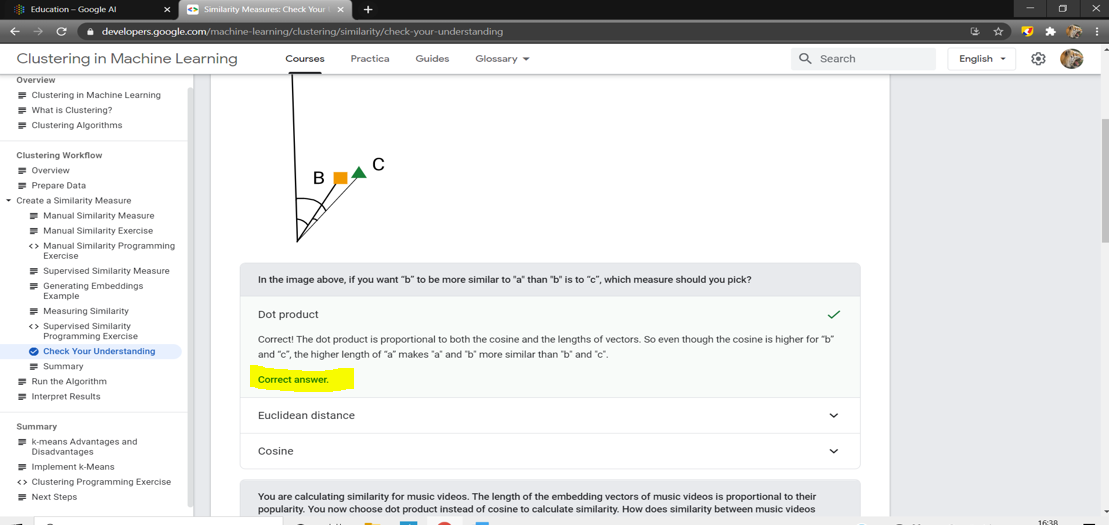
***Google AI for Education & Learning:***

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Name of the Course or Learning methodology | Link / Website | Objectives/Main Purpose | Topics Covered | Advantages and Recommendations/Review |
| 1) AI for Social Good | <https://ai.google/social-good> | * Use of AI in the challenges faced by the society and people * To predict natural disasters and track contagious disease * To help people with disabilities | * AI Impact challenge * Machine Learning Basics * Forecasting Floods accurately * Open Covid-19 data repository * AI4SG Google Research India | * Used for solving humanitarian and natural challenges * This is one of the best challenging programs created across the industries as it can allow organizations to submit their ideas on how to use AI in facing the environment challenges * By collecting data about environmental challenges, we can improve the efficiency in handling and predicting the natural calamities |
| 2) Federated learning | <https://federated.withgoogle.com/#learn>  <https://medium.com/@ODSC/what-is-federated-learning-99c7fc9bc4f5> | * Building better on device models with security as the default priority | * FL out of vocabulary words * FL in Google products * Secure aggregation * FL Architecture * FL theory | * Using local data, it downloads the on going model and validates the updated model on the device itself * This is one of the best Edge computing techniques used by Google * Better security and privacy |
| 3) People AI Guidebook | <https://pair.withgoogle.com/guidebook> | * To help professionals working in UX and product managers to use AI in a human based approach | * Data collection and Evaluation * User Needs and Defining success * Mental Models * Errors and dignity failures * Feedback control | * This is one of the best data related guide books as it is based on inferences from various experts in the industry and research teams * It will guide you completely on the flow of product development |
| 4) AI explorables | <https://pair.withgoogle.com/explorables/> | * To know exactly how machine learning models work * To find whether or not the predictions made by the machine learning models are fair enough * To find the ways of training the model in the best possible way | * Reason for data leakage and how to fix it? * Handling sensitive information with some standard privacy policies * Underlying facts about biased model and biased data * Evaluating the model performance | * One of the best ways to find how the model is understanding the problem statement * AI explorables focuses mainly on handling the sensitive information which is often ignored by the developers and organizations when using huge amount of data as it is difficult to manage * It also finds a way to check the model bias |
| 5) Clustering | <https://developers.google.com/machine-learning/clustering> | * Clustering techniques are used to segment the data into groups when there is no output label provided * Since there is no output label for the categories of the data, it is called as Unsupervised machine learning model | * Different Clustering algorithms * Clustering the data based on the iterations of centroid positions * Clustering the data based on Gaussian distributions * Hierarchical clustering * K-Means Clustering * Interpreting the results based on quality of clustering, measuring the similarity performance and optimal number of clusters | * Used in categorising the people when there is no previous evidence or information based on distance or similarities between the observations of the data using multiple techniques like centroid positions of the categories etc. * K-means is the most popular clustering algorithm used in real world * DBSCAN is used to get the clusters perfectly and it is one of the clearest ways of clustering the data |
| 6) Recommendation Systems | <https://developers.google.com/machine-learning/recommendation/> | * Different techniques and models used for developing recommendation systems including cosine similarity and deep learning networks | * Home page vs Similar items recommendations * Co-sine Similarity Recommendations * Content similarity filtering * Collaborative Filtering * Building Recommendations systems using Neural Networks | * It shows how to use TensorFlow in order to develop recommendation systems * It demonstrates many ways of developing recommendation systems like SoftMax, Matrix factorization, content-based ad collaborative filtering systems * It will explain the underlying components like candidate generation, scoring and re-ranking |
| 7) Testing and Debugging in Machine Learning | <https://developers.google.com/machine-learning/testing-debugging/> | * It describes how to debug our machine learning model all the way to the monitoring our machine learning pipeline in the deployment | * Debugging the model ensures to increase the model performance by optimizing it in various no. of ways * Data debugging like feature selection, Hyper-parameter tuning, Abnormal observations and Feature Engineering * Model Optimization * Validating the model for deployment * Testing the model in production | * It includes the various no. of ways to increase the model performance both in development and deployment stages * Key chapters like model debugging, optimization, pipeline management and monitoring the model will help us to know the complete cycle of a machine learning model development |
| 8) Introduction to Machine Learning Problem Framing | <https://developers.google.com/machine-learning/problem-framing/?utm_source=googleAI&utm_medium=card-image&utm_campaign=training-hub&utm_term&utm_content=problem-framing> | * Helps to frame a machine learning model on the given problem statement | * What are the different types of machine learning models? * Supervised vs unsupervised learning * Clustering * Anomaly detection * Which machine learning model to choose for the given problem statement | * Contains exercises for each category of machine learning * It also contains several uses cases and instance4s for using the particular type of machine learning model over the other * Includes a clear description of general methods of Machine learning problem solving approach |
| 9) Data preparation and Feature Engineering in Machine Learning | <https://developers.google.com/machine-learning/data-prep/?utm_source=googleAI&utm_medium=card-image&utm_campaign=training-hub&utm_term&utm_content=data-prep> | * In order for the model to work efficiently we have to pass the data which is clean enough to understand the underlying patterns in it * So, this course contains various data cleaning and data transformation methods to use it later for building a machine learning model | * Various methods of collecting the data * How to handle imbalanced data that could lead to biased predictions by the machine learning model especially when we are dealing with classification of the output variable * Splitting the data into training and testing * Data transformation techniques like Categorical Encoding techniques | * It gives the detailed analysis and importance of data preparation on order for the model to work more accurately * It explains all the major steps involved in data handling like Categorical Encoding, Feature transformations and Splitting the data before the model to be trained on this data |
| 10) Machine Learning Guides | <https://developers.google.com/machine-learning/guides> | * Step-by-Step explanation and common practices involved in machine learning using different case studies for each type of problem statement | * People + AI guidebook which contains the detailed explanation of different tools and methods used for developing the machine learning models * Rules of ML which contains the set of rules or procedure one should follow while building the model * Text Classification * Good data analysis which explains the various tricks that can be applied on huge datasets to find the insights from the data | * Rules of Machine learning is one of the most useful guides across the web as it contains the detailed explanation of machine learning phases like Pipelines, Feature Engineering and Optimization of the models * Good data analysis is the guide where we can find all the important tricks that can be applied on the datasets in order to visualize and analyse the data in the best possible way |
| 11) Machine Learning Crash Course | <https://developers.google.com/machine-learning/crash-course/ml-intro> | * A practical self study guide for ML aspirants which explains all the concepts of machine learning using TensorFlow | * Framing a Machine Learning Problem * Checking how well the model generalized on the data leaving no space for bias * Testing the model performance by validating on the validation part of the dataset and followed by final testing on testing part of the dataset * Regularizing thereby reducing the bias and variance of the model * Neural Networks * Real world uses cases of ML systems | * Every section has detailed explanation about various steps and concepts involved in Machine Learning cycle * Multiple use cases have been used to explain the process more clearly * All the major machine learning terms have been covered and some of them are A/B testing, accuracy and Area Under the Curve (AUC) * In deep learning section of the course, it shows how the back propagation works along with the batch normalization * In Natural Language Processing, it describes the usage of Bag of words * It also explains the Bayesian Optimization technique especially when we are using deep learning models * One of the fascinating sections of the course is explanation about handling imbalanced dataset in the classification problems there by using down sampling or SMOTE techniques to solve this issue |
| 12) Fairness | <https://developers.google.com/machine-learning/crash-course/fairness/video-lecture> | * To check how fairly the model is performing * Evaluating the model bias * To identify the model bias and different strategies to remove this bias in order for the model to generalize effectively | * Different types of bias * Different strategies to find the model bias * Validating and evaluating the impact of bias on the new data | * This fairness course is very much useful especially when we are building a machine learning model * The model performance is directly proportional to how good the model is generalizing the data * If there is any bias in the model then the predictions or inferences which we get form the model are not accurate and often misunderstood |
| 13) Responsible AI practices | <https://ai.google/responsibilities/responsible-ai-practices/> | * A brief description about current AI practices using which google has been trying to improve the lives of the people across the globe in each and every industry they have been involved into | * Responsible AI with TensorFlow * Quality of search in rating guidelines * PAIR + AI platform which constantly publishes the necessary insights and approaches in order to design the AI models * Different machine learning model cards toolkit | * All these practices in AI at google benefits the machine learning community when building a most reliable and powerful yet efficient machine learning models * They have created a unique human centered approach in designing this AI models and their methodologies * Using different metrics in order to evaluate the model thereby boosting the performance of the model accordingly |
| 14) Seedbank – Machine Learning Discovery | <https://aihub.cloud.google.com/u/0/s?category=notebook> | * To provide a platform for ML aspirants and give them a hands-on experience on various real world use cases along with datasets and codes of the projects | * It has over 655000 datasets and over 170 coded project notebooks in order to make the rookies in ML to understand the different use cases of Machine learning | * One of the best resources to kick-start careers in machine learning is AI-HUB by google * It has over 655 thousand datasets and over 170 sample projects covering all the niche skills of Artificial Intelligence like NLP, ML, and DL |
| 15) Machine Learning Recipes with Josh Gordon | <https://www.youtube.com/playlist?list=PLOU2XLYxmsIIuiBfYad6rFYQU_jL2ryal> | * These are short YouTube videos taught by Josh Gordon which gives a gist of important phases of machine learning development cycle | * Visualizing Decision Trees * Finding the best features and the underlying reasons * Machine learning pipelines * Sample Classification project on Hand-written digit classification | * You will understand how to train and build a classification model from scratch using TensorFlow * It also explains the several feature engineering techniques that can be used to find the underlying information in each of the features available in our dataset |
| 16) TensorFlow workshops | <https://github.com/tensorflow/workshops> | * A hands-on workshop platform where we can find the different use cases of Machine learning projects like Text Classification and Text Analytics along with Jupyter Notebooks | * TensorFlow pipelines for creating Text classification problems * Distributed Principal Component Analysis in TensorFlow * Text Pipeline for BERT processing TensorFlow model | * This Hands-On workshop platform will give us a brief idea about how to use TensorFlow for developing machine learning problems * There are multiple use cases and also contains different packages already installed in TensorFlow so that we can start building the projects in TF right away after data acquisition |
| 17) Machine Learning Glossary | <https://developers.google.com/machine-learning/glossary?utm_source=google-ai&utm_medium=card-image&utm_campaign=training-hub&utm_content=ml-glossary> | * A glossary showing us all the different terms used in Machine Learning and some of the terms which are very specific to TensorFlow | * All the data pre-processing terms including Scaling, Normalization, Categorical Encoding and Feature Selection * Different Data Analysis terms and evaluation metrics like MSE, RMSE, MAE and R2 score * Different Deep learning terminologies like Batches, epochs, call-backs and Back propagation | * This glossary will be really helpful to know about all the methods involved in machine learning * There is a detailed explanation to all the terms used in machine learning along with accurate instances to understand them clearly |
| 18) Deep Playground | <http://playground.tensorflow.org/#activation=tanh&batchSize=10&dataset=circle&regDataset=reg-plane&learningRate=0.03&regularizationRate=0&noise=0&networkShape=4,2&seed=0.57728&showTestData=false&discretize=false&percTrainData=50&x=true&y=true&xTimesY=false&xSquared=false&ySquared=false&cosX=false&sinX=false&cosY=false&sinY=false&collectStats=false&problem=classification&initZero=false&hideText=false> | * To see how the deep learning model inferences change with a change in some of the parameters like activation function, learning rate etc. | * We can see how the weights of hidden layers change with a change in the activation functions and various other parameters involved in a Neural-Network * There are different options we can change according to our requirement to see the model interaction with the parameters | * This is just a browser tool but has huge advantage in understanding the several underlying factors involved in a deep learning neural network * We can change the activation function and click on the nodes to highlight the process of weights and bias evolution from that particular node to the next hidden layers or output layer |
| 19) AI adventures: 7 steps of Machine Learning | <https://www.youtube.com/watch?v=nKW8Ndu7Mjw&list=PLMUdwhLW7A0icUPn6LT5Zj0rDqZ2OOP8Y&index=4> | * A short YouTube video which explains all the 7 major steps of machine learning development from the scratch | * Data Acquisition * Data preparation * Model selection * Model Training * Model Evaluation * Tuning the Hyper parameters * Predictions on the New data | * One should have a clear understanding of all the 7 major steps involved in a machine learning model development * This video will tell us about all the steps which is very essential before we start building any machine learning model |

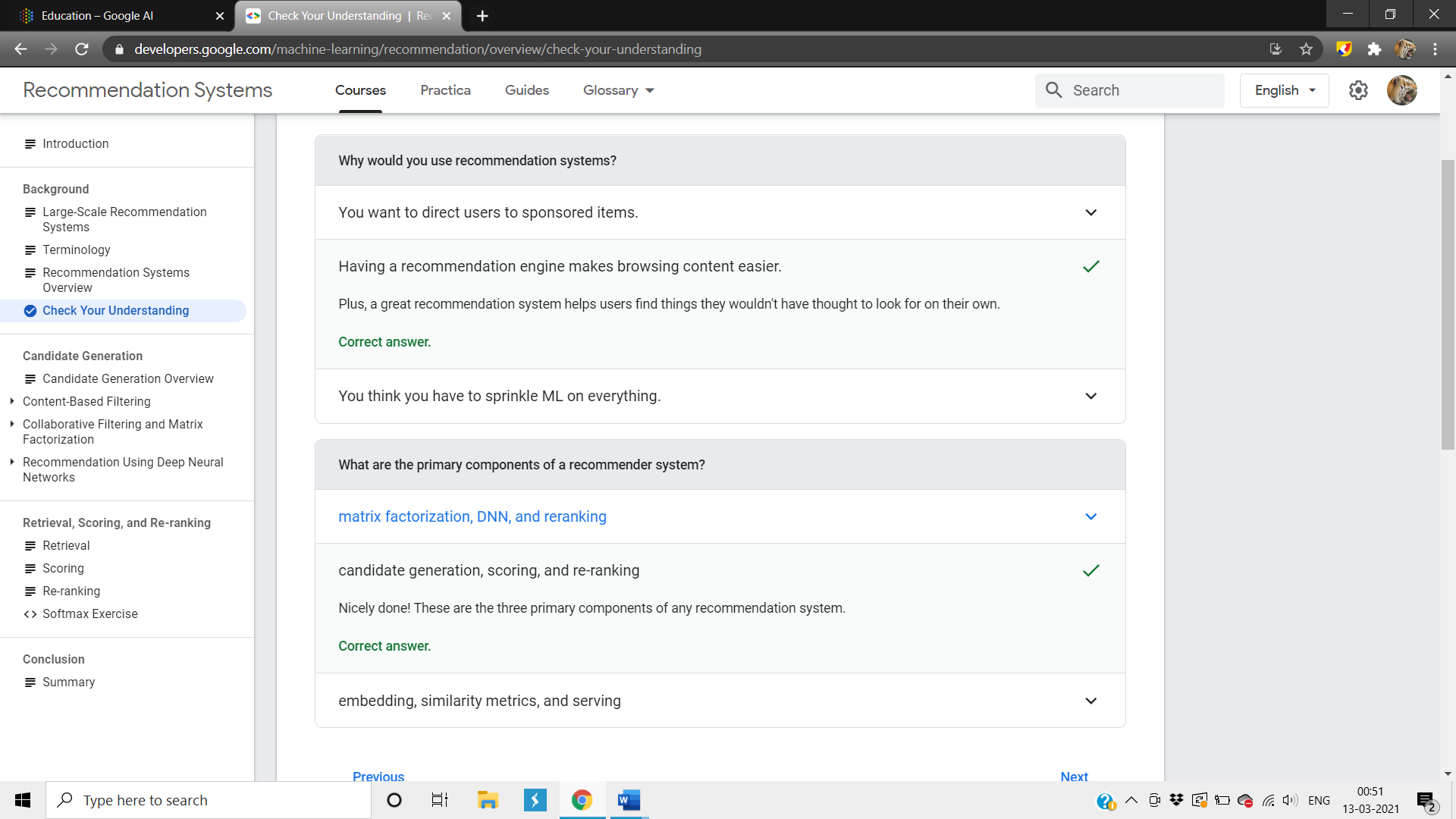
***Proof of Concept (POC): On my Machine***

******

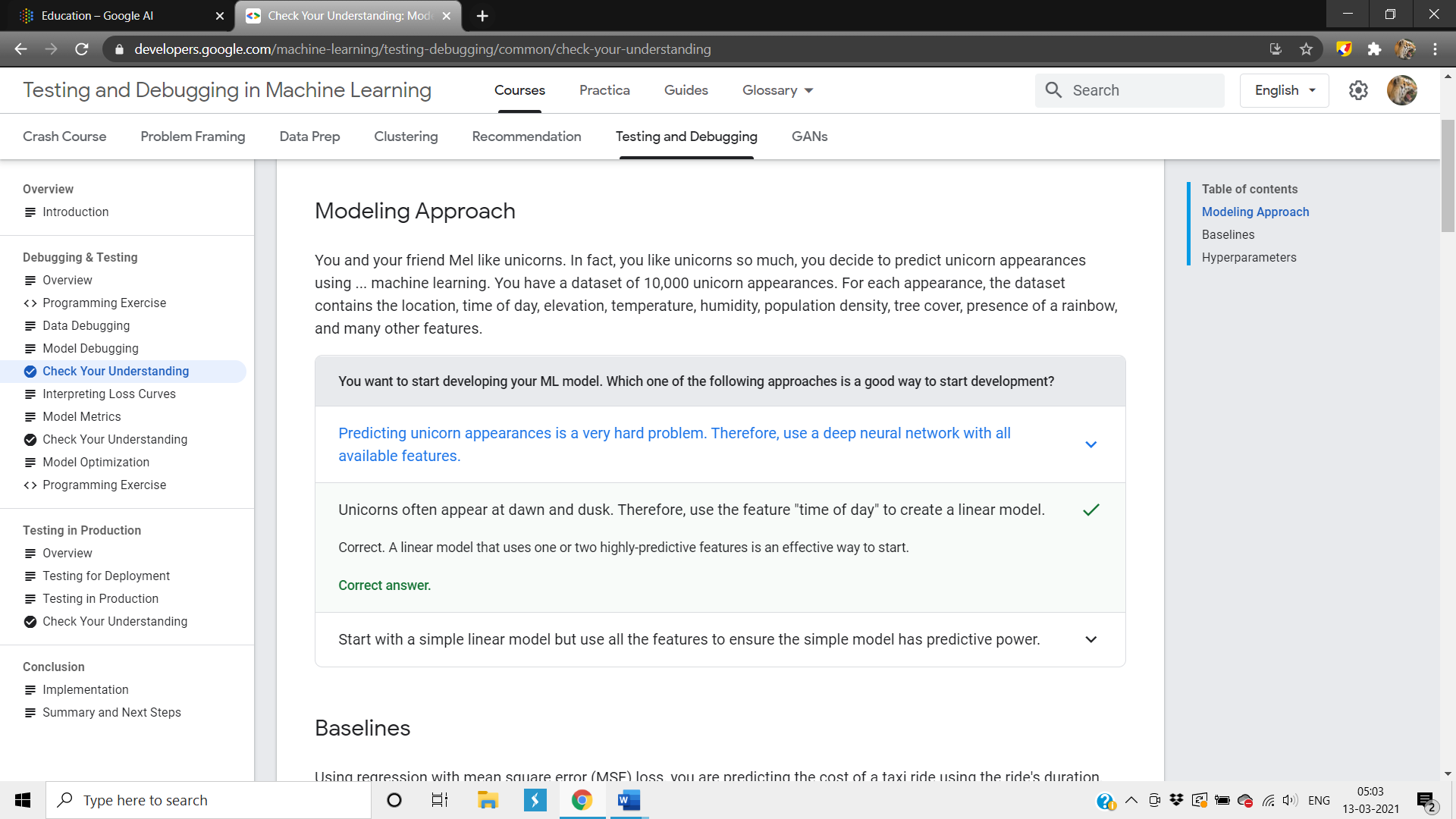
***POC for Clustering:***

******

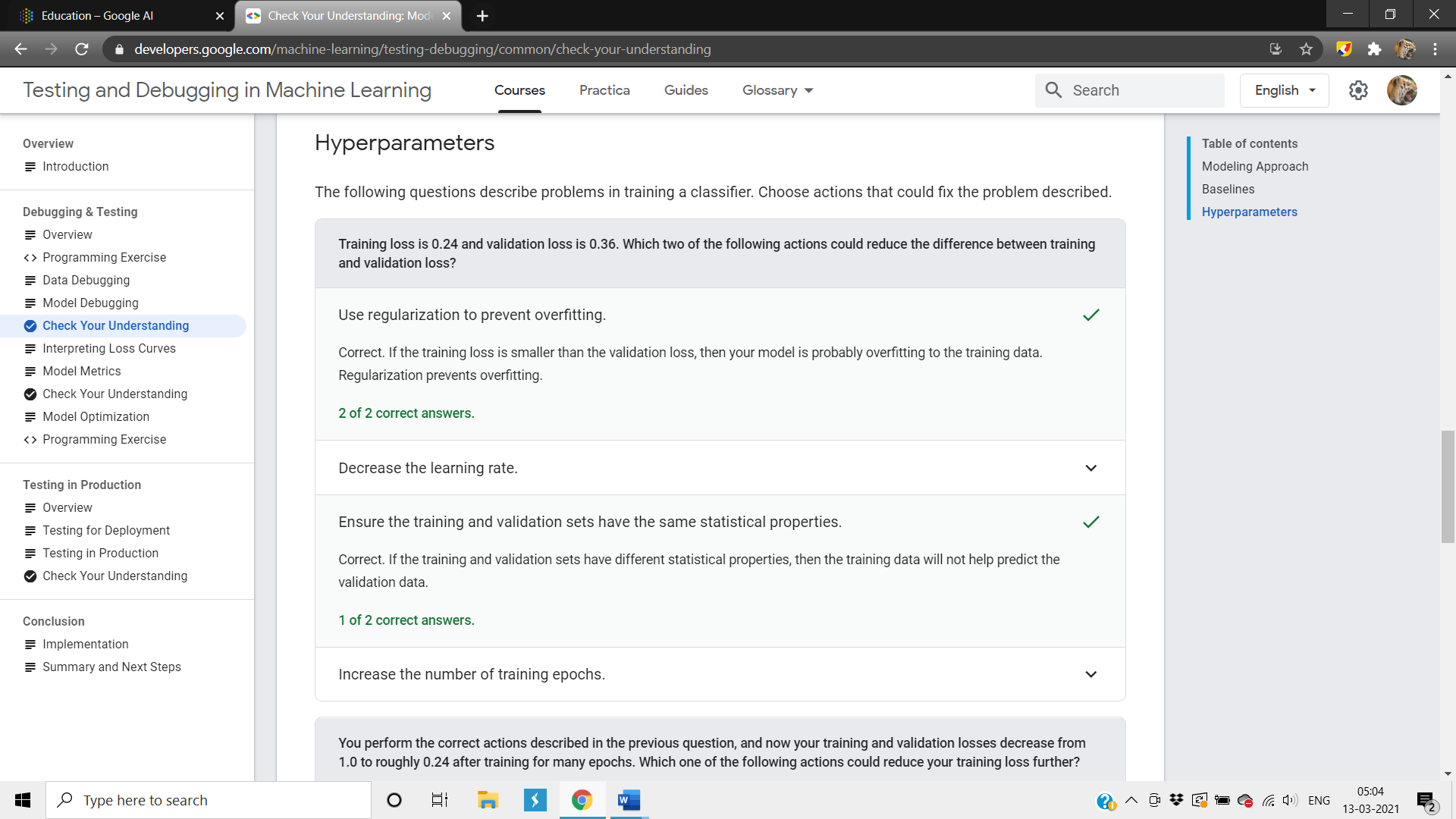
***POC for Recommendation systems:***

******

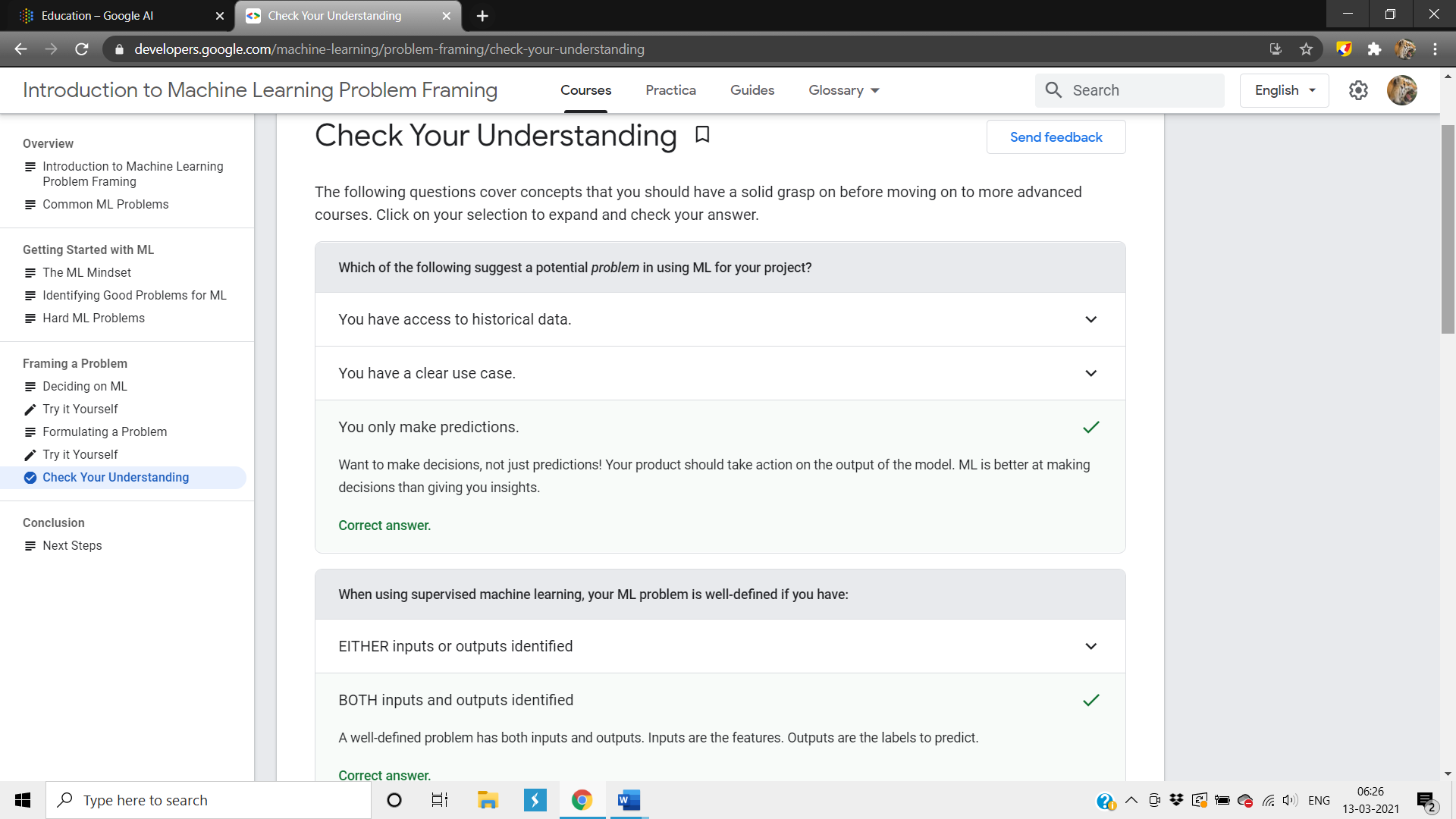
***POC for testing and debugging in ML:***

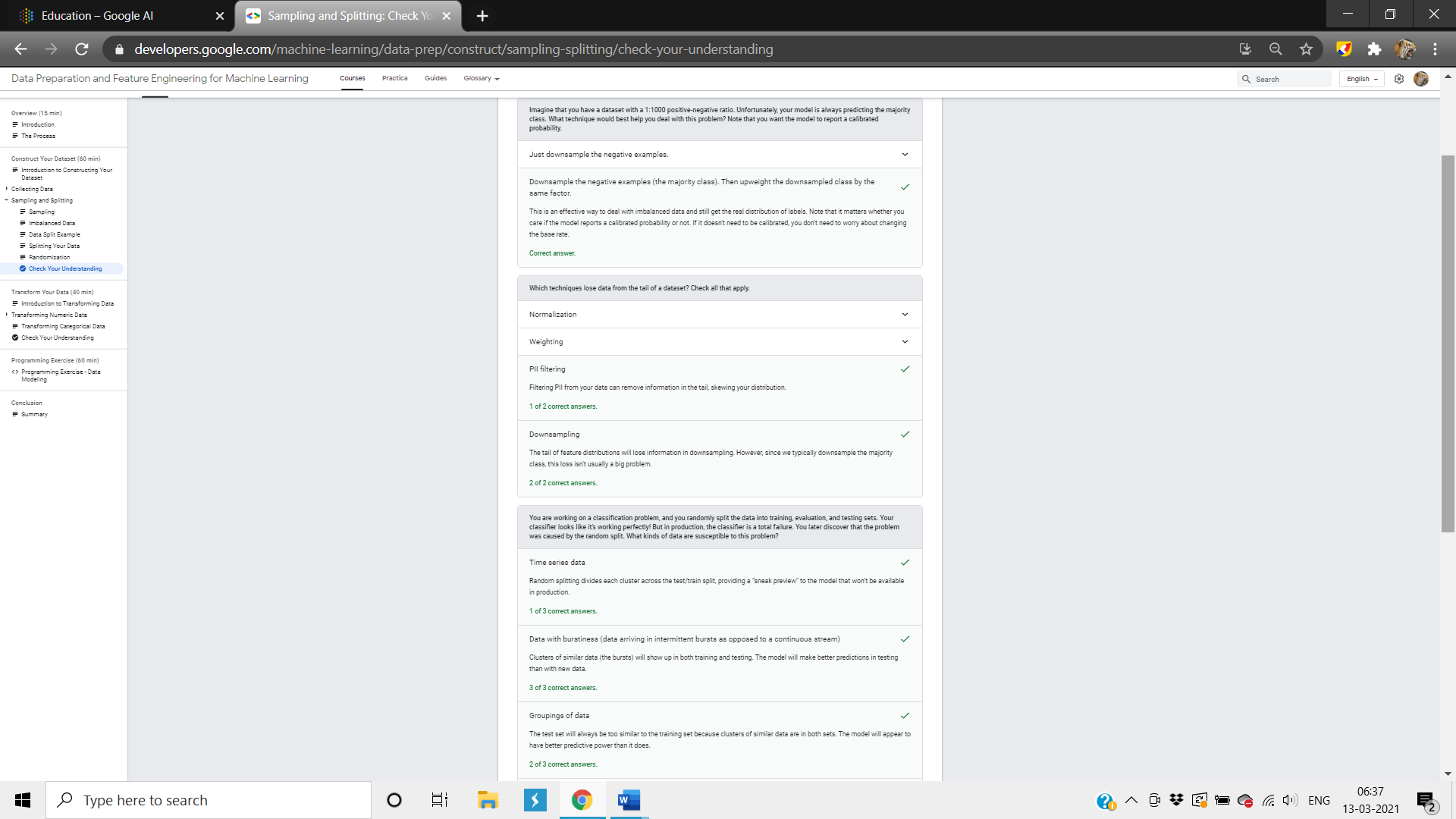
******

***POC for Hyper parameters:***

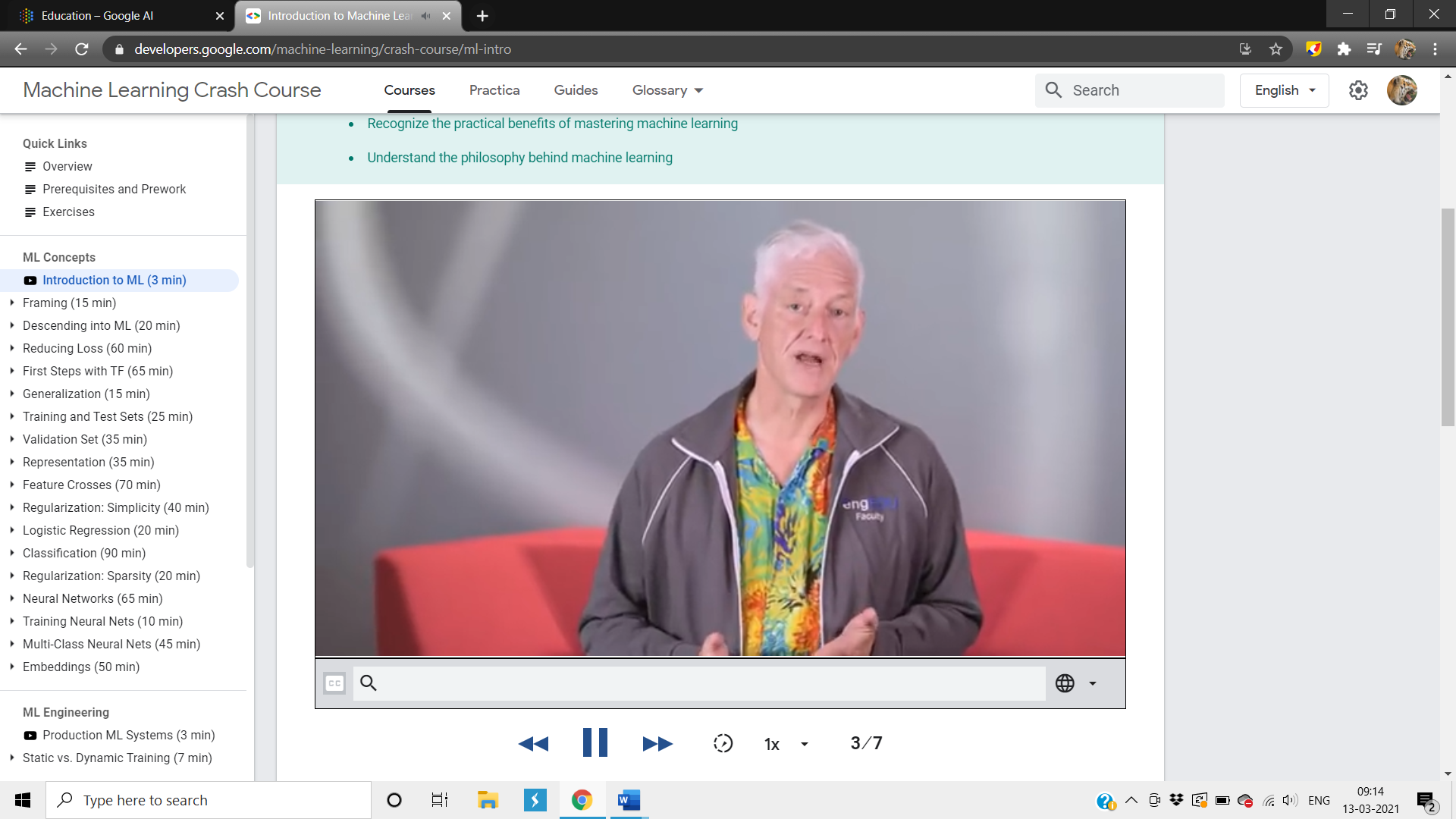
******

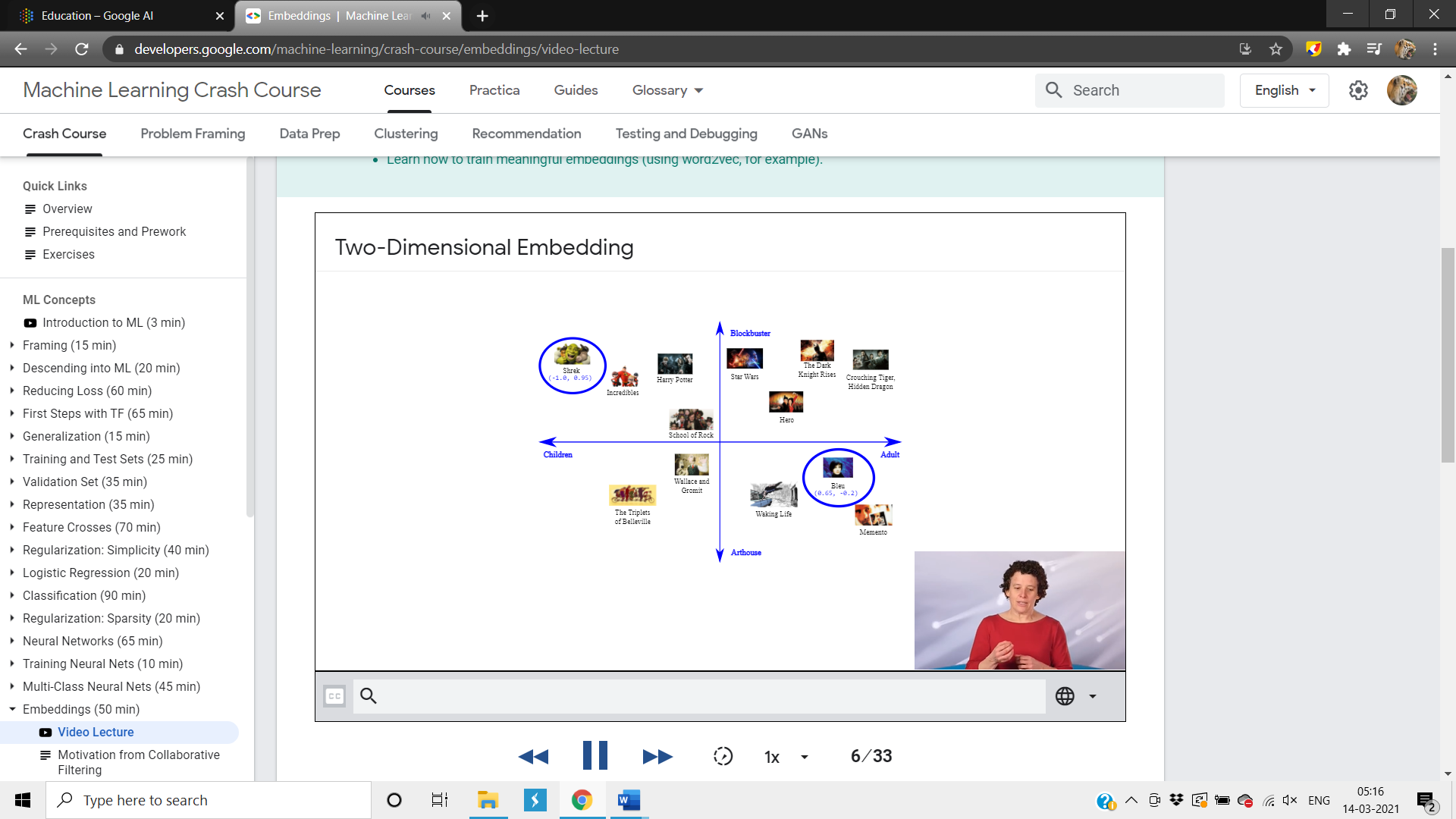
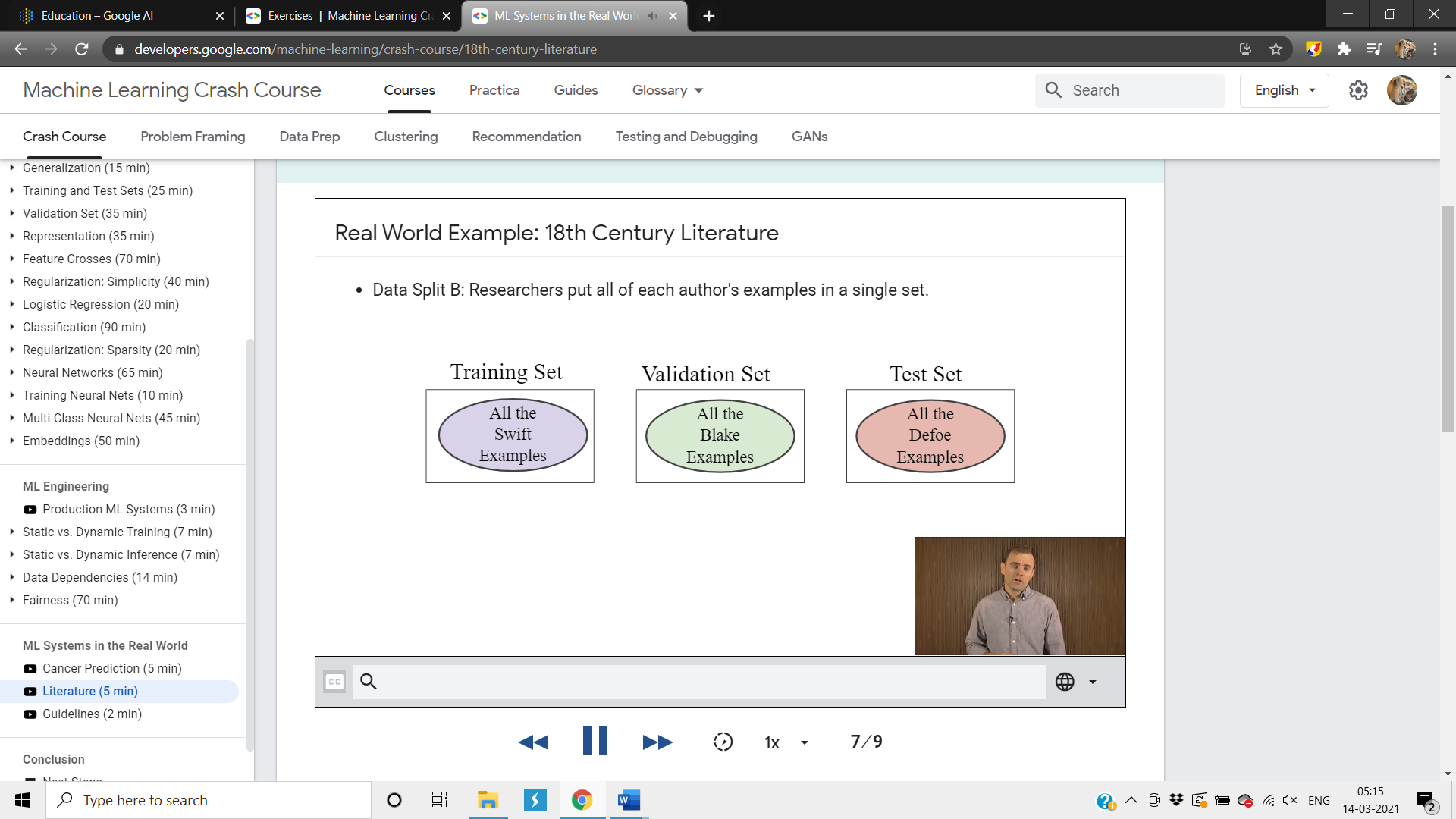
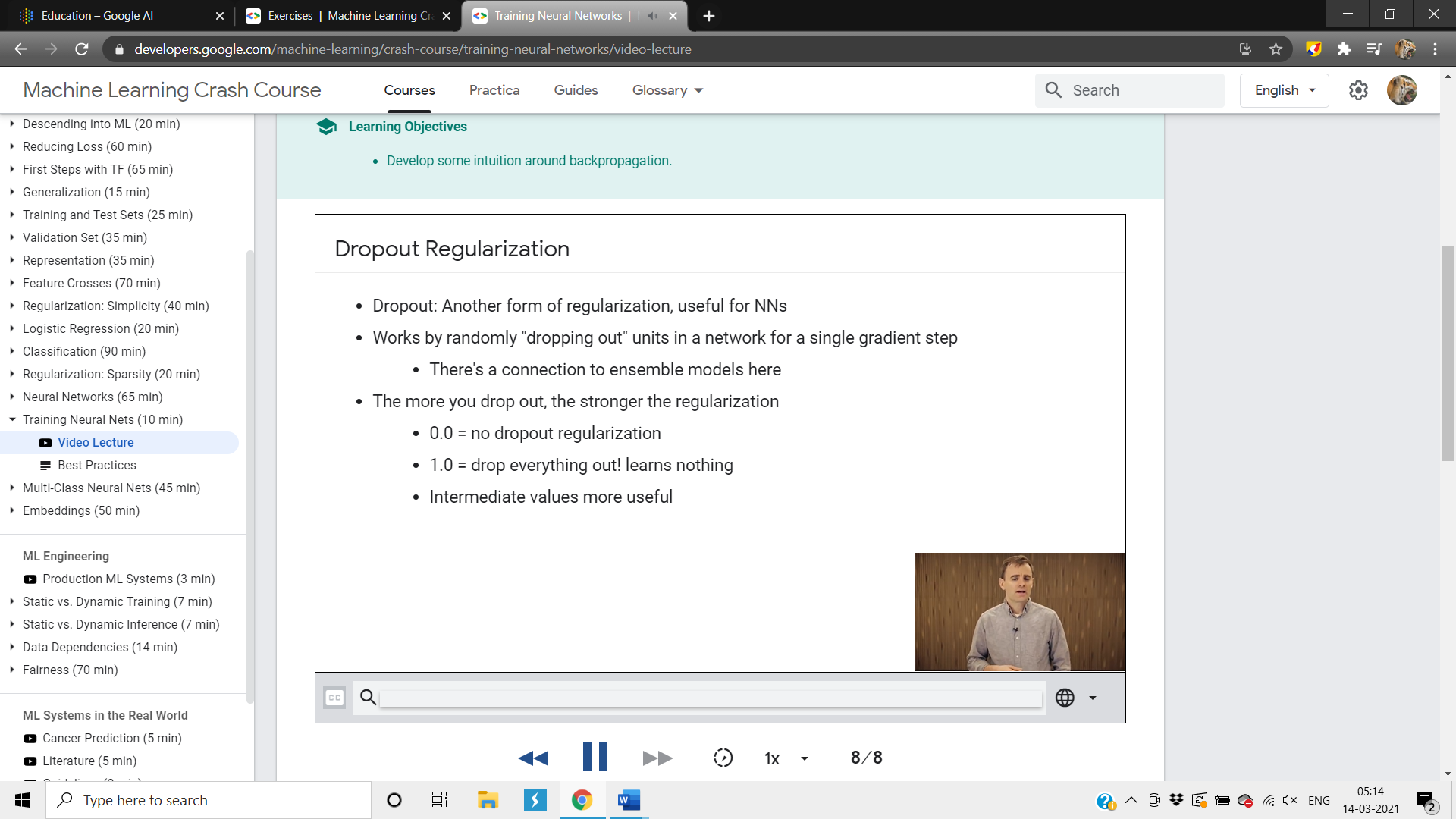
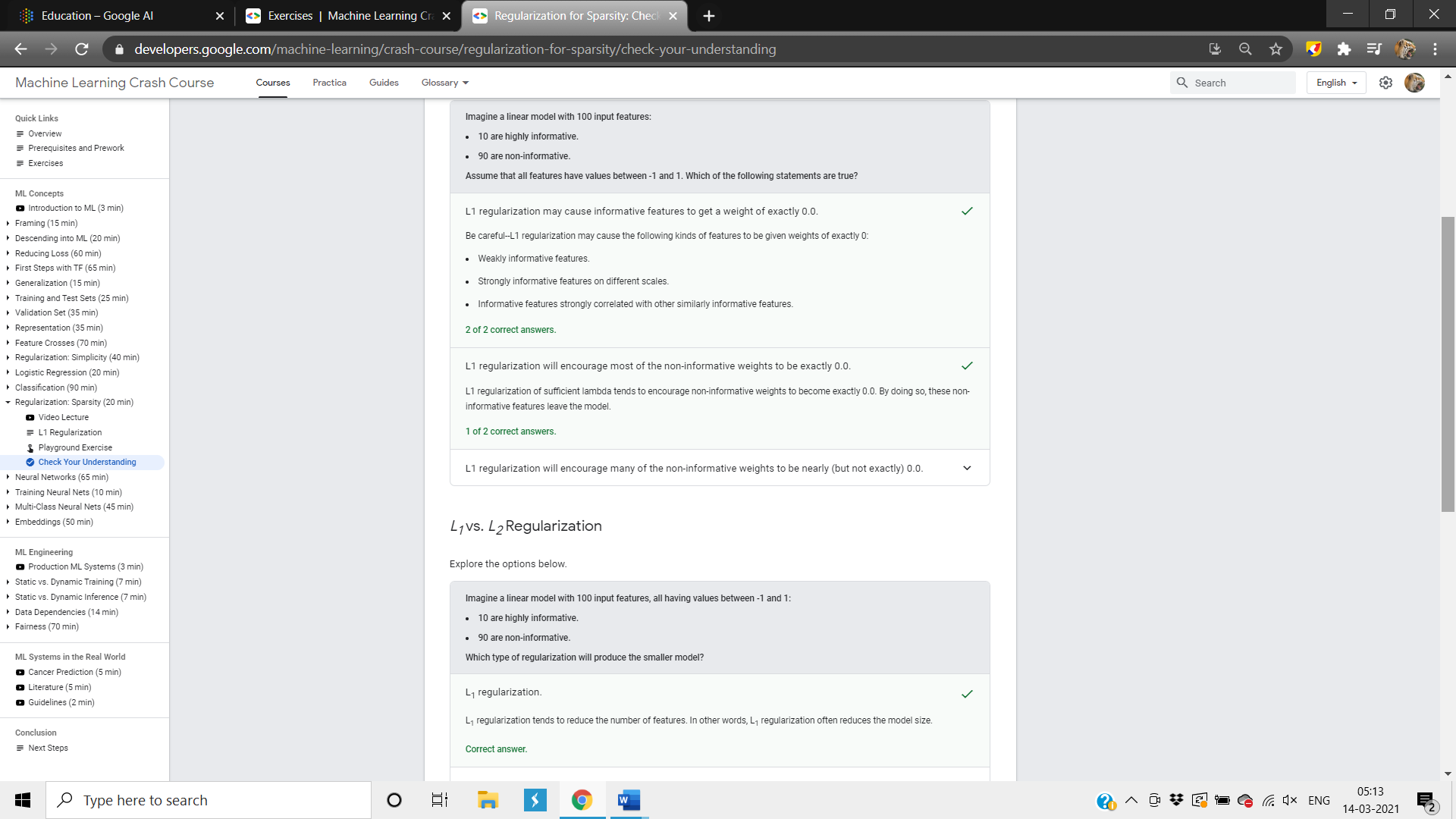
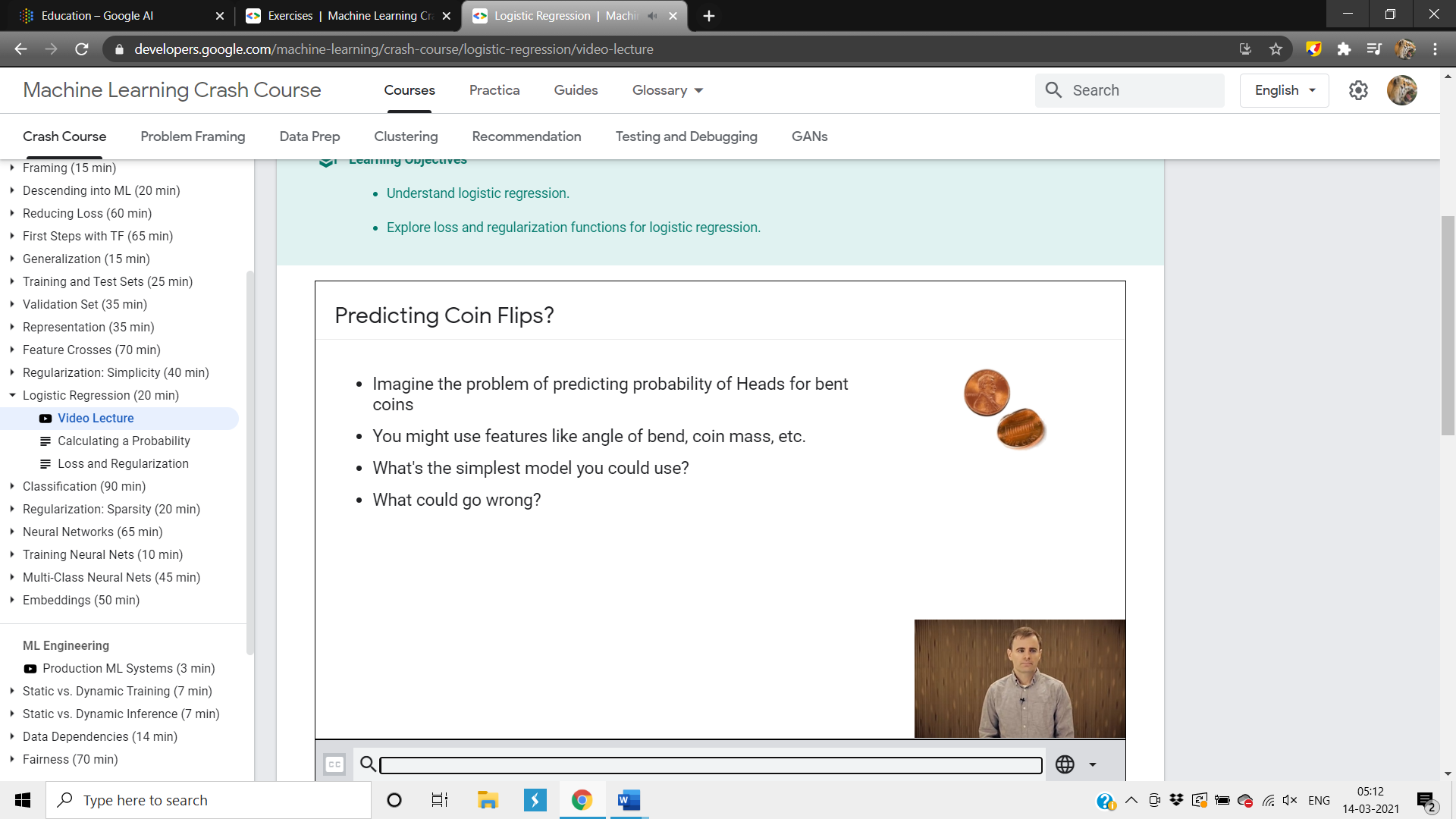
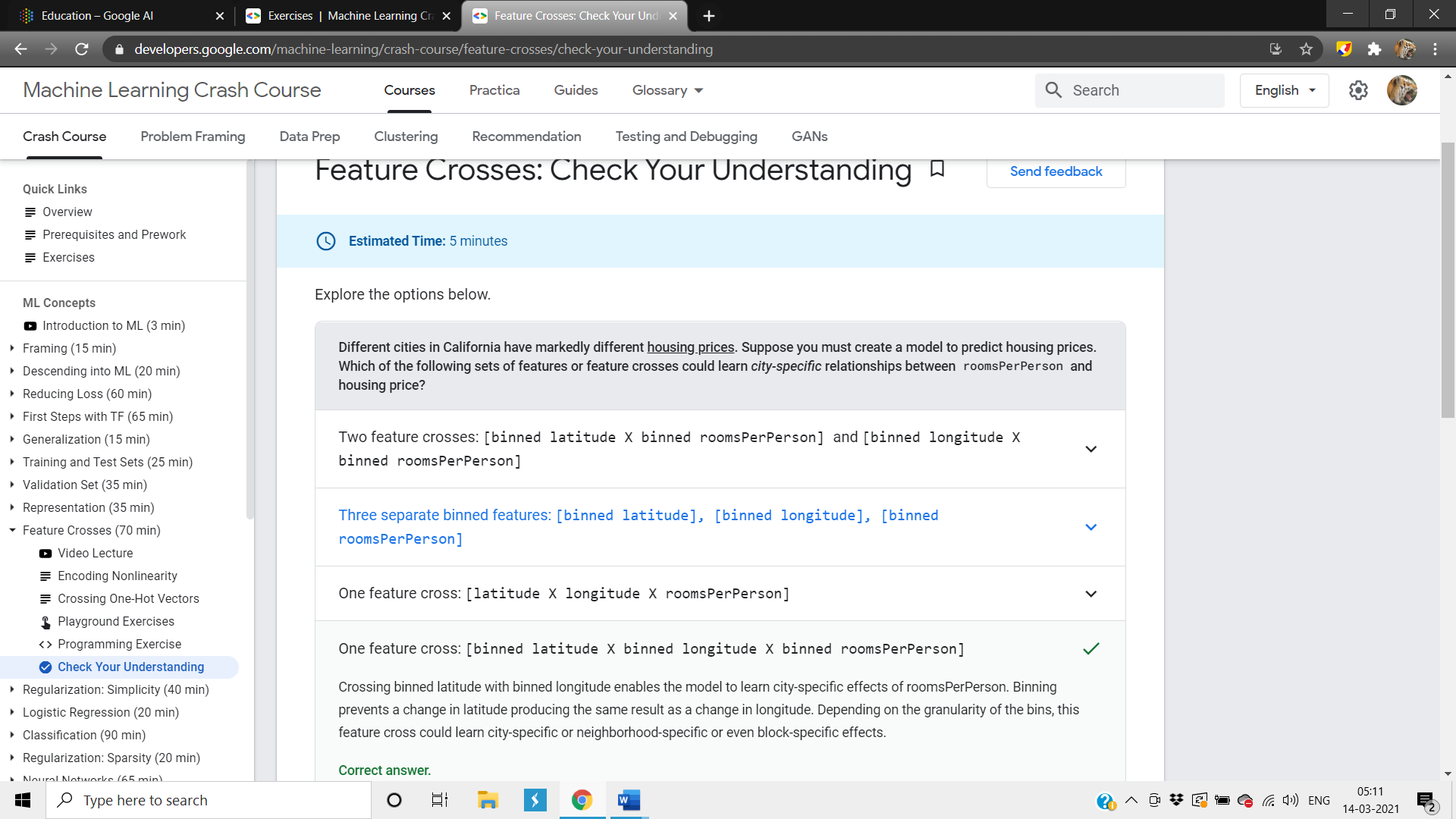
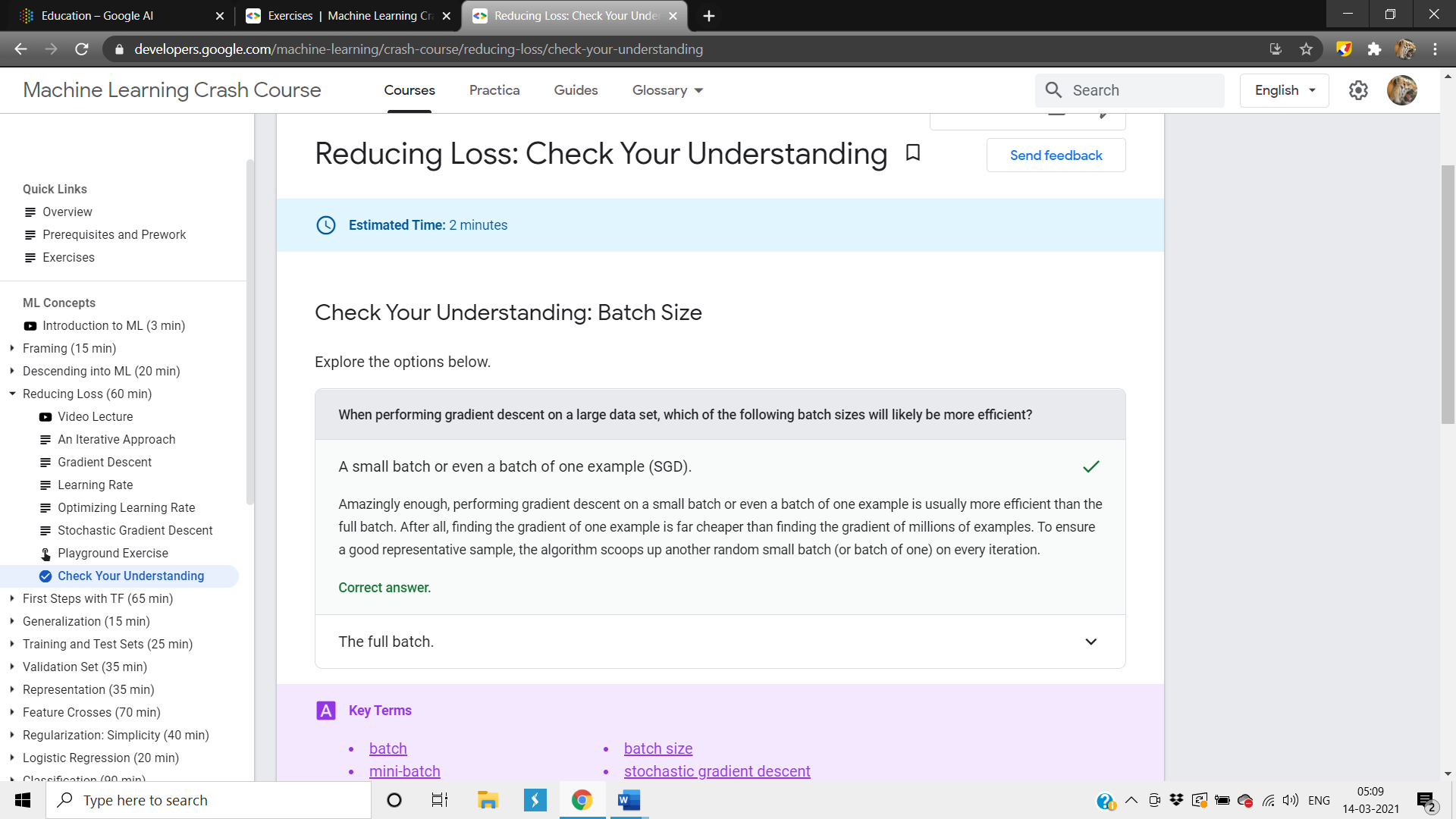
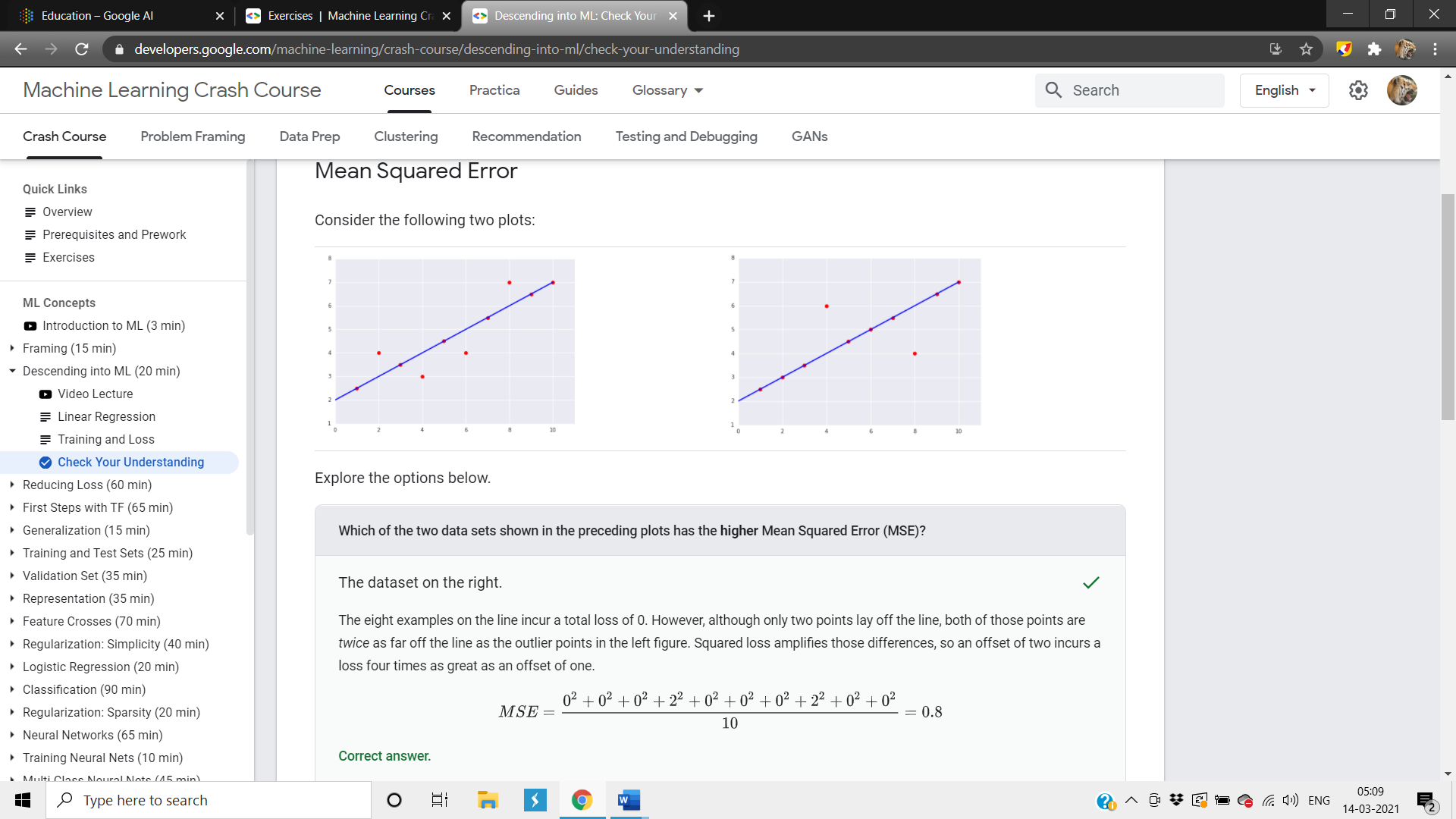
***POC for Machine Learning Problem Framing:***

******

***POC for Data Preparation and Feature Engineering: ***

***POC for Machine Learning Crash Course:***

******

******

***Google AI Datasets:***

|  |  |  |  |
| --- | --- | --- | --- |
| Dataset Name | Link / Website | Main  Purpose and Overview | Brief Recommendation of How and where to use it |
| 1. YouTube-8M | <https://research.google.com/youtube8m/> | * A complex large-scale dataset with an output variable and consists of millions of YouTube video IDs with several annotations * Contains 6.1 million IDs, and 3862 features | * Used for deep learning algorithms * Used for annotation learning or representation * Video domain segmentation model building * Transfer learning deep learning model applications |
| 1. YouTube Speakers | <http://people.csail.mit.edu/ludwigs/data/youtube_speakers_2013.txt> | * A big list of videos selected from the channel named GoogleTechTalks * These videos are segmented into a group based on Speaker | * It consists of 2 columns with speaker id and link of the video * The speaker id column is unique * Used for transfer learning deep learning models * Video segmentation analysis |
| 1. YouTube Bounding Boxes | <https://research.google.com/youtube-bb/> | * A complex large-scale dataset which consists of set of URLs for the videos * These are densely sampled single object BB annotations with high quality * Consists of 10.5 million annotations of Human and 23 different types of objects | * Used for Video classification model building * Used to advance the quality of Video Understanding model building applications |
| 1. Word2vec | <https://code.google.com/archive/p/word2vec/> | * It has two variants of word vector models * First contains 3 million words and 300 dimensional vectors of word embeddings represented as phrases * Second contains 1.4 million machine IDs with 1000-dimensional vector embeddings | * Used for NLP applications * Used to vectorize the words there by using it as re-usable code in pre-trained models |
| 1. Wiki Reading | <https://github.com/google-research-datasets/wiki-reading> | * A large scale NLU dataset with over 18 million observations * Data obtained from Wikipedia and wiki data | * Used to predict values for the text from the structured Wikipedia by reading the corresponding Wikipedia text in the articles published on their platform * Used for building the pre-trained models in text processing model buildings |
| 1. Wikipedia and arXiv similarity triplets | <https://research.google/tools/datasets/> | * Allows developers and researchers to validate and compare the different representation methods of the documents | * This is a dataset for evaluating document similarity models. In each file, each line consists of a triplet of URLs, either all from Wikipedia or all from arXiv.org. * The triplets in the file 'wikipedia-hand-triplets-release.txt' were hand generated whereas the other two files were generated automatically be examining Wikipedia and arXiv.org document categories. |
| 1. Wikilinks: 40 million Entities in Context | <https://github.com/google-research-datasets/wiki-links> | * An entity representation solving dataset consisting of 10 million web pages with over 40 million entities | * The links inserted in the pages across the web by the authors can be used to disambiguate mentions |
| 1. What’s Cookin’ | <https://research.google/tools/datasets/> | * A list of all cooking related videos IDs on YouTube along with the time stamps of the events | * Used for developing any machine learning model on self cooking and utensil classification by tracking the objects in the video |
| 1. Venezuelan Spanish [es-ve] multi-speaker speech dataset | <https://research.google/tools/datasets/venezuelan-spanish-tts/> | * The audio is recorded in a quiet room with high quality * It contains over 3350 sentences | * Used for speech technology * Used to develop some Natural Language processing models in order to translate one language to another language |
| 1. The Universal Dependency Treebank Project | <https://universaldependencies.org/> | * A treebank set used for annotating multiple languages in the same Stanford style dependencies | * The latest version 2.7 has 183 treebanks with 104 languages * This is very much useful for building pre-trained models and Natural Language processing models |
| 1. Telugu [te-in] multi-speaker speech | <https://research.google/tools/datasets/telugu-tts/> | * The audio is recorded in a quiet room with high quality * It contains over 4450 sentences | * Used for speech technology * Used to develop some Natural Language processing models in order to translate one language to another language |
| 1. Tamil [te-in] multi-speaker speech | <https://research.google/tools/datasets/tamil-tts/> | * The audio is recorded in a quiet room with high quality * It contains over 4250 sentences | * Used for speech technology * Used to develop some Natural Language processing models in order to translate one language to another language |
| 1. Syntactic Ngrams over Time | <http://commondatastorage.googleapis.com/books/syntactic-ngrams/index.html> | * It contains a corpus of 350 billion words which are analyzed from 350 million English language books | * These are analyzed words from several English language books which means instead of just translating the words of one language to another language they have been modified in such a way that a new word can be understood easily in the same language * Used for NLP translation models |
| 1. Sundanese [su-id] multi-speaker speech | <https://research.google/tools/datasets/sundanese-tts/> | * The audio is recorded in a quiet room with high quality * It contains over 4200 sentences | * Used for speech technology * Used to develop some Natural Language processing models in order to translate one language to another language |
| 1. Sinhala [si-lk] multi-speaker speech | <https://research.google/tools/datasets/sinhala-tts/> | * The audio is recorded in a quiet room with high quality * It contains over 2000 sentences | * Used for speech technology * Used to develop some Natural Language processing models in order to translate one language to another language |
| 1. SemCor and Masc documents annotated with NOAD word senses | <https://github.com/google-research-datasets/word_sense_disambigation_corpora> | * This package contains SemCor and Masc documents annotated with word senses of New Oxford American Dictionary * These are stored in XML format | * They use almost all the parts-of-speech available in the English corpus * Some of those tags are: ADJ Adjectives, ADV Adverbs etc. * Used for more accurate NLP translations as there is a deep understanding of the language |
| 1. Google cluster data 2019 | <https://research.google/tools/datasets/google-cluster-workload-traces-2019/> | * This is basically a trace that describe: Submission of jobs, Data used in these jobs, Make schedules for the decisions * The trace that is made in 2011 in the month of May, generated analysis and studies in hundreds | * Used for developing some mission comparison tool specially to deal with job tasks * Used to cluster the data based on the timestamps of the jobs given |
| 1. Audio set | <https://research.google.com/audioset/> | * Contains 632 event classes * Include 10 second sound clips taken from YouTube * There are 2084320 sound clips that are labeled by human * Have 2.1 million video clips * Defines the models that are hierarchal and that too exist between the sounds | * Used for the asynchronous audio processing model estimation with respect to the video clips using machine learning models * Used for predicting the type of Audio and video clips using classification models |
| 1. AVA dataset | <https://research.google.com/ava/> | * This dataset basically provides the annotation in audio-visual format * These are the annotations present in this dataset: AVA kinetics dataset, AVA actions dataset, AVA spoken activity dataset | * This is done in order to make the improvements in the understandings regarding the activities done by human * Used for both audio and video computer vision deep learning models |
| 1. Android smartphones high accuracy GNSS dataset | <https://research.google/tools/datasets/android-accuracy/> | * Dataset contains 39 logs * Each log is of 20 minutes * Google released the collection of the dataset from the android smartphones * There are 39 traces in data collected | * Acts as a bridge between advancement of the techniques of positioning and the applications of the positioning * Used for Geospatial Analysis and Object tracking model development |
| 1. Objectron | <https://research.google/tools/datasets/objectron/> | * It is the collection of object-centric videos that are of short duration. * These videos contain the objects that are captured from the different angles. * There are 15k videos in the dataset along with 4m images. * This collection is made from 10 different countries and 5 continents. | * Used for Object detection computer vision algorithms * Used to ensure the geo-diversity |
| 1. Dataset search: dataset for metadata | <https://research.google/tools/datasets/dataset-search/> | * The size of the dataset is 3.5 M * It is basically the collection of metadata from schema.org | * Aggregation, reconciliation and cleaning can be performed on data in order to display the search results of the datasets * This is very useful to obtain the optimal search results in the searching algorithms |
| 1. UK and Ireland English dialects | <https://research.google/tools/datasets/uk-ireland-english-dialects/> | * The collection of the data is made from the native speakers of the UK an Ireland area * Recordings which are available in the data are done by the native speakers in a quiet room | * Used for speech recognition technology * Used to build Natural Language Processing original language translation models |
| 1. DQN replay dataset | <https://research.google/tools/datasets/dqn-replay/> | * The size of the dataset is 3.07 TB. * Each of the game replay dataset is a way larger than the ImageNet dataset. It is approximately 3.5 times larger than ImageNet. | * These datasets are used to provide the training to the RL agents that are offline. * Any type of interaction in the process of training is not allowed. |
| 1. Taskmaster1 | <https://research.google/tools/datasets/taskmaster-1/> | * This dataset consists of the dialogs of English language that are only task based and they are 13215 in numbers. * These sentences are divided into two procedures. * There are 5507 dialogs that are spoken and 7708 are the written ones. * The type of the dataset is the text annotations. | * Used for text processing * Used for obtaining the necessary vector transformations of the words and later use it for the model building |
| 1. Dataset metadata for CORD-19 | <https://research.google/tools/datasets/dataset-metadata-for-cord-19/> | * This dataset consists of the pairs of the datasets that are paper based. * These dataset pairs are defined in the CORD-19 papers. * There are 16k rows in this dataset. | * It is defined as the collaboration in order to speed up the research regarding the coronavirus. * This collaboration is made between multiple institutes that works on the research of coronavirus. |
| 1. CoVid-19 open dataset | <https://research.google/tools/datasets/covid-19-open-data/> | * In this dataset, data is collected from the different countries all around the world regarding the COVID-19. * There are different aggregation levels that are applied on data. These are given below:  1. Countries 2. State 3. locally equivalent 4. municipality 5. locality  * There are different types of data available that is: * outcome data * static covariate data * dynamic covariate data | * Used to track the progress of the disease across the countries * Used to help the research organizations and pharma industries to be aware of the current situation there by scheduling their plans accordingly |
| 1. Taskmaster 2 | <https://research.google/tools/datasets/taskmaster-2/> | * The type of this dataset is text annotation. * There are 17289 dialogs available in the dataset. These are again divided into different domains. * There are more than 17000 conversations. | * Used for the user to believe that they are interacting with the machine that works automatically but they are actually interacting with the human. |
| 1. Coached conversational preference elicitations | <https://research.google/tools/datasets/coached-conversational-preference-elicitation/> | * This dataset contains almost 500 conversations. * There are 502 dialogs available in the English language. * In this there are some annotated utterances between an assistant and a user. * Both are discussing about the preferences of the movies in the natural language. | * The assistant asked some questions that are designed in order to minimize the bias * Used to tune the conversational applications more efficiently and reduce the bias and variance of the conversational model created |
| 1. Disco Fuse | <https://research.google/tools/datasets/discofuse/> | * The type of the dataset is the text annotation. * There are almost 60 million sentences in the dataset. | * The task in the dataset is to perform the join operation on the multiple sentences into a single text. * Used to reduce the amount of information available and get the gist of the data with small and accurate information thereby making model more robust towards the efficiency * This approach has been applied to the following documents collection: * Wikipedia * Sports article |
| 1. Room-across-Room | <https://research.google/tools/datasets/room-across-room/> | * It is the first language-based dataset that came into existence. * It contains 126069 navigation instructions in it. * Data is collected in three languages: * English * Hindi * Telugu | * Used for navigation in native languages * For getting navigation instructions, we can ask guide annotators to move along a path and generate the navigation instructions accordingly |
| 1. Open images | <https://storage.googleapis.com/openimages/web/index.html> | * Contains 16M bounding boxes under 600 categories. * In this dataset approximately 9M of the images are available with the annotations. * There are instance segmentations over 350 categories. | * It has been considered the largest existing dataset with object annotations. * Used for Image segmentation models there by using it for particular tracking of the objects with in the limits of the image |
| 1. Bengali ASR | <https://research.google/tools/datasets/bengali-asr/> | * The collection of the dataset is made from the native speakers of the Bengal. * Recordings in the dataset is done by the native speakers on their smartphones. * There are some quality checks that are performed on the dataset. * There are 21800 sentences in the dataset. | * Used to provide support to speech technology research * Used in the machine learning models which are used to translate the meaning from one language to another language * Hence used in Conversational applications |
| 1. Sinhala ASR | <https://research.google/tools/datasets/sinhala-asr/> | * The collection of the dataset is made from the native speakers of the Sinhala. * Recordings in the dataset is done by the native speakers on their smartphones. * There are some quality checks that are performed on the dataset. * There are 18500 sentences in the dataset. | * Used to provide support to speech technology research * Used in the machine learning models which are used to translate the meaning from one language to another language * Hence used in Conversational applications |
| 1. Kaggle datasets | <https://www.kaggle.com/datasets> | * It allows us to explore the dataset, share the dataset and create a dataset. * It also gives us platform to learn about data and its types and collaborating work. * It provides us the opportunity to learn the courses related to python and dataset. | * Kaggle datasets can be integrated with many platforms like GitHub and google colab * The main advantage of using Kaggle datasets is that we can directly use them on Kaggle kernels with GPU accelerated option provided by Kaggle for their users |
| 1. Nepali multi speaker speech | <https://research.google/tools/datasets/nepali-tts/> | * The collection of the dataset is made from the native speakers of the Nepal. * Recordings in the dataset is done by the native speakers voluntary. * There are 2000 sentences in the dataset. | * Used for speech technology * Used to develop some Natural Language processing models in order to translate one language to another language |
| 1. Khmer multi speaker speech | <https://research.google/tools/datasets/khmer-tts/> | * The collection of the dataset is made from the native speakers of the Khmer. * Recordings in the dataset is done by the native speakers voluntary. * There are 2900 sentences in the dataset. | * Used for speech technology * Used to develop some Natural Language processing models in order to translate one language to another language |
| 1. Javanese multi speaker speech | <https://research.google/tools/datasets/javanese-tts/> | * The collection of the dataset is made from the native speakers of the Khmer. * Recordings in the dataset is done by the native speakers voluntary. * There are 5800 sentences in the dataset. | * Used for speech technology * Used to develop some Natural Language processing models in order to translate one language to another language |
| 1. Bengali multi speaker speech | <https://research.google/tools/datasets/bengali-tts/> | * The collection of the dataset is made from the native speakers of the Bengal. * Recordings in the dataset is done by the native speakers voluntary. * There are 1850 sentences in the dataset. | * Used for speech technology * Used to develop some Natural Language processing models in order to translate one language to another language |
| 1. Javanese ASR | <https://research.google/tools/datasets/javanese-asr/> | * The collection of the dataset is made from the native speakers of the Sundanese. * Recordings in the dataset is done by the native speakers on their smartphones. * There are some quality checks that are performed on the dataset. * There are 185000 sentences in the dataset. | * Used for speech technology * Used to develop some Natural Language processing models in order to translate one language to another language |
| 1. NSynth | <https://magenta.tensorflow.org/datasets/nsynth> | * A very large dataset which is made up of high quality of annotated musical notes. * The audio data set contains almost 305,979 musical notes which is unique have different timber and envelope. * These musical notes are from different 1006 musical instruments. | * Helps us to build a model which identifies and can separate different musical notes. |
| 1. Nigerian English[en-ng] multi-speaker speech | <https://research.google/tools/datasets/nigerian-english-tts/> | * This data set is made up 3,350 transcribed audios which are of high quality. * These audios are in Nigerian and English languages which are recorded by permission of volunteers. * The audio in the dataset is of 48 kHz which is 16bit mono wave audio. | * Used in models which are made for text and speech translation and recognition between Nigerian and English languages. * Used for speech technology * Used to develop some Natural Language processing models in order to translate one language to another language |
| 1. Colombian Spanish[es-cl] multi-speaker speech | <https://research.google/tools/datasets/colombian-spanish-tts/> | * Contains 4,900 high quality transcribed audios * The audio in the dataset is of 48 kHz which is 16bit mono wave audio. | * Used in models which are made for text and speech translation and recognition between Colombian and Spanish languages. * Used for speech technology * Used to develop some Natural Language processing models in order to translate one language to another language |
| 1. Conceptual Captions | <https://ai.google.com/research/ConceptualCaptions/> | * Contains nearly 3.3 million images which are annotated with captions which are harvested from the web. * The raw descriptions which are harvested from the web are from Alt-Text attribute which is associated with web images. | * Helps us to perform an automatic image captioning which produces description for an image. * This can help people who have visual impairments by explaining images with the help of text to speech systems. |
| 1. Peruvian Spanish[es-pe] multi-speaker speech | <https://research.google/tools/datasets/peruvian-spanish-tts/> | * Contains 5,450 transcribed audios which are of high quality. * The audio in the dataset is of 48 kHz which is 16bit mono wave audio. | * Used for speech technology * Used to develop some Natural Language processing models in order to translate one language to another language |
| 1. Burmese[my-my] multi-speaker speech | <https://research.google/tools/datasets/burmese-tts/> | * Contains 2,500 high quality transcribed audios * These audios are in Burmese language which is recorded by permission of volunteers. * The audio in the dataset is of 48 kHz which is 16bit mono wave audio. | * Used in models which are made for text and speech translation and recognition of Burmese language. * Used for speech technology * Used to develop some Natural Language processing models in order to translate one language to another language |
| 1. Blocking Cross-Site Documents in Chrome | <https://www.chromium.org/developers/design-documents/blocking-cross-site-documents> | * Contains 5 percent values which are collected from Chrome browser usage data. * It shows the user requests which should be blocked and with a proposed security reason. | * Helps users unintentionally made request the harmful data with a valid security reasons to keep the user safe and secured. * This data also reflects on the actions which are made by the Chrome user base. |
| 1. HDR+ Burst Photography Dataset | <https://hdrplusdata.org/dataset.html> | * It contains Google HDR+ camera software images * It consists of 3640 burst images which are made from 28461 images | * Helps us to compare the images which are taken from Google HDR+ software and raw images. * The final result of the pipeline is provided in JPG format. |
| 1. Catalan [ca-es] multi-speaker speech | <https://research.google/tools/datasets/catalan-tts/> | * It consists of 4,200 transcribed audios which are of high quality. * These audios are in Catalan language which is recorded by permission of volunteers. * The audio in the dataset is of 48 kHz which is 16bit mono wave audio. | * Used for speech technology * Used to develop some Natural Language processing models in order to translate one language to another language |
| 1. RealEstate 10K | <https://google.github.io/realestate10k/> | * This data set consists of camera trajectories. * Evolved from YouTube videos | * Helps the researchers who work in 3D computer vision, graphics and view synthesis. |
| 1. ETA Exploration Traces | <https://github.com/google/cluster-data/blob/master/ETAExplorationTraces.md> | * Developed by Omega which is one of the cluster management systems at Google * tlp.trace: does the scheduling work and it is one of the parts of compressed tar file | * Used to find the bugs that are provoked events which are caused due to non-deterministic parameters such as time and asynchronous |

***Google AI Tools & Features:***

|  |  |  |  |
| --- | --- | --- | --- |
| Tools and Features | Link / Website | Overview | Recommendation of How to use it and where to use it |
| 1. AI Explanations | <https://cloud.google.com/ai-platform#section-6> | * To know the gist of how each independent variable contributed to the model’s inferences | * Used to find the inter correlations * Used to find the correlations of the input features with respect to the output feature * The most significant features can be selected for the model building there by reducing the noise of the model |
| 1. Auto ML | <https://cloud.google.com/ai-platform#section-6> | * To develop high performance machine learning models without even training the model in the trained data | * Auto ML contains several pre-trained models for which the models have been trained on the huge training data * So, we can use this tool if we have to develop a model with out training the data thereby reducing huge amount of time and cost * We can directly get the predictions from the pre-trained models on the new data * It also contains hyper-parameter tuning search feature which allows our model to get the optimal parameter values |
| 1. Continuous Evaluation | <https://cloud.google.com/ai-platform#section-6> | * To evaluate the machine learning models with appropriate metrics * Compare the predicted values with the actual values and optimize the performance accordingly | * Used to evaluate and validate our model on the new data * Based on the type of machine learning model, it will select the metrics * For instance, RMSE and MSE are used as metrics for regression problems whereas Accuracy and F1-score are used as metrics for classification problems |
| 1. ML Kit | <https://developers.google.com/ml-kit> | * A package the brings the machine learning expertise for mobile app developers * So, with ML Kit the mobile app developers can easily build machine learning apps by connecting APIs to the various applications using TensorFlow | * Highly optimized for the usage of mobile machine learning apps * We can use the following APIs based on the requirement and type of the projects: * Vision APIs for Face detection, Image Labelling etc. * Natural Language APIs for Entity Extraction, Translation etc. |
| 1. Fairness Indicators | <https://github.com/tensorflow/fairness-indicators> | * To know whether the model is performing without any bias or not * Fairness metrics for multi classifiers. * TensorFlow toolkit for fairness indication | * To see how fairly our model worked on the data * To know whether or not the predictions obtained from the model are fairly predicted * Helps to find whether or not out model generalized the data fairly * Used to Evaluate the distribution of the data * Used to deep dive into the model to explore new causes which can improve our model. |
| 1. Data Labelling Services | <https://cloud.google.com/ai-platform#section-6> | * To get the most appropriate labels which are labelled by the humans | * Labelling the data is very useful to understand the features more efficiently * It’s very important to label the data in order to perform the feature engineering phase of the machine learning development cycle more efficiently |
| 1. Deep Learning Containers | <https://cloud.google.com/ai-platform#section-6> | * A portable and convenient environment for building and deploying the models | * Using Containers, we can deploy the model after building and monitor its performance * Containers allow us to deploy the model without any model performance deviation problems after passing some new data to the model * Its highly recommended and used by several organizations across the industries to deploy their machine learning models in containers |
| 1. TensorFlow.js | <https://www.tensorflow.org/js/> | * To develop ML models in Java Script and use them directly on browsers | * We can convert the TF models to run in a browser or under Node.js by using JavaScript models * Load and save the models in TensorFlow.js. * Run the existing models and re-train them * Build and train models in JavaScript directly using APIs * Deploy TensorFlow.js Node Project on cloud |
| 1. Neural Network Architecture Search | <https://cloud.google.com/ai-platform#section-6> | * To build an entirely new application specific models with more accuracy, optimized power usage * Also, to improve an existing model with more accuracy and efficiency | * With NNAS, developers can design a model which is more powerful in terms of performance, efficiency in terms of power consumption etc. * Allows developers to improve the already existing model to increase its performance or tune its performance without disturbing the original model’s performance |
| 1. Colaboratory | <https://colab.research.google.com/notebooks/intro.ipynb#recent=true> | * To build projects or models directly on the browser * Train and develop neural networks. | * Google Colab allows us to develop models on the browser * We can change the TPUs and also use GPUs for the given free amount of RAM accessible to everyone * We can share our work directly to the google drive or integrate and share the work to GitHub and Kaggle |
| 1. Cloud Auto ML | <https://cloud.google.com/automl/> | * To develop high performance machine learning models without even training the model in the trained data | * We can build and deploy models containing different types of data like image, video, text, etc. * We can derive insights from object detection and image classification. * To know the structure and meaning of text data used for model * Detect and translate between languages using ML dynamically * So, we can use this tool if we have to develop a model with out training the data thereby reducing huge amount of time and cost on the cloud |
| 1. Tensor Flow | <https://www.tensorflow.org/> | * A deep learning tool or framework where can build deep learning models using its tools and packages | * TensorFlow contains all the necessary packages and libraries required for building deep learning models * TensorFlow interactive visualization package helps us to visualize the weights, bias and activation functions and how they are changing between the layers * It also contains call backs package that allows to call back certain phases of deep learning model during training considering the condition given to that model * TensorFlow uses Keras syntax and t acts like an API. So, every time we write a code in TF the interpreter will wraps the API to get the code execution done behind that and displays the output which is readable to us * We can easily train and deploy our model in the cloud. |

***Google AI Recommendation & Analysis:***

* Google provides ML and AI services to its customers through two different ways:

Google Cloud Machine Learning

Google Cloud Auto ML

* Google Cloud Auto ML is a cloud platform designed for in experienced users
* Developers can upload their datasets, train the models and deploy them on the websites or on the cloud
* Models which are trained can be deployed using Rest APIs
* Some of the different features and services offered by Google AI are as follows:
  + - **AI Explanations**: To know how each independent variable is affecting the model’s output
    - **Auto ML**: We can easily develop high performance machine learning models with out training
    - **Continuous Evaluation**: We can use the best metrics for evaluating and validating the model
    - **Data Labelling Services**: Labelling the data appropriately helps us to get the underlying patterns
    - **DL containers**: We can easily develop and deploy the model in a portable yet flexible containers
    - **Neural Architecture Search**: Build a new high-performance model or improve and tune the existing model
    - **Notebooks**: We can create and manage Virtual Machines with JupyterLab
    - **Predictions:** We can easily deploy the model and get the predictions from it which can be either batch predictions or normal predictions
* With ML kit, we can easily develop machine learning apps on mobile devices.
* We can use and integrate our models with various APIs using ML Kit in order to work them for a particular problem. Some of those problem statements with API integration are as follows:
  + - ***Vision APIs:***
      * Image and Video processing or analysis to label the entities
      * Barcode Scanning
      * Face detection
      * Image Labelling
      * Object detection and tracking
      * Text Recognition
      * Digital Ink Recognition
      * Pose detection
      * Selfie Segmentation
    - ***Natural Language Processing APIs:***
      * Using NLP APIs, we can identify and translate over 50 languages and answer the queries
      * Language ID
      * On-Device translation
      * Smart Reply
      * Entity Extraction
* Google Cloud Machine Learning Engine is designed for experienced data scientists
* GCP is testing its capabilities on various algorithms like XG Boost, Scikit-learn and Keras
* TensorFlow another GCP product, an open-source machine learning library used for developing deep learning models
* Let’s compare the Google AI platform with some other competitors in the market of Cloud:



Source: <https://www.altexsoft.com/blog/datascience/comparing-machine-learning-as-a-service-amazon-microsoft-azure-google-cloud-ai-ibm-watson/>

* Now from the above picture we can see that Google is able to process the speech and text data for over 120 languages for which it towers above the competition
* ***Conclusion:***
  + - GCP is offering some of the best services compared to its competitors in the market
    - One of the most fascinating things using GCP is its TensorFlow library
    - Using TensorFlow and its Keras syntax, I have created many deep learning models with out facing any trouble
    - Some of the deep learning models I have created using TF and Keras are as follows:

House Sales Predictions using Artificial Neural Networks

Pathogen cells detection or Image classification using Convolutional Neural Networks

Time Series Forecasting using Recurrent Neural Networks

Poem generation based on the input word using RNN

* + - To conclude, GCP is the best cloud service for anyone to start building an AI empire

References:

1. <https://developers.google.com/ml-kit>
2. <https://cloud.google.com/ai-platform#section-6>
3. <https://www.altexsoft.com/blog/datascience/comparing-machine-learning-as-a-service-amazon-microsoft-azure-google-cloud-ai-ibm-watson/>
4. <https://ai.google/education/>