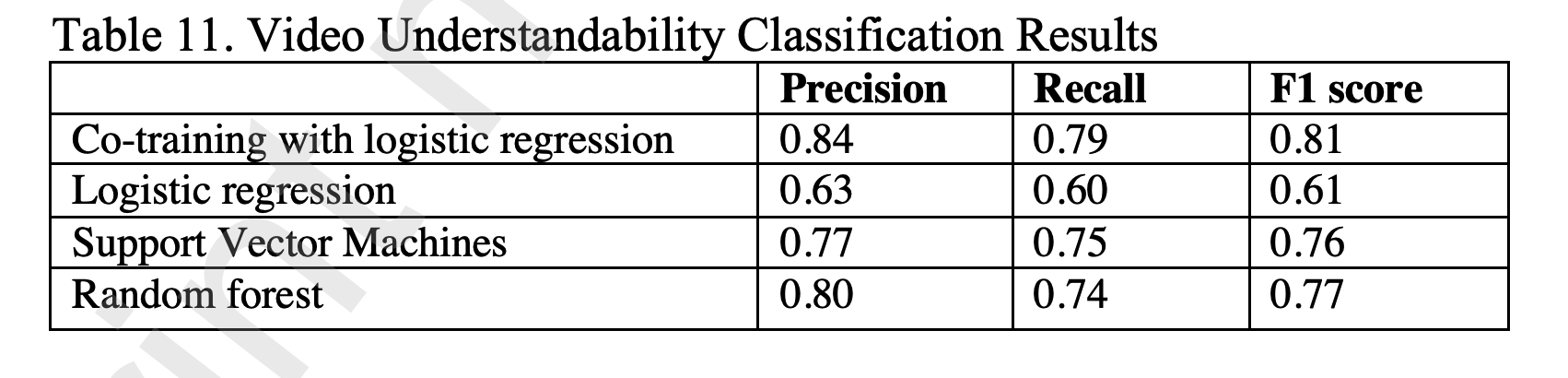
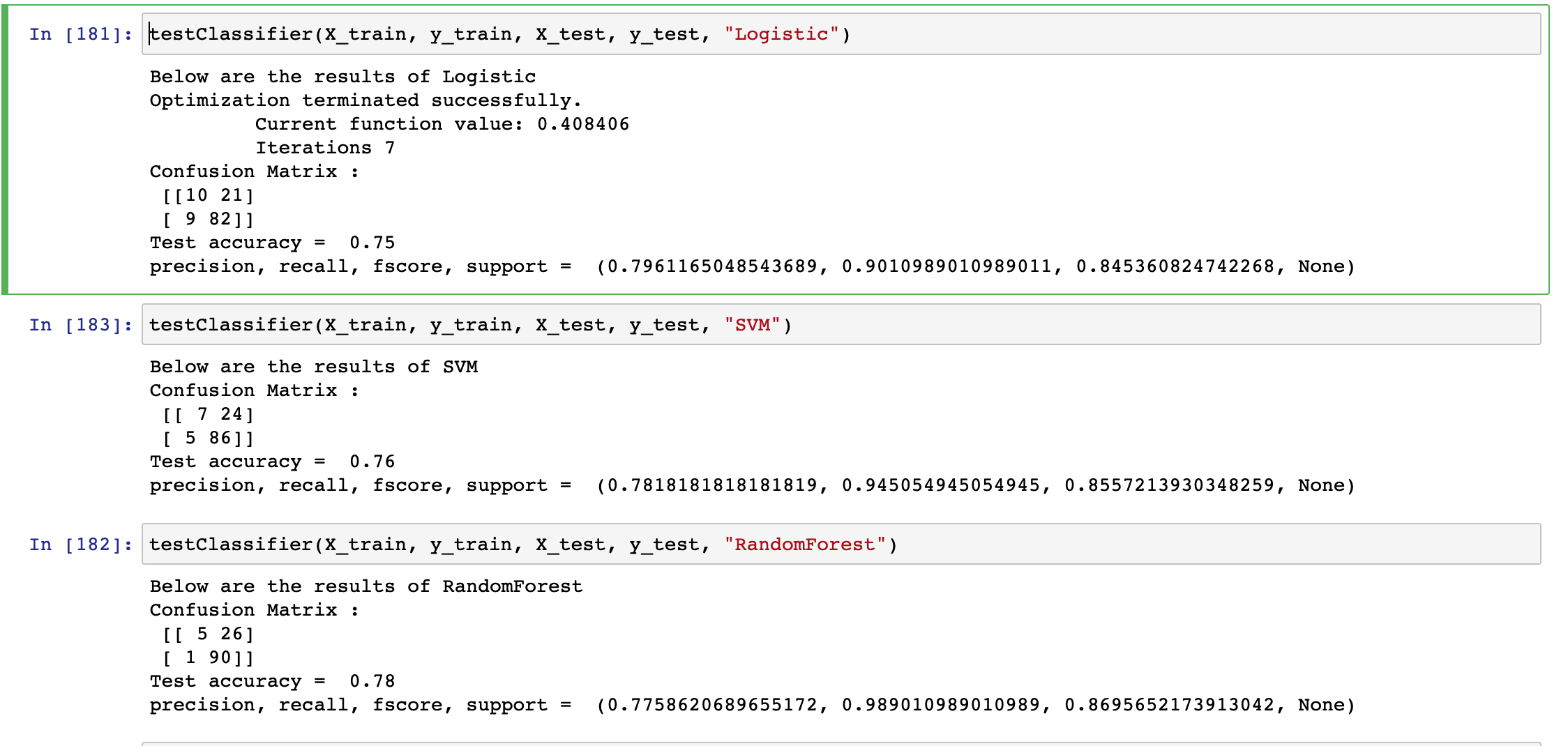
Peem Lerdputtipongporn

February 28th, 2022

**Questions I have about the ADA dataset**

1. **Many videos unrelated to diabetes appear in the dataset**. The dataset consists of top 50 videos under specific keywords. However, because some keywords have multiple interpretation (e.g, “starlix” is a diabetes drug, but is also a name of a Youtube channel), non-diabetes videos show up (see Appendix for details).
2. **Values calculated from text features are out of their appropriate range.**
   1. The library used to calculate cosine similarity between texts cannot handle special symbols. Because comments are stored as list represented as strings (think: “[What a great talk!, Hooray]”), every comment contains “[..]”, throwing off the value.
   2. The library used to calculate readability indices —ARI, Flesch’s reading ease, Kincaid — cannot handle sentence-less description. For instance, if “18740” is the description, Kincaid will be extremely high (one word, one sentence). On the other hand, some descriptions contain no full stop (many words, one sentence), making readability indices extremely low.
   3. If description starts with a link (https://…), word count is automatically set to zero even though there exists following description.
3. **Some relations between variables that should hold do not.** 
   1. Sentiment analysis should be applied to individual comments. Therefore, positive, negative, and neutral comment counts should add up to total number of comments. However, the former’s maximum values are 100 (so their sum are at most 300), whereas the latter’s maximum is 319,000.
   2. A few comments I have looked through have incorrect number of comments due to special symbols.
4. **Same videos appear multiple times under different keywords.** 
   1. How did the paper handle this issue when they built their models?
5. **I couldn’t replicate the results in** [**Ask Your Doctor’s paper**](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3711751)**:** 
   1. I used only the numerical features in this dataset (fewer variables that those in the paper). Theoretically, my result should be worse than what is shown in the paper. However, with logistic regression, I already achieved 75% in accuracy; 79.6% precision; 90% recall; and 84%. There’s no substantial gain in moving to SVM, randomForests.
   2. See my results versus the paper for details.



**Appendix: Distribution of relevantTopicIds**

Everything from Gaming has no connection to diabetes.

{'Society (parent topic)': 5812,

'Knowledge': 1401,

'Lifestyle (parent topic)': 5638,

'Health': 2035,

'Food': 580,

'Fitness': 193,

'Music (parent topic)': 1162,

'TV shows': 216,

'Movies': 166,

'Entertainment (parent topic)': 643,

'Technology': 158,

'Gaming (parent topic)': 94,

'Electronic music': 41,

'Hobby': 48,

'Vehicles': 31,

'Sports (parent topic)': 25,

'Rock music': 53,

'Pop music': 35,

'Hip hop music': 54,

'Military': 5,

'Politics': 12,

'Christian music': 4,

'Religion': 8,

'Performing arts': 65,

'Reggae': 7,

'Soul music': 3,

'Action-adventure game': 1,

'Role-playing video game': 17,

'Action game': 38,

'Pets': 49,

'Rhythm and blues': 1,

'Golf': 2,

'Cricket': 1,

'Fashion': 4,

'Simulation video game': 2,

'Music of Asia': 7,

'Motorsport': 4,

'Basketball': 7,

'Independent music': 16,

'Racing video game': 2,

'Music of Latin America': 23,

'Sports game': 3,

'Football': 1,

'Strategy video game': 1,

'Business': 3,

'Humor': 2,

'Boxing': 2,

'Classical music': 1,

'Jazz': 1,

'Baseball': 2}