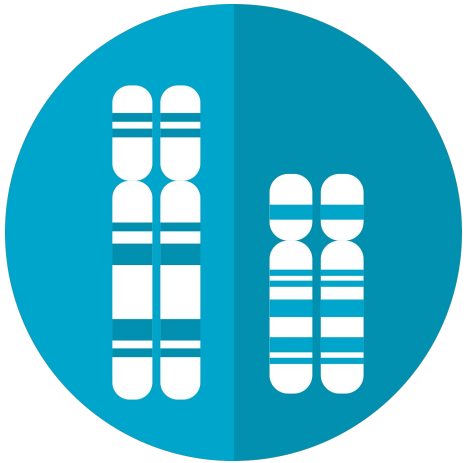
Brain Tumor Identification

Sprint 2

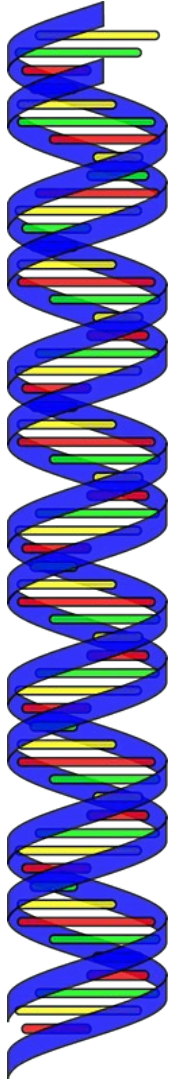


Low Grade Glioma

&

Glioblastoma Multiforme

~~Noah C. Strevell~~



Advancements

In Early Diagnosis

"Time is so meaningful in the face of a terminal diagnosis"
-Dr. Daniel Orringer

"To expedite diagnosis, Dr. Orringer has developed a device that uses advanced imaging techniques and artificial intelligence to predict a tumor's genetic makeup. The process takes just three minutes. So far, his research has found the results to be 93 percent as accurate as the current methodology"
- NYU Langone Staff

Dr. Daniel Orringer

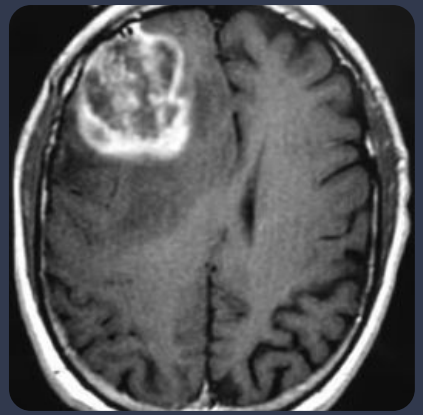
Article Link

Project Value

Patient Value	Drug Manufacturer Value	Health Insurance
• Accurate Diagnosis • Prognositc Information • Personalized Treatment • Clinical Trial Access • Reduced Uncertainty • Family Risk Assessment • Monitoring and Surveillance	• Targeted Drug Development • Market Differentiation and Competitive Advantage • Faster Regulatory Approval • Improved Drug Efficacy • Market Access and Reimbursement Support	• Cost-Effective Treatment Selection • Improved Patient Outcomes • Reduce Trial-and-Error Medicine • Efficient Use of Resources • Long-Term Cost Savings • Preventative and Predictive Care

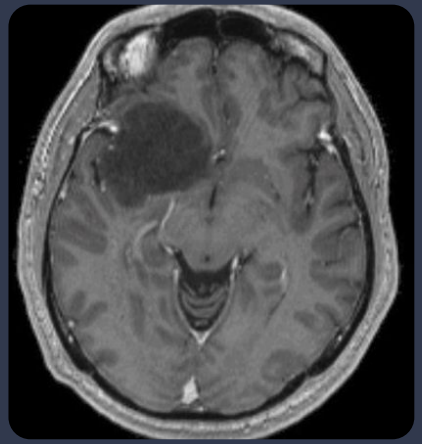
Target Variable

- Glioblastoma Mutiforme (1) or Low-Grade Glioma (0)
- Both tumors take place in the glial cells
- Glioblasotmas are identified as a Stage IV Glioma
- Low-Grade Gliomas have been known to transform into Glioblastomas.
- While molecularly visually similar, genetically different.
- Random occurrence except for
  - Secondary tumor
  - Genetic predisposition



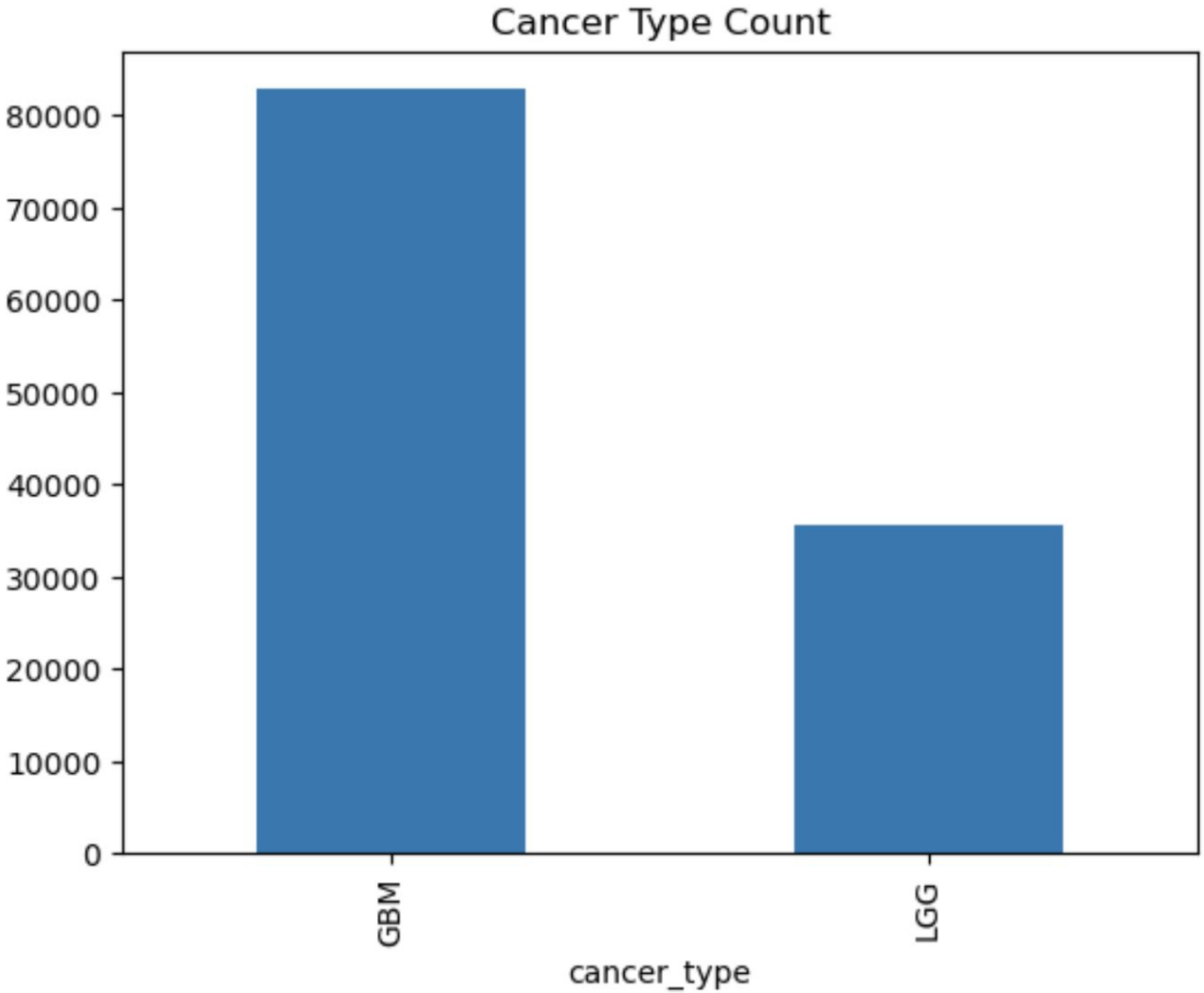
Glioblastoma Multiforme

Grade IV tumor. Grow and spread rapidly with low survival rates (median survival time of 15 months). The most aggressive type of glioma.



Low-Grade Glioma

Grade I-II tumor. Grow slowly and not as aggressive.


Cancer Type Count
80000
70000
60000
50000
40000
30000
20000
10000
0
GBM
LGG
cancer_type

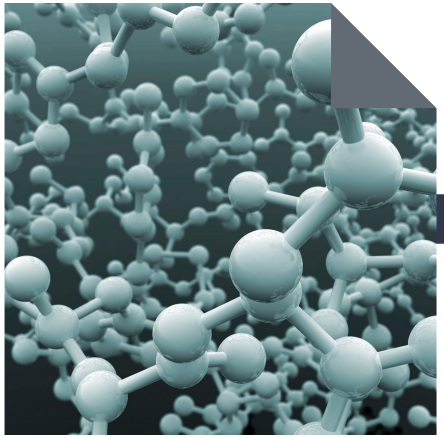
TCGA Pan-Cancer Atlas Selection
(118,322 records - 37 features)

Information	Column Names	TOTAL
Cancer Type	cancer_type	1
Chromosome Information	chrom, chromStart, chromEnd, chromStarts	4
Mutation Information / Description	name, freq, Variant_Classification, Variant_Type	4
Codon Information	thickStart, thickEnd	2
Allele Information	Reference_Allele, Tumor_Seq_Allele1, Tumor_Seq_Allele2	3
Gene Location / Info	blockCount, blockSizes, Hugo_Symbol, Entrez_Gene_Id	4
Further Investigation	strand	1

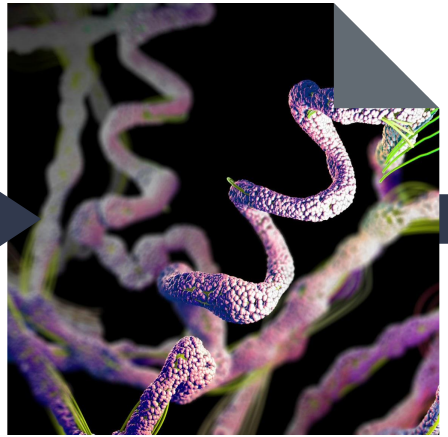
Information	Column Names	TOTAL
Patient Information	days_to _death, cigarettes_per_day, weight, alcohol_history, alcohol_intensity, bmi, years_smoked, height, gender, ethnicity	9

Information	Column Names	TOTAL
Unknown / Irrelevant	score, reserved, sampleCount, dbSNP_RS, dbSNP_Val_Status, project_id, Tumor_Sample_Barcode, Matched_Norm_Sample_Barcode, case_id	9

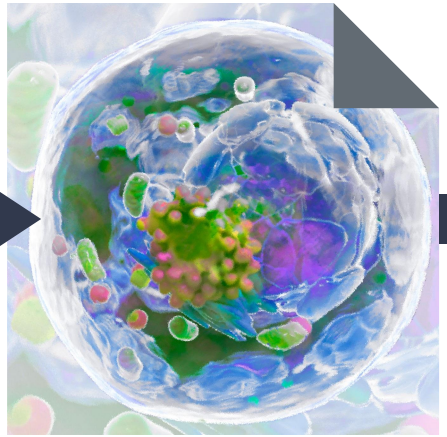
Amino Acids & Proteins



Amino Acids



Proteins

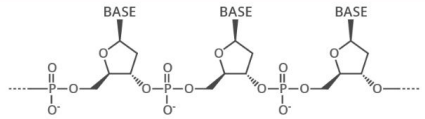


Living Cells

THE CHEMICAL STRUCTURE OF DNA

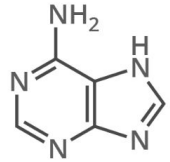
DNA (deoxyribonucleic acid) carries genetic information in all multicellular forms of life. It carries instructions for the creation of proteins, which carry out a wide range of roles in the body.

THE SUGAR PHOSPHATE 'BACKBONE'

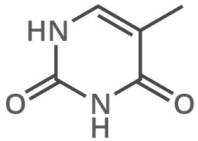


DNA is a polymer made up of units called nucleotides. The nucleotides are made of three different components: a sugar group, a phosphate group, and a base. There are four different bases: adenine, thymine, guanine & cytosine.

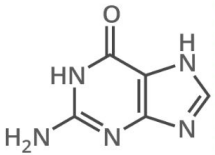
A ADENINE



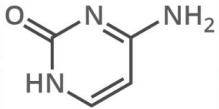
T THYMINE

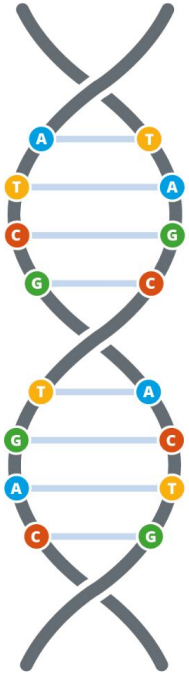


G GUANINE



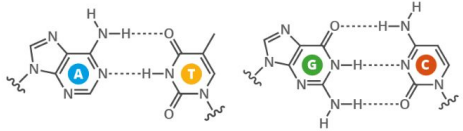
C CYTOSINE





WHAT HOLDS DNA STRANDS TOGETHER?

DNA strands are held together by hydrogen bonds between bases on adjacent strands. Adenine (A) always pairs with thymine (T), whilst guanine (G) always pairs with cytosine (C).



FROM DNA TO PROTEINS

The bases along a single strand of DNA act as a code. The letters form three letter 'words', or codons, which code for different amino acids - the building blocks of proteins.

An enzyme, RNA polymerase, transcribes DNA into mRNA (messenger ribonucleic acid). It does this by splitting apart the two strands that form the double helix, then reading a strand and copying the sequence of nucleotides. The only difference between the RNA and the original DNA is that in the place of thymine (T), another base with a similar structure is used: uracil (U).

In multicellular organisms, the mRNA carries genetic code out of the nucleus, to the cell's cytoplasm. Here, protein synthesis takes place. 'Translation' is the process of converting turning the mRNA's 'code' into proteins. Molecules called ribosomes carry out this process, building up proteins from the amino acids coded for.

Feature Engineering – Part 1

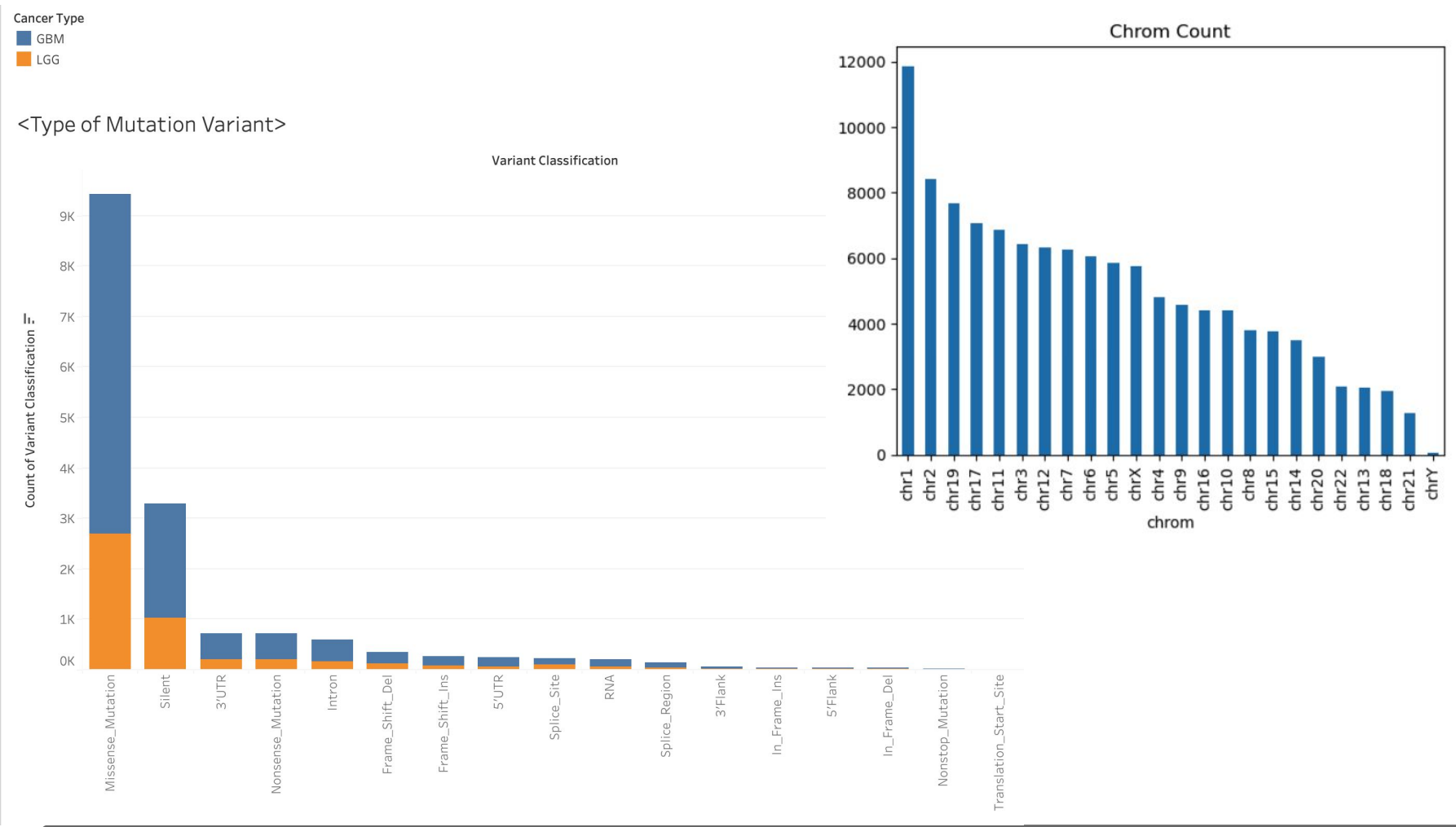
Original Feature Name	Transformation	Number of Features	Initial Column Dropped	New Feature Name(s) Example
cancer_type	Binary Encoding	1	No	cancer_type
chrom	Dummy Variables	22	Yes	chr19
chromStart / chromEnd	chromStart - chromEnd = chrom_len -> ordinal encode	7	Yes	Mut_Len_1-9
Variant_Classification	Dummy Variables	18	Yes	5'Flank
Variant_Type	Dummy Variables	3	Yes	SNP
Hugo_Symbol	Dropped	0	Yes	n/a
gender	Binary Encoding	1	No	gender
chrom	2 new boolean features, 1 if sex chromosome, 1 if autosome	2	Yes	autosome / sex_chromosome

Feature Engineering – Part 2

Original Feature Name	Transformation	Number of Features	Initial Column Dropped	New Feature Name(s) Example
‘Reference_Allele’	Binarize. 1 =- at CpG site (often hypermutable) 0 = not at CpG site	1	Yes	CpG_Site : 0 / 1
‘Variant_Classification’	Ordinal Encode based on variant severity	1 (High Risk - Low Risk)	Yes	Variant_Severity: 1 - 4
‘name’	Convert to boolean column. 1=transition. 0=transversion	1	Yes	transition_transversion: : 0 / 1

Number of features before vs after feature engineering

37 to 40


Cancer Type
GBM
LGG
<Type of Mutation Variant>
Variant Classification
Count of Variant Classification
9K
8K
7K
6K
5K
4K
3K
2K
1K
0K
Missense_Mutation
Silent
3'UTR
Nonsense_Mutation
Intron
Frame_Shift_Del
Frame_Shift_Ins
5'UTR
Splice_Site
RNA
Splice_Region
3'Flank
In_Frame_Ins
5'Flank
In_Frame_Del
Nonstop_Mutation
Translation_Start_Site
Chrom Count
12000
10000
8000
6000
4000
2000
0
chr1
chr2
chr19
chr17
chr11
chr3
chr12
chr7
chr6
chr5
chrX
chr4
chr9
chr16
chr10
chr8
chr15
chr14
chr20
chr22
chr13
chr18
chr21
chrY
chrom

Model Results

Since this is a binary classification problem, the following three model types were used: Logistic Regression, Decision Tree Classifier, and K-Nearest Neighbors. A pipeline was used to determine the best hyperparameters for each model.

Logistic Regression

- Accuracy = 70.40%
- Precision = 70.59%
- Recall = 99.25%
- F1 Score = 83%

	Predicted LGG	Predicted GBM
True LGG	147	6879
True GBM	123	16516

Decision Tree Classifier

- Accuracy = 70.46%
- Precision = 70.69%
- Recall = 98.79%
- F1 Score = 82.47%

	Predicted LGG	Predicted GBM
True LGG	236	6790
True GBM	200	16439

K-Nearest Neighbors

- Accuracy = 70.50%
- Precision = 70.76%
- Recall = 98.93%
- F1 Score = 82.50%

	Predicted LGG	Predicted GBM
True LGG	223	6803
True GBM	178	16461

What’s Next?

Noah C. Strevell
https://github.com/NoahNoahCS/Capstone